Unsupervised Term Selection using Entropy

Héctor Jiménez-Salazar, Mauricio Castro, Franco Rojas, Ernesto Miñón, David Pinto & [1]Helena F. Carcedo

Facultad de Ciencias de la Computación,
[1]Instituto de Ciencias,
B. Universidad Autónoma de Puebla,
hjimenez@aleteya.cs.buap.mx, mcastro@cs.buap.mx, flopez@cs.buap.mx,
eminon@cs.buap.mx, dpinto@cs.buap.mx, fcarcedo@hotmail.com

Abstract. Term selection is a very important problem for text representation in Natural Language Processing. Several methods have used entropy, a measure of the information, in order to select "best" features; although, most of the time, such selection is done a posteriori, namely supervised. Supervised methods have better results with respect to unsupervised methods, but they require past experience. In this paper we propose an unsupervised term selection method using entropy. This method enhances F_1 measure reducing the dimensionality of term space of an Information Retrieval System using TREC-5 Spanish collection.

1 Introduction

Text representation is necessary for many Natural Language Processing (NLP) tasks. State of the art, in text representation dealing with a texts collection, reveals most of the techniques using "bag of words" approach; i.e. the use of terms contained in the text, in order to represent it. This approach is based in the Vector Space Model (VSM) [3], in which a weight is assigned to each term, regarding the frequency of terms in each document and the number of texts that use such term: $tf_{ij} \cdot idf_j$; thus, each dimension is related with a term. Typically, even a moderately sized collection of text has tens or hundreds of thousands of terms. Hence, the document vectors are high-dimensional. Because of this high-dimensionality of term space, term selection methods have been developed.

As an instance, in Text Categorization (TC), defined as classification of documents into a set of predefined categories [7], three well known term selection methods are: DF (Document Frequency), the number of documents in which a term t_i occurs (df_i); CHI (CHI-statistic), measures the lack of independence between a term and the category, then CHImax is computed for a term obtaining the maximum on all categories; IG (Information Gain), measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term in a document, IGsum for a term represents the expected information gain on categories. These three term selection methods are effective in the TC task [11]. Furthermore, same methods, but not limited to them, are used in diverse tasks as in Text Clustering [8].

Term selection methods are divided into supervised, taking into account a training set; and unsupervised, without previous knowledge to solve a kind of problems. DF is, perhaps, the most simple and effective unsupervised method. Besides, CHI and IG are supervised, and they obtain better results in almost all cases. Certainly, supervised methods are preferred due to the high performance attained. However, a great number of problems cannot dispose of a training set; perhaps because the problem uses a very specialized corpus, or the results would be not possible to transport from a domain to another one. Also, supervised and unsupervised term selection methods may cooperate to select better terms in applications of NLP and others (bioinformatics, pattern recognition, etc.).

In this paper, entropy, a measure of information of each word of the vocabulary of a given text collection, was calculated in order to select terms in the framework of an Information Retrieval System (IRS); each text is represented both, using all terms contained in the text and only those which were selected by our method. We will show this method permits to replace documents of text collection by the terms selected with equiparable performance, and in some cases it outperforms the classical representation.

Next two sections of this paper deal with two approaches on the behavior of terms, the first one is anchored on some results of manual text analysis provided from a project in course; second one presents the main result of Information Theory. Section four presents the results of our experiments, and, finally, we construe a brief discussion of findings.

2 Empirical Framework

An study on lexical qualification of text writers was carry out at BUAP university. A corpus composed by 135 cases of words from a sample of 1311 university students was collected. The analysis of such corpus is now runing. However, due to semantic extension of terms are constrained by the context, it is confirmed that precision about meaning is related with frequency of occurrence of terms. From this claim we conjectured a proposition concerned with relevance of terms in a text collection. Some details on the study are presented in the remainder of this section.

