Dimensionality Reduction for Information Retrieval

Franco Rojas López[1], Héctor Jiménez-Salazar[1]
David Pinto[1,2], A. López-López[3]

[1]Facultad de Ciencias de la Computación,
Benemérita Universidad Autónoma de Puebla,
14 sur y Av. San Claudio, Ciudad Universitaria, Edif. 135
Puebla, Pue., México, 72570
{frl99, hgimenezs}@gmail.com
[2]Departamento de Sistemas Informáticos y Computacionales,
Universidad Politécnica de Valencia,
Camino de Vera s/n,
Valencia, España, 46006
davideduardopinto@gmail.com
[3]Laboratorio de Tecnologías del Lenguaje
Instituto Nacional de Astrofísica Óptica y Electrónica
allopez@inaoep.mx

Abstract. This work presents a variation of the traditional text representation based on the vector space model, used in Informational Retrieval. In particular, a representation is proposed, intended to select terms for indexing and weighting them according to their importance. These two tasks are performed taking into account the terms with medium frequency, that have shown an advantage to reveal keywords. The results of experiments using an information retrieval system on the TREC-5 collection show that the proposed representation outperforms term weighting using $tf \cdot idf$, reducing simultaneously the dimensionality of terms to less than 12%.

1 Introduction

Vector Space Model (VSM) was proposed by Salton [10] in the 1970's. This model states a simple way to represent documents of a collection; using vectors with weights according to the terms appearing in each document. Even though several other approaches have been tried, such as the use of representative pairs [7] or the tokens of documents, vector representation based on terms remains a topic of interest, since some other applications of Natural Language Processing (NLP) use it, for instance, text categorization, clustering, summarization and so on.

In Information Retrieval (IR), a commonly used representation is the vector space model. In this model, each document is represented as a vector whose entries are terms of the vocabulary obtained from the text collection. Specifically,

given a text collection $\{D_1, \ldots, D_M\}$ with vocabulary $V = \{t_1, \ldots, t_n\}$, the vector $\overrightarrow{D_i}$ of dimension n, corresponding to document D_i, has entries d_{ij}, where the value of an entry d_{ij} is the weight of term t_j in D_i:

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

where tf_{ij} is the frequency of term t_j in document D_i, $idf_j = \log_2(\frac{2M}{df_j})$, and df_j is the number of documents using term t_j. In collections of hundreds of documents, the dimension of the vector space can be of tens of thousands.

A key element in text representation is basically the adequate election of important terms, i.e. those that do not affect the process of retrieval, clustering, and categorization, implicit in the application. Besides, they have to reduce the dimensionality without affecting the effectiveness. It is important, from the reason just explained, to explore new mechanisms to represent text, based on terms appearing in the text. There are several methods to select terms or keywords from a text, many of them affordable in terms of efficiency but not very effective. R. Urbizagástegui [12] used the *Transition Point* (TP) to show its usefulness in text indexing. Moreover, the transition point has shown to work properly in term selection for text categorization [4] [5] [6]. TP is the frequency of a term that divides a text vocabulary in terms of high and low frequency. This means that terms close to the TP, of both high and low frequency, can be used as keywords that represent the text content. A formula to calculate TP is:

$$TP = \frac{\sqrt{1 + 8 \cdot I_1} - 1}{2}, \tag{2}$$

where I_1 represents the number of words having frequency 1. Alternatively, TP can be found as the lowest frequency, from the highest, that does not repeat, since a feature of low frequencies is that they tend to repeat.

This work explores an alternative to the classic representation based on the vector space model for IR. Basically, the proposed representation is the result of doing a term selection, oriented to index the document collection and, in addition, a weighting scheme according to the term importance. Both tasks are based on terms allegedly having a high semantic content, and their frequencies are within a neighborhood of the transition point.

Following sections present the term weighting scheme, experiments done using TREC5 collection, results, and a discussion with conclusions.

2 Term Selection and Weighting

The central idea behind the weighting scheme proposed here is that important terms are those whose frequencies are close to the TP. Accordingly, term with frequency very "close" to TP get a high weight, and those "far" from TP get a weight close to zero. To determine the nearness to TP, we proceed empirically: selecting terms with frequency within a neighborhood of TP; where each neighborhood was defined by a threshold u.

