A Competitive Term Selection Method for Information Retrieval[*]

Franco Rojas López[1], Héctor Jiménez-Salazar[1], and David Pinto[1,2]

[1] Faculty of Computer Science,
BUAP, Puebla, 72570
Ciudad Universitaria, Mexico
{frl99, hgimenezs}@gmail.com

[2] Department of Information Systems and Computation,
UPV, Valencia 46022,
Camino de Vera s/n, Spain
davideduardopinto@gmail.com

Abstract. Term selection process is a very necessary component for most natural language processing tasks. Although different unsupervised techniques have been proposed, the best results are obtained with a high computational cost, for instance, those based on the use of entropy. The aim of this paper is to propose an unsupervised term selection technique based on the use of a bigram-enriched version of the transition point. Our approach reduces the corpus vocabulary size by using the transition point technique and, thereafter, it expands the reduced corpus with bigrams obtained from the same corpus, i.e., without external knowledge sources. This approach provides a considerable dimensionality reduction of the TREC-5 collection and, also has shown to improve precision for some entropy-based methods.

1 Introduction

Vector Space Model (VSM) was proposed by Salton [15] in the 1970's. This model states a simple way to represent documents of a collection by using vectors with weights according to the terms appearing in each document. Even though several other approaches have been tried, such as representative pairs [10] or documents tokens, terms vector representation remains a topic of interest. Main attraction stills on VSM because it provides a framework for several applications of Natural Language Processing (NLP) such as text categorization, clustering, summarization and so on. Particularly, in Information Retrieval (IR), several experiments have shown a sucessful use of VSM. In this model, each document is represented as a vector whose entries are weights of terms of the vocabulary obtained from a text collection. Specifically, given a text collection $\{D_1, \ldots, D_M\}$ with vocabulary $V = \{w_1, \ldots, w_n\}$, the vector $\overrightarrow{D_i}$ of dimension n, corresponding

[*] This work has been partially supported by the BUAP-701 PROMEP/103.5/05/1536 grant and FCC-VIEP-BUAP.

A. Gelbukh (Ed.): CICLing 2007, LNCS 4394, pp. 468–475, 2007.

to the document D_i, has entries d_{ij} representing the weight of the term w_j in D_i:

$$d_{ij} = tf_{ij} \cdot idf_j, \tag{1}$$

where tf_{ij} is the frequency of term w_j in document D_i, $idf_j = \log_2(\frac{2M}{df_j})$, and df_j is the number of documents in which w_j appears. In collections of hundreds of documents, the dimension of the vector space can be of tens of thousands. Therefore, a key element in text representation consists basically of the adequate selection of important terms, i.e., those that do not affect the retrieval, clustering, or categorization process, implicit in the application. Besides, a reduction of the vocabulary dimensionality without affecting the effectiveness is expected. It is important, from the reason just explained, to explore new mechanisms to represent texts, with the minimal number of terms and, the maximum tradeoff of precision and recall.

In [17] , for instance, R. Urbizagástegui used the *Transition Point* (TP) to show its usefulness in text indexing. TP is a frequency value that splits the vocabulary of a text into two sets of terms (low and high frequency). This technique is based on the Zipf Law of Word Ocurrences [19] and also on the refined studies of Booth [2]. These studies are meant to demonstrate that mid-frequency terms are closely related to the conceptual content of a document. Therefore, it is possible to hypothesize that terms closer to TP can be used as index terms of a document. A typical formula used to obtain this value is: $TP = \frac{-1+\sqrt{8*I_1+1}}{2}$, where I_1 represents the number of words with frequency equal to 1 (see [17]). Alternatively, TP can be found as the first frequency that is not repeated from a non-increasing frequency-sorted vocabulary; since a feature of low frequencies is that they tend to repeat [2]. Particularly, in the experiments we have carried out, we used this approach. Additionaly, the Transition Point technique has shown a good performance in term selection for text categorization [9] and clustering [12].