Manual inspection of the use of terms in documents of academic issue, entails to propose a scale of values based on the inversely proportional relationship between frequence of a term, and the amount of information that this term has. Three ranges of this scale were identified: an informal set, coloquial or popular type (from 0 to 0.33 of relative frequency); standard set (from 0.34 to 0.66); and specific sublanguage, academic, scientific or technical type (from 0.67 to 0.99). We established a measurement scale based on the relation between frequency of occurrence of a term (related with its significance extension), and the amount of information of such term (related with the precision in the use of the term, which is typical for scientific and technical documents), limited by the following extremal points:

- More extension, high frequency, which implies a low requeriment of ocurrences.
- More precision, high amount of information, which implies a high number of ocurrences.

If we consider the scale partitioned from 1 to 9 for frequency and amount of information (in inverse relationship), we can divide terms into three sets:

First Set : It is composed by extensive terms, with a high probability to be applicable in a diversity of environments, which implies a low number of ocurrences (it must be included in no more than 0.33 of pertinent lexical manage).

Second Set : Term frequence relation is equilibrated and this corresponds to the standard registry. Featured by an active and non-active manage of a little bit higher occurrences, derived by a higher degree of precision, and therefore, less extension and a higher number of occurrences (0.34-0.66 of pertinent use of lexical corresponds to this segment).

Third Set : Scientific and technical registry; corresponds to the academic sublanguage, with a use between 0.67 and 0.99. It has a high level of precision. The number of occurrences of words required is higher than the other two sets, due to the high precision, which transform them in uncommon words or infrequent (because of their less significance extension).

Partial findings of this study aimed to take advantage from frequence of occurrence of words. Equilibrated use of words conveys they are used at the transit from extensive terms to precise terms. So, the high information ported by precise terms is taken from the context and it is provided by terms of second set which may have high entropy. Thus, an observer, outside of the context, gains information of texts from this kind of terms. Next section exposes the mathematical tool that formalizes previous remarks.

3 Theoretical Framework

Theory of Information was presented in *Bell System Technical Journal* in october of 1948 [1]; a little time after, Warren Weaver redacted an essay named *Recent Contributions to the Mathematical Theory of Communication* with the goal to emphasize benefits of that theory, developed by Shannon. One year later, a book named *The Mathematical Theory of Communication* was published by the Illinois University (1949) discussing communication issues (Shannon's formula) from a mathematical view. The idea that sublies in the work of Shannon is to consider an information measure from the different possibilities that a system has. This leds to define the information as follows:

Let X be a random variable that can take the values $x_1, x_2...x_n$, each one with probabilities $p_1, p_2, ...p_n$, then, the amount of information contained in the system is named entropy and it is defined as follows:

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$

This means that a deterministic system (i.e., a system that does not depends of any random variable), lacks of information $H(x) = 0$, since $p(x_i) = 1$ (the same argument is valid if $P(x_i) = 0$). On the other hand, a system whose random variables have the same probability ($p(x_i = 1/n)$) will have the maximum of information $H(X) = \log n$. We must not to confuse the information that a system has, and the information that can be extracted from it; in other words, as less information we have from a system, bigger amount of information the system will have (the information of a system is a measurement of our ignorance).

Despite the strong criticism process that communication people have applied to this approach, it have had a strong impact on several problems related with communications, such as, storing and processing information in physical media. If we establish that the same message can be written in different ways (with different sequence of words), then it is possible to divide words belonging to the message in: a set of words W_1, that have a high probability to appear in all the documents ($P(W_1) = 1$); a set of words that are not uniformly distributed in the collection of texts, i.e., those which are concentrated in some document. And finally, those words that are uniformly distributed in a corpus. The last one, has a high value of significance in terms of information, compared with the two former. Recently, Marcelo A. Montemurro [2] used the entropy concept for sorting sets of words, based on the roll that these words play in a set of documents of literature. He did an statistical analysis of these words without knowledge of the gramatical structure of the documents analized.