Given a document D_i, we build its vocabulary from the frequency, fr, of each word: $V_i = \{(x, y)|x \in D_i, y = fr(x)\}$. From the vocabulary, we can calculate $I_1 = \#\{(x, y) \in V_i|y = 1\}$ for D_i. So, using equation 2, TP of D_i is determined (denoted as PT_i), as well as a neighborhood of important terms selected to represent document D_i:

$$R_i = \{x|(x, y) \in V_i, TP_i \cdot (1 - u) \leq y \leq TP_i \cdot (1 + u)\}, \tag{3}$$

where u is a value in $[0, 1]$.

The important terms of document D_i are weighted in the following way. For each term $t_{ij} \in R_i$, its weight, given by equation 1, is altered according to the distance between its frequency and the transition point:

$$tf'_{ij} = \#R_i - |TP_i - tf_{ij}|. \tag{4}$$

3 Data Description

TREC-5 collection consists of 57,868 documents in Spanish, and 50 topics (queries). 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 used in the experiments were taken from TREC-5.

4 Experiments

Two experiments were performed, the first aimed to determine the size of the neighborhood u (eq. 3) and, the second was oriented to measure the effectiveness of the proposed scheme on the whole collection TREC-5. In these experiments, we applied standard measures; *i. e.*, precision (P), recall (R), and F_1 measure [13] defined as follow.

$$P = \frac{\#\text{relevant docs. obtained by the system}}{\#\text{docs. obtained by the system}}, \tag{5}$$

$$R = \frac{\#\text{relevant docs. obtained by the system}}{\#\text{ relevant documents}}, \tag{6}$$

$$F_1 = \frac{2 \cdot P \cdot R}{P + R}. \tag{7}$$

4.1 Neighborhood Determination

Two subsets of TREC-5 were extracted, S_1 and S_2 sub-collections, with 933 and 817 documents, respectively. Each one contains documents relevant to two topics, in addition to non relevant documents selected randomly, in a double rate to relevant documents. Several threshold values were tested, whose results are displayed in Figure 1.

Fig. 1. Values of F_1 using three thresholds in two sub-collections of TREC-5.

Sub-collection	u 0.3	0.4	0.5
S_1	0.34	0.37	0.39
S_2	0.28	0.34	0.38

Other values of u led to F_1 values less or equal to those showed in the table of Figure 1. We picked $u = 0.4$, even though this does not produce the maximum F_1, but allows to determine a lower bound of the performance of the proposed term selection.

4.2 Term Selection and Weighting Performance

Document indexing was done using formulas 3 and 4, in addition to classic term weighting (eq. 1) in the whole TREC-5 collection, and submitting the 50 queries. Retrieved documents were sorted according to their assessed similarity to the query (*ranking*). For a vector query $\overrightarrow{q}$, and a document $\overrightarrow{D_i}$, its similarity was calculated using the cosine formula. Finally, to assess the effectiveness, we calculate average precision at standard recall levels, as shown in (fig. 3) [1] for classical and proposed (using TP) weighting.

Figure 2 summarizes the number of terms in the vocabulary for the whole collection, average number of terms per document, and the percentage of terms generated by the proposed indexing with respect to those produced by the classical representation.

Fig. 2. Vocabulary for the Two Representations.

Total/partial	**Classic**	**TP**	**%**
TREC-5	235,808	28,111	**11.92**
Average × doc.	191.94	6.86	**3.57**

5 Discussion

G. P. Luhn based on the argument that high frequency terms are very general and can lead to low precision, while those of low frequency result in low recall, proposed with insight what has been confirmed empirically. As stated above, the problem of determining adequate words to index documents is of interest in several tasks.

The use of transition points for the problem of term selection has shown effectiveness in some contexts [6] [9] [8].