TP is derived from items underlying in the signifier form, because of its intrinsic property of statistical regularity in the texts. By using ontologies, dictionaries and other lexical resources, it is possible to affect the signifier substance [4]. Thus, TP can be used to affect form and substance by using terms related to it. However, the use of some lexical resources, such as WordNet, would not be factible because of its wide domain, carrying out to discard several terms belonging to the specific application domain. Regarding the usefulness of the later remark, the set of terms selected by TP may be increased with related terms, namely TP enriched approach [14].

On the other hand, M. A. Montemurro [6] did a statistical analysis of some set of words without knowledge of the grammatical structure of the documents analized. He used the entropy concept for sorting sets of words, based on the role that these words play in a set of documents from the literature domain. Entropy measures the amount of information contained in a system [3]. So, given a system S with $s_1, \ldots, s_n$ states, the entropy of S is $H(S) = -\sum_i p_i \log(p_i)$, being $p_i = \Pr(s_i)$. This means that a deterministic system lacks of information if $H(S) = 0$, since $p_i = 1$ (the same argument is valid if $p_i = 0$). On the other hand, a system whose states have the same probability ($p_i = 1/n$) will have the maximum of information $H(S) = \log n$. We must not confuse the information that a system has with the information that can be extracted from it; in other words, the less information we have from a system, the bigger amount of information the system will have; the information of a system is a measurement of our ignorance. In a text collection we can consider a) the words, w, that have high probability to appear in all the documents ($\Pr(w) \approx 1$); b) words that are not uniformly distributed in the collection of texts, i.e., those which are concentrated in some document; and, c) 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 and may be used to represent the text.

This work explores an alternative to the classic representation based on the VSM for IR. Entropy was used in [5] for IR processes on a small text collection, but results were not conclusive. TP has been also used in this context [13], obtaining good results: reducing dimensionality and outperforming classical representation. In [14] an enrichment of the term selected by TP for cross-lingual information retrieval was presented, but results were not indicative of better performance due to the noisy terms in multilingual collections. Our contribution here consists in clarify the uselfulness of each of these methods by using the TREC-5 standard collection as a common reference. Besides, we have tried to enhance the obtained results by providing a combination of such approaches.

Following sections present the term selection and weighting schemata, experiments done by using the TREC-5 collection, results, and a discussion with conclusions.

2 Term Selection and Weighting

In this section we describe in detail each dimensionality reduction method explored in our experiments. The description of the method is presented first and, thereafter, an explanation of the representation schema is given.

2.1 Entropy

Determination of a set of words that characterize a set of documents given, is the focus of our work. Given a set of documents $D = \{D_1, D_2, ..., D_M\}$, and N_i the number of words in the document D_i, the relative frequency of the word w_j in D_i is defined as follows:

$$f_{ij} = \frac{tf_{ij}}{N_i^{tf_{ij}}}, \tag{2}$$

and

$$p_{ij} = \frac{f_{ij}}{\sum_{j=1}^{m} f_{ij}} \tag{3}$$

is the probability of the word w_j be in D_i. Thus, entropy of w_j can be calculated as:

$$H(w_j) = -\sum_{i=1}^{M} p_{ij} \log p_{ij}. \tag{4}$$

The representation of a document D_i is given by the VSM, whenever terms have high entropy. Let H_{max} be the maximum value of entropy on all the terms, $H_{max} = \max_j H(w_j)$, the representation based on entropy of D_i is

$$H_i = [w_j \in D_i | H(w_j) > H_{max} \cdot u], \tag{5}$$

where u is a threshold which defines the level of high entropy. In our experiments we have set $u = 0.5$.

2.2 Transition Point

Given a document D_i and its vocabulary $V_i = \{(w_j, tf_i(w_j)) | w_j \in D_i\}$, where $tf_i(w_j) = tf_{ij}$, let TP_i be the transition point of D_i. A set of important terms which will represent the document D_i may be calculated as follows:

$$R_i = \{w_j | ((w_j, tf_{ij}) \in V_i), (TP_i \cdot (1-u) \leq tf_{ij} \leq TP_i \cdot (1+u))\}, \tag{6}$$

where u is a value in $[0, 1]$. Some experiments presented in [13] have shown that $u = 0.4$ is a good value for this threshold.