Determination of a set of words that characterize a set of documents given, is matter of our work. Given a set of documents $D = \{D_1, D_2, ..., D_m\}$, N_i the number of words in the document i (D_i), and tf_{ij} the frequency of the word j in D_i, the relative frequency of the word j in D_i is:

$$f_{ij} = \frac{tf_{ij}}{N_i}, \tag{1}$$

and

$$p_{ij} = \frac{f_{ij}}{\sum_{j=1}^{m} f_{ij}} \tag{2}$$

is the probability that word j be in D_i. Thus, entropy of word j is given by

$$S_j = -\sum_{i=1}^{m} p_{ij} \log p_{ij}. \tag{3}$$

4 Text Representation in Information Retrieval

In order to remark the difference between the proposed methods and others based on entropy, let us take IG method applied to TC. Let D be the training documents set, and $\{c_k\}_{k=1}^{M}$ be the set of categories. The information gain of a term t_i is defined as

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

the probabilities are interpreted on an event space of documents, and are estimated by counting occurrences in the training set. IG_j score of terms allows ordering of the vocabulary and to select terms with higher score; commonly 10% of terms is a good percentage. As we can see, this method uses probability on categories; which means using a posteriori probabilities.

We carry out two experiments aimed to know the efficacy of using entropy, then we select terms of high entropy value in order to represent each text. Such experiments were posed in the Information Retrieval (IR) task; i.e. we evaluate the efficacy of a representation giving a query to the IR system, and we measure the F_1 coefficient. In this way, the classical representation and the reduced representation by the term selected were compared.

As we have said, our concern are medium-frequency high-precise terms, because they contain more information than others. Therefore, it is necessary to identify medium frequency terms. Transition point (TP) is the frequency that divides the vocabulary of a text into low and high frequency terms [4], [5], [6]. Through the formula $PT = (\sqrt{8 \cdot I_1 - 1} + 1)/2$, where I_1 is the number of words with frequency 1, TP may be computed. Second requirement is supported by equation 3.

Some subsets of TREC-5 Spanish collection were used in order to carry out the experiments. Table 1 shows the composition of these corpora, and table 2 the corresponding queries for each corpus.

Subset	Query	#Docs	#Relev.
2	c_4, c_5	1048	97, 257
3	c_{10}, c_{11}	933	206, 105

Table 1. TREC-5 subset for 4 topics.

We used precision (P), recall(R) and $F1$ coefficient in order to evaluate our results [9]: $P = a/b$, $R = a/c$, and $F_1 = \frac{2 \cdot P \cdot R}{P+R}$, considering a as the number of relevant documents obtained by the IRS, b as the number of documents obtained by the IRS, and c as the number of relevant documents.

Id	Query
c_4	Papel de México en la OEA.
c_5	Maquiladoras en la economía mexicana.
c_{10}	México es importante país de tránsito en la guerra antinarcótica.
c_{11}	Derechos a las aguas de los ríos en la región fronteriza entre Mexico y los Estados Unidos.

Table 2. Queries considered in our subsets of TREC-5.

4.1 Experiment 1

Representation Schema The representation of a document D_i is given by the boolean representation, whenever terms have high entropy. Let $D'_i = [1_j | t_{ij} \in D_i]$ be the boolean representation of a document, and H_{max} the maximum value of entropy, the representation based on entropy of D_i is

$$D''_i = [1_j | t_{ij} \in D_i, H(t_{ij}) > H_{max} \cdot u], \tag{4}$$

where u is a threshold to define the level of high entropy.

Results Figure 1 displays precision for each recall level from the response of IR system, when $u = 0.05, 0.1, 0.15, 0.2$ were used in equation 4, with the query c_{10} in the subset 2, and the curve for boolean model (BM). The number of terms for each threshold was: 17,901, 15,861, 11,757, 9,072, for $u = 0.05, 0.1, 0.15, 0.2$, respectively; and the total number of terms (boolean model) was 35,596. The best threshold in this test was 0.15, using only 33% of terms and equiparable performance in terms of F_1.

