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,

© *A. Gelbukh, S. Torres, I. López (Eds.)*
Advances in Artificial Intelligence
Research in Computing Science 20, 2006, pp. 107-112

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-		u	
collection	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].

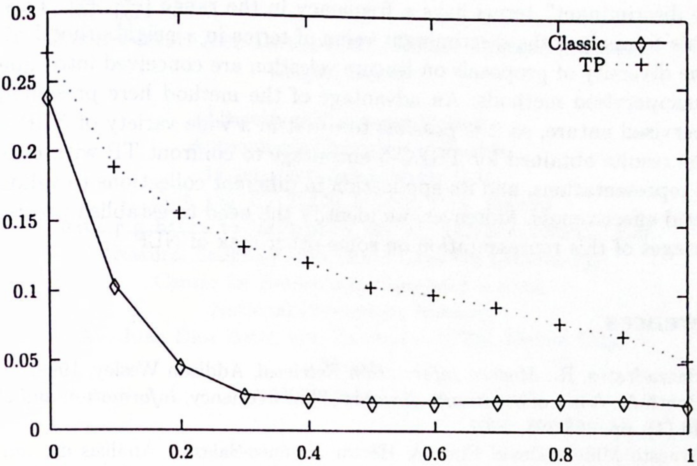

Fig. 3. Average Precision at standard Recall levels, using classical and proposed weighting.

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 where tf_{ij} is frequency of the term j in the document i, while idf_j refers to the number of documents that use the term j, df_j. It is calculated as

$$idf_j = \log_2(\frac{2 \cdot M}{df_j}).$$

This can be explained as following: the term with high frequency in the text has large weight only if it occurs in a small number of documents. If a high frequency term occurs in many documents, then it does not convey real information because it does not allow for distinguishing between the documents.

The idea that we present in the paper is based on the fact that the terms with medium frequency should have the major weight, and while the terms are more distant from the range of medium frequencies, the less their weight is. There are descriptions of experiments that demonstrate that the usage of variations of weights (Eq. 1) using threshold and transition point (TP) is promising [7] [4]. In this paper, we propose a formula for determining medium frequencies without a threshold. It is shown that this formula allows for obtaining equal or better performance than the TP.

The work has the following structure: first we describe how to determine the range of medium frequencies of a text, then we present the suggested weighting formula, after this, the extraction of the representative sentences from a text is described, and finally, the results of the experiments are discussed.

2 Transition Point Range

One of the important tasks of text representation is the selection of a subset of terms that are good representation of a text and permit operations of categorization, clustering, searching, etc. using the selected subset instead of a whole document. There are various methods for selection of indexing terms or key words, for example, Urbizagástegui [2] used the transition point for showing the usefulness of text indexation.

Transition point is a frequency of a term that divides text vocabulary into terms of low and high frequencies. The terms that are useful for text representation are situated around the TP, because it is supposed that they have high semantic content. The formula for the calculation of TP is as follows:

$$TP = \frac{\sqrt{1 + 8 * I_1} - 1}{2}, \qquad (2)$$

where I_1 represents the number of terms with frequency 1. This empiric formula seeks the identification of a frequency that is neither low, nor high. Usually, many terms correspond to low frequencies; say, more that 50% of terms in an average text have frequency 1, etc. This formula excludes them from the consideration explicitly. The calculated frequency (TP) is the lowest of the high frequencies. The alternative calculation of TP is seeking the lowest frequency that is not repeated, i.e., the first frequency that corresponds to exactly one term. It is justified by the fact that several terms usually correspond to values of low frequencies.

In this paper, we suggest to use two transition points basing on the idea of repetition of frequencies. The first TP is the lowest of the high frequencies that is repeated, TPa, i.e., we start form the highest frequency and go downwards until we find the first repetition. The second TP is the highest of the low frequencies that is not repeated, TPb, i.e., we start from the lowest frequency and go upwards until we find the first term with unique (non-repeated) frequency. Thus, we define the range of medium frequencies, namely, *transition range*, $[TPb, TPa]$.

In the following sections, we describe the application of the transition range to information retrieval and extraction of the representative sentences tasks.

3 Weighting of Terms

As we mentioned before, the documents can be represented by the weighted terms. In [7] a scheme of term weighting which takes into account the TP is presented. The method proposed there is different from Eq. 1, namely

$$d_{ij} = IDPT_{ij} \times DPTC_i, \tag{3}$$

where $IDPT_{ij} = 1/|TP_j - tf_{ji}|$ is the inverse distance of the term i to the TP of the document j, and $DPTC_i = |TP - fr_i|$, is the distance between the term i and TP, calculated for the whole collection.