For the representation schema, we consider that the important terms are those whose frequencies are closer to the TP. Therefore, a term with frequency very "close" to TP will get a high weight, and those "far" to TP will get a weight close to zero. For each term $w_j \in R_i$, its weight, given by Equation (1), is modified according to the distance between its frequency and the transition point, obtaining a new value for its "term frequency" (see Equation (7)).

$$tf'_{ij} = \|R_i\| - |TP_i - tf_{ij}| \tag{7}$$

2.3 Term Enrichment

Although TP certainly reduces space dimensionality by increasing precision, it obtains a low recall. Due to this fact we are proposing to enrich the terms selected by this method with those which have similar characteristics, by using a co-ocurrence bigrams-based formula. Formally, given a document D_i made up of only those terms selected by using the TP approach (R_i), the new important terms for D_i will be obtained as follows:

$$R'_i = R_i \cup \{w' | (w_j \in R_i), (v = w'w_j \text{ or } v = w_j w'), (v \in D_i), (tf_i(v) > 1)\}. \tag{8}$$

That is, we only used a window of size one around each term of R_i, and a minimum frequency of two for each bigram was required as condition to include new terms.

As R_i, weighting for enriched terms follows Equations (1) and (7). Terms $\{w' | w' \in R'_i \wedge w' \notin R_i\}$ will use directly the Equation (1).

2.4 Union of Entropy and TP

This representation takes advantage of the benefit of both approaches, TP and entropy. TP represents text independently, whereas entropy obtains better discriminant terms, therefore, we have selected those terms that satisfy either of these two conditions. The representation of a document D_i is then given by:

$$H'_i = H_i \cup R_i \tag{9}$$

In this approach two weighting criteria were adopted for the representation schema. Terms provided by H_i (Equation (5)) and R_i (Equation (6)) are weighted by Equations (1) and (7), respectively. The procedure for determining H'_i was to add, to the set R_i, all the terms that satisfy H_i. Thereafter, terms $w_j \in H_i \cap R_i$ are weighted by Equation (7).

3 Experiments

Three experiments were performed in this work, first we determined the performance of the entropy schema, H; then we used an enrichment of TP, TP'; finally the union of the both TP and H was done. The dataset and the results obtained are described in the following subsections.

3.1 Data Description

We have used the TREC Spanish Corpora, produced by the Linguistic Data Consortium (LDC)[1], for our experiments. Particularly, one corpus of the TREC-5 collection which consists of 50 topics (queries) and 57,868 documents in Spanish language from the "El Norte" mexican newspaper was selected. The average size of vocabulary of each document is 191.94 terms. Each of the topics has associated its set of relevant documents. On average, the number of relevant documents per topic is 139.36. The documents, queries and relevance judgements (qrels) used in the experiments were all taken from TREC-5.

3.2 Results

Figure 1 shows an interpolation of the average precision at different standard recall levels [1]. Two of these curves were previously presented: the classical VSM and TP [13]; therefore, we are using them as a reference for our own results. The three remained curves were obtained by using the representation schemas presented at section 2: H, terms obtained by using entropy; TP', enriched terms by bigrams; and H+TP, the union of H and TP.