4.2 Experiment 2

Representation Schema Given a document D_i, let $V_i = \{(x, y) | x \in D_i, y = fr(x)\}$ be its vocabulary with frequencies fr. Let us define a $u-$neighborhood around transition point, TP_i, as: $R_i = \{x | (x, y) \in V_i, TP_i \cdot (1-u) \leq y \leq TP_i \cdot (1+u)\}$, where u is normalized in $[0, 1]$. The terms that lie in the $u-$neighborhood of D_i are re-weighted in the following way. We used the classical weighting formula $tf_{ij} \cdot idf_j$ for each term t_j of document D_i, changing tf_{ij} by tf'_{ij}, which is defined as:

$$tf'_{ij} = \#R_i - |TP_i - tf_{ij}|. \tag{5}$$

This weighting schema is based on importance of terms; the more closeness of term t_j to TP_i, the more weight for t_j. Previous selection and weighting schema is local to each document, but it is blinded about the whole of the text collection. In the collection, distinguished terms will be those which are used in a "balanced" manner in each text where they appear; this is to say, an equiprobable ocurrence,

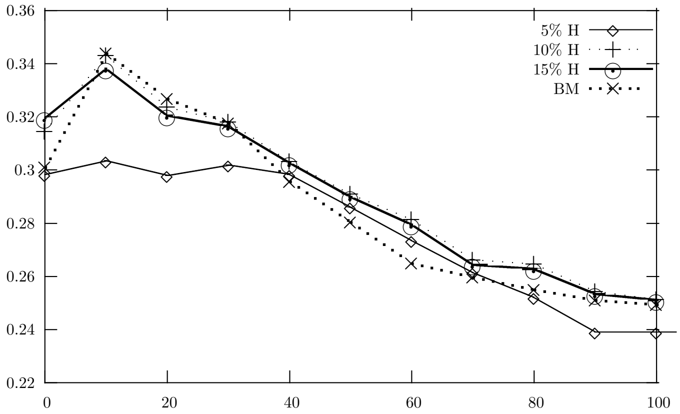

Fig. 1. Precision per recall levels for query c10.

or maximum uncertainty. So, we "cut" terms if they have a low value of entropy. Specifically, in this experiment, we replace equation 1 by

$$f_{ij} = \frac{tf_{ij}}{N_i^{tf_{ij}}}, \tag{6}$$

and H' will denotate computation of H according to equations 6 and 3.

Term selection was done considering terms t of a document D_i, such that $H'(t) \geq H'_{max}/2$ were retained to represent D_i; where H'_{max} is the maximum entropy value in the collection using equation 6.

Results We used subsets 2 and 3 of TREC-5 Spanish text collection which are depicted in table 1.

Given a text collection, each document was represented using the above schema, and a new text collection was formed. The corresponding queries (see table 2) was done to the IRS. Table 3 shows F_1 coefficient and dimensionality of term space (divided by a slash symbol) for representation based on selection using PT and entropy (column 2); using entropy alone, i.e. terms hold an entropy greater than $H'_{max}/2$ (column 3); and the classical model (column 4). Figures 2 and 3 show curves of precision for recall levels using entropy selection *vs.* Vector Space Model.

5 Discussion

A method based on entropy to select terms was proposed. Tests on an Information Retrieval System using subsets of TREC-5 Spanish texts collection were done. The results conveys to analyze a new text representation where reduced

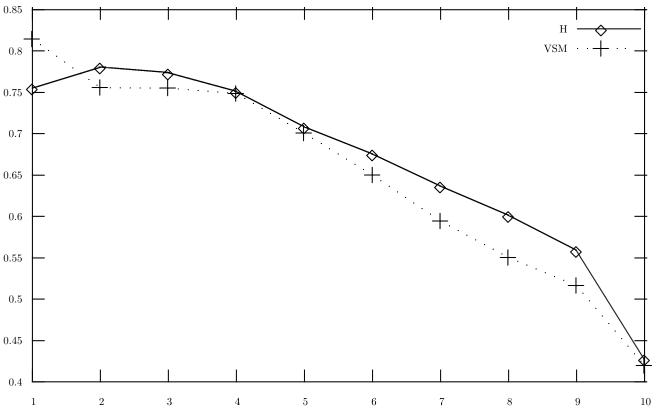

Fig. 2. Precision per recall levels using subset 1.