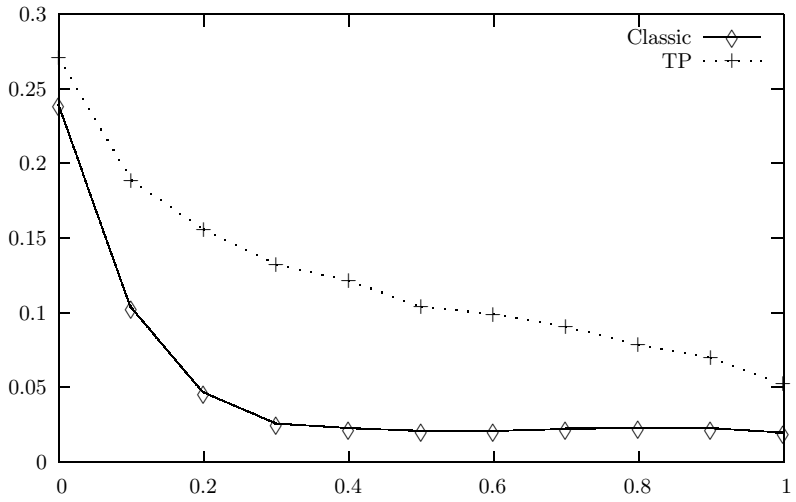

Fig. 3. Average Precision at standard Recall levels, using classical and proposed weighting.

The base on which lies the identification of medium frequencies has been taken from the formulae presented in 1967 by Booth [2], intended to determine a value that it was not high or low. From this formulae, the TP began to be used in the identification of keywords in a text [12]. In the present work, TP was used in a particular way: term weighting considering a neighborhood of frequencies around TP. The formula that determines such neighborhood (eq. 3) comes from the characteristics assumed for the TP [3]. It is not the same case for the weighting equation (4) which modifies the classical weighting (eq. 1). In the former, it is implicit the fact of repeating the terms that occur in the neighborhood as many times as their complementary distance to TP. That is the rationale of the replacement of tf_{ij} in eq. 1 by the proposed weighting (eq. 4). This repetition is a simple way to reinforce the importance of a term whose frequency is in the TP neighborhood.

The text representation problem, using the VSM, implies the selection of index terms and their weighting. 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 a well known empirical fact that using all terms of a text commonly produces a noisy effect in the representation [11]. The high dimensionality of the term space has led to a index term analysis. For instance, Salton et al. [10] 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 $[\frac{M}{100}, \frac{M}{10}]$. This result suggests to analyze the discriminant value of terms in a neighborhood of TP.

The diversity of proposals on feature selection are conceived into supervised and unsupervised methods. An advantage of the method here presented is its unsupervised nature, so it is possible to use it in a wide variety of NLP tasks.

The results obtained for TREC-5 encourage to confront TP with other proposed representations, and its application in different collections to validate the observed effectiveness. Moreover, we identify the need to establish precisely the advantages of this representation on some other task of NLP.

References

1. Baeza-Yates, R.: *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. Ernesto Miñón; David Pinto & Héctor Jiménez-Salazar: Análisis de una representación de textos mediante su extracto, *Avances en la Ciencia de la Computación*, pp. 107-111, 2005.
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.: Experiments in Text Categorization Using Term Selection by Distance to Transition Point, *Proc. of CIC-04, XIII Congreso Internacional de Computación,* pp. 139-145, 2004.
6. 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.
7. Pérez-Carballo, José & Strzalkowski, Tomek: Natural Language Information Retrieval: progress report, *Information Processing and Management* v.36(1), Elsevier, pp. 155-178, 2000.
8. Pinto D.; Jiménez-Salazar, H. & Paolo Rosso: Clustering Abstracts of Scientific Texts using the Transition Point Technique, *Lecture Notes in Computer Science*, Vol. 3878, pp. 536-546, 2006.
9. David Pinto, Héctor Jiménez-Salazar, Paolo Rosso & Emilio Sanchis: BUAP-UPV TPIRS: A System for Document Indexing Reduction at WebCLEF. Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers. Carol Peters, Fredric C. Gey, Julio Gonzalo, Gareth J.F. Jones, Michael Kluck, Bernardo Magnini, Henning Müller, Maarten de Rijke (Eds.), *Lecture Notes in Computer Science* Vol. 4022 (forthcoming).
10. Salton, G., Wong, A. & Yang, C.: A Vector Space Model for Automatic Indexing, *Communications of the ACM*, 18(11) pp. 613-620, 1975.
11. Sebastiani, F.: Machine Learning in Automated Text Categorization, *ACM Computing Surveys*, 34(1), pp. 1-47, 2002.
12. Urbizagástegui, A.R.: Las Posibilidades de la Ley de Zipf en la Indización Automática, `http://www.geocities.com/ResearchTriangle /2851/RUBEN2.htm`, 1999.
13. van Rijsbergen, C.J.: *Information Retrieval*. London, Butterworths, 1999.