The TP-based method shows a better performance than the classical VSM by using low computational resources. On the other hand, the entropy-based method has a very good performance but with a higher computational cost. The

[1] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2000T51

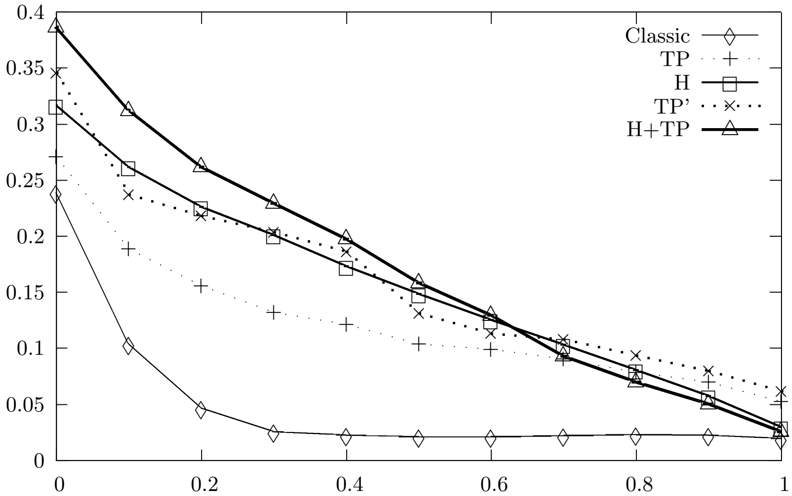

Fig. 1. Performance of term selection using entropy (H) and transition point (TP)

TP approach, enriched with bigrams, obtained a similar performance than the entropy. Finally, the union of entropy and TP curve may indicate that weighting procedure (by using both Equation (1) and (7)) did not give an adequated importance to terms, since precision diminished after of 0.6 recall level.

The vocabulary size for each method is shown in Table 1. Entropy did the highest reduction (it just uses 3.3% from original term space). TP enrichment obtained the highest vocabulary size, except for VSM, but its results are competitive with the entropy method and, with so much light computation consumption than entropy does.

Table 1. Term reduction methods and the vocabulary size obtained for TREC-5

Method name	**Vocabulary size**	**Percentage of reduction**
VSM	235,808	0.00
TP	28,111	88.08
H	7,870	96.70
TP'	36,442	84.55
H+TP	29,117	87.66

4 Discussion

Text representation, by using the VSM, implies the problem of selecting the minimal set of index terms and, thereafter, the calculation of their weights. Despite the fact that VSM and the classical weighting have several decades of existence, nowadays they are in essence being used in a diversity of NLP tasks; e.g., text categorization, text clustering, and summarization. It is well known the empirical fact that by using all terms of a text commonly produces a noisy effect in the representation [16]. Besides, the high dimensionality of the term space has led to an index term analysis. For instance, Salton et al. [15] proposed a measurement of discrimination for index terms, i.e., terms defining vectors in the space that better discerned what documents answer a particular query. They concluded that, given a collection of M documents, the "more discriminant" terms have a frequency in the range $[\frac{M}{100}, \frac{M}{10}]$. A similar experiment was carried out in [8], showing that term frequencies around TP overlap the above range. This result suggested to analyze the discriminant value of terms in a neighborhood of TP [7]. TP have a good performance due to the use of mid-frequencies terms, however, many important terms in a document have a frequency far from TP. In this work, such terms were included in the document representation through a very simple procedure (bigrams), outperforming the TP method.

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. We have seen, that the H method had very good performance, but the computation of the entropy for each term of the collection has a very high computational cost.

Conjecture, formuled in [5], established that *terms with balanced use through the texts collection is a characteristic related with the Zipf's Law [19]: minimum effort to write a text entails a moderate use on some words, which is revealed by entropy.* When dealing with many texts, it may be interpreted as *preserving the regularity of occurrence of such words, as if they were relevant because of their role in the texts as pivots.* In fact, from the experiments carried out in this work, it was shown that TP enrichment performed in similar manner as the entropy method. Besides, in the experiment which joins entropy and TP, the most of the terms selected by entropy were also selected by TP (87.21%). Furthermore, just the 0.78% of the H-terms do not belong to the set provided by TP'. This fact is confirmed by comparing the TP' precision-recall curve with the H curve (Fig. 1). However, there is a high amount of TP'-terms (6,711) that do not belong to neither, the TP-term nor the H-term set. This set of terms introduces an unstable behaviour at TP' curve: good terms and noisy terms distribute relevant and non relevant texts throughout of all retrieved results.