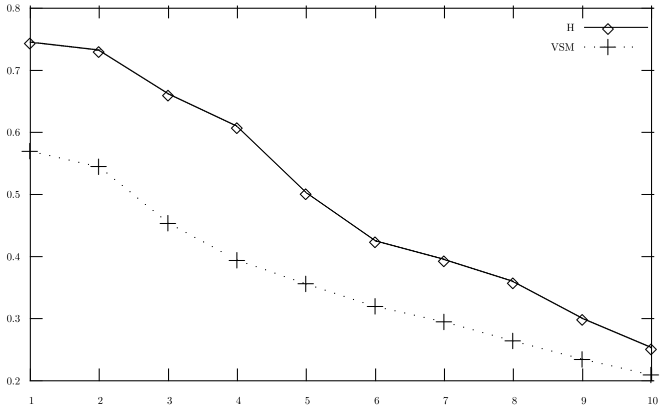

Fig. 3. Precision per recall levels using subset 2.

Subset	PT&*H*	*H*	VSM
2	0.398/820	0.625/1,118	0.410/39,967
3	0.316/800	0.309/1,462	0.286/35,596

Table 3. F_1 values and number of term used by each representation.

text using terms selected by the method may replace original texts. Entropy value of terms used as a selector of terms behaves well using terms t with $H(t) > H_{max} \cdot 0.15$, reducing 33% the term space. It is not clear yet if transition point combined with entropy helps to the representation; in the first test TP decrements F_1 value 2.9%, and in the second one increments by 9.4%, but using only entropy increments in 52.4% and 8% respectively.

Entropy property of reaching maximum value with equiprobable outcomes says that the terms are used, among texts, with a relative constant frequency. This is an indicator supported by intertextual frequency on a text collection. Therefore, it would not be possible to apply the method on isolated texts or heterogeneous texts collections.

Our conjecture is that terms with balanced use through the texts collection is related with Zipf's Law [12]: minimum effort to write a text entails a moderate use on some words revealed by entropy. Dealing with several texts, we interpret preserving regularity of occurrence of such words as if they were relevant because of its role in the texts as pivots.

Nevertheless the results obtained by the methods here presented show better performance than classical representation of text; we cannot conclude anything about the proposed method, rather the encouragement to run experiments using a variety of queries, and, by the fact of being unsupervised, applying the method on diverse NLP tasks, such as Text Categorization and Text Clustering, confirming the scope and limitations of the present proposal.

References

1. C. E. Shannon, *The Bell System Technical Journal* 27, 379 (1948)
2. Marcelo A. Montemurro, Entropic Analysis of the role of the words in literaty texts. arXiv:cond-mat/0109218 v1 12 sep 2001.
3. Salton, G., Wong, A. & Yang, C.: A Vector Space Model for Automatic Indexing, *Communications of the ACM*, 18(11) pp 613-620, 1975.
4. Moyotl, E. & Jiménez, H.: An Analysis on Frecuency of Terms for Text Categorization, *Proc. of SEPLN-04, XX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*, pp 141-146, 2004.
5. Moyotl, E. & Jiménez, H.: Enhancement of DPT Feature Selection Method for Text Categorization, *Proc. of CICLing-2005, Sixth International Conference on Intelligent Text Processing and Computational Linguistics*, pp 706-709, 2005.
6. Héctor Jiménez-Salazar, David Pinto & Paolo Rosso: El uso del punto de transición en la selección de términos índice para agrupamiento de textos cortos, *XXI Congreso de la SEPLN*, en prensa, 2005.
7. Sebastiani, F.: Machine Learning in Automated Text Categorization, *ACM Computing Surveys*, Vol. 34(1), pp 1-47, 2002.
8. T. Liu, S. Liu, Z. Chen & W. Ma: An Evaluation on Feature Selection for Text Clustering, *ICML*, p. 488-495, 2003.
9. van Rijsbergen, C.J.: *Information Retrieval.* London, Butterworths, 1999.
10. F. Rojas, D. Pinto, H. Jiménez, Resultados Preliminares de una Ponderacion de Términos para la Recuperación de Información, sent to *Workshop on Tecnologías del Lenguaje Humano, ENC 05*, 2005.

11. 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.
12. Zipf, G.K.: *Human Behaviour and the Principle of Least Effort*, Addison-Wesley, (1949).