Up to now, we have tested the methods proposed in only one collection, but further investigations should consider other datasets in order to see if the given conclusions carry out in those as well.

A clear advantage of the methods presented in this paper are their unsupervised nature and language independence which makes them suitable for their use in a wide variety of NLP tasks.

References

1. Baeza-Yates, R. & Ribeiro, N.: *Modern Information Retrieval*, Addison Wesley, 1999.
2. Booth A.: A law of occurrence of words of low frequency, *Information and Control*, 10 (4), pp. 383–396, 1967.
3. C. E. Shannon, *The Bell System Technical Journal* 27, 379 (1948).
4. Gelbukh, A.; Sidorov, G. & Guzman-Arenas, A.: Use of a weighted topic hierarchy for text retrieval and classification, *LNCS 1692*, pp 130-135, 1999.
5. Jiménez-Salazar, H.; Castro, M.; Rojas, F.; Miñón, E.; Pinto, D. & F. Carcedo: Unsupervised Term Selection using Entropy, *Research on Computing Science* 14, ISSN 1665-9899, pp. 163–172, México, 2005.
6. Montemurro, M.A. & Zanette D. H.: Entropic Analysis of the role of the words in literaty texts, CoRR, arXiv:cond-mat/0109218, v1 12, sep 2001.
7. Moyotl, E.: *DPT: un método de selección de términos para categorización de textos*, Master in Computer Science Thesis, FCC-BUAP, 2005 (*In spanish*).
8. E. Moyotl & H. Jiménez: An Analysis on Frequency of Terms for Text Categorization, *Procesamiento del Lenguaje Natural*, ISSN 1135-9948, pp 141-146, España.
9. Moyotl, E. & Jiménez, H.: Enhancement of DPT Feature Selection Method for Text Categorization, LNCS 3406, pp. 706–709, 2005.
10. Pérez-Carballo, J. & Strzalkowski, T.: Natural Language Information Retrieval: progress report, *Information Processing and Management* v.36(1), Elsevier, pp. 155–178, 2000.
11. Pinto, D.; Jiménez-Salazar, H.; Rosso P. & Sanchis, E.: BUAP-UPV TPIRS: A System for Document Indexing Reduction at WebCLEF. Accessing Multilingual Information Repositories, CLEF 2005, *LNCS 4022*, 2006.
12. Pinto D.; Jiménez-Salazar, H. & Paolo Rosso: Clustering Abstracts of Scientific Texts using the Transition Point Technique, *LNCS 3878*, pp. 536–546, 2006.
13. Rojas, F.; Jiménez, H.; Pinto, D. & Aurelio López: Dimensionality reduction for Information Retrieval, *Research on Computing Science*, Vol 20, pp 107–112 2006.
14. Rojas, F.; Jiménez, H. & Pinto, D.: Text Reduction-Enrichment at WebCLEF, In *Proceedings of CLEF 2006*, pp. 53, 2006.
15. Salton, G., Wong, A. & Yang, C.: A Vector Space Model for Automatic Indexing, *Communications of the ACM*, 18(11) pp. 613–620, 1975.
16. Sebastiani, F.: Machine Learning in Automated Text Categorization, *ACM Computing Surveys*, 34(1), pp. 1–47, 2002.
17. Urbizagástegui, A.R.: Las Posibilidades de la Ley de Zipf en la Indización Automática, `http://www.geocities.com/ResearchTriangle /2851/RUBEN2.htm`, 1999 (*In spanish*).
18. Yang, Y., Pedersen, P.: A Comparative Study on Feature Selection in Text Categorization, *Proc. of ICML-97, 14th Int. Conf. on Machine Learning*, pp. 412–420, 1997.
19. Zipf, G.K.: *Human Behaviour and the Principle of Least Effort*, Addison-Wesley, 1949.