For our experiment, the definition of $IDPT_{ij}$ is given by:

$$IDPT_{ij} = \begin{cases} 1 & \text{if } tf_{ji} \in [TPb_j, TPa_j], \\ 1/(tf_{ji} - TPa_j) & \text{if } TPa_j < tf_{ji}, \\ 1/(TPb_j - tf_{ji}) & \text{if } TPb_j > tf_{ji}. \end{cases} \tag{4}$$

where $[PTb_j, PTa_j]$ is the transition range of the document j. $DPTC_i$ also is adapted to the two frequencies of transition that are global now:

$$DPTC_i = \begin{cases} 1 & \text{if } fr_i \in [TPb, TPa], \\ fr_i - TPa & \text{if } TPa < fr_i, \\ TPb - fr_i & \text{if } TPb > fr_i. \end{cases} \tag{5}$$

Here $[TPa, TPb]$ constitutes the transition range of the whole collection and fr_i is the frequency of term i in the collection.

4 Representative Sentences in Texts

Let us consider the task of selection of the most "representative" sentences of a text. We base on the work [8], where the terms near TP are considered for assigning scores to sentences and generate an extract composed by three sentences with major scores. The proposed approach is as follows:

1. Preprocessing. Document splitting into sentences is performed, taking into account abbreviations, etc. The words from the stop list are eliminated from the sentences. These are words like prepositions, articles, etc.
2. Vocabulary extraction. All terms are extracted and their frequencies are calculated.
3. Transition range. The transition range is calculated according to the procedure described above. The "virtual paragraph" (VP) is generated, i.e., the paragraph, to which all terms that belong to the transition range are added.
4. Assignment of scores to sentences. Each sentence is assigned a score according to its similarity to the VP.
5. Extraction of representative sentences. Three sentences with major scores according to their similarity to the VP are taken.

The extract quality is verified by its comparison with the complete document. One of the ways of doing this is the usage of the extract instead of the full text in certain tasks like Information Retrieval. If the IR system performs in the same way, then the quality of the extract is good. For our experiments, we used Jaccard's formula to calculate the similarity between the query and each document in the collection, as in [8].

$$sim(D,q) = \frac{\#(D \cap q)}{\#(D \cup q)}.$$

5 Obtained Results

We conducted experiments with term weights assignment based on the transition range and detection of representative sentences. Several subcollections of TREC-5 were used allowing comparison of results with previous works. Further we describe subcollections and then the obtained results.

5.1 Data Description

Collection TREC-5 is a text collection of more than 50,000 documents in Spanish and 50 topics (possible queries). Each topic is assigned a set of documents that correspond to it, i.e., are relevant for this topic. The TREC-5 documents, queries, and relevance criteria were used in our experiments. We defined three subcollections from the documents according to the following algorithm: for a given topic, we add all relevant documents to the subcollection, and then add twice as many non-relevant documents. Table 1 contains the number of documents in subcollections.

The subcollections were preprocessed and words from stop lists were eliminated. The queries were preprocessed as well in the same way. Besides, all letters in queries were changed to lower case. The topics are shown in Table 2.

Table 1. TREC-5 subcollections for 6 topics.

Subcollection	Topics	#	#Relevant
1	$c_1 : c_3$	1117	211 : 164
2	$c_{10} : c_{11}$	933	206 : 105
3	$c_{14} : c_{15}$	817	281 : 6

Table 2. Topics used in evaluation.

c1	mexican oposition FTA (free trade agreement)
c3	pollution mexico city
c10	mexico important country transit war antidrug
c11	water rights rivers frontier region mexico unites states
c14	monopoly oil pemex has great influence mexico
c15	dispute fishing caused capture fishing ships unites staes

5.2 Results

In Fig. 1, the results are presented for each subcollection and for each method. The Column 1 refers to the method. For three first rows, the method based on weighting was used, while for three last rows the extract generation was applied. For each subcollection, we calculated the values of precision P, recall R, and F_1 measure, see, for example, [3].

$$P = \frac{\text{\#relevant documents obtained by the system}}{\text{\#total documents obtained by the system}}, \qquad (6)$$

$$R = \frac{\text{\#relevant documents obtained by the system}}{\text{\#total relevant documents}}, \qquad (7)$$

$$F_1 = (2 \cdot P \cdot R)/(P + R). \qquad (8)$$

The methods that are referred to as TR use transition range, while those referred to as TP are based on transition point as it is explained in sections 3 and 4.

Fig. 1. Transition range.

Method	Subcol. 1			Subcol. 2			Subcol. 3		
	P	R	F_1	P	R	F_1	P	R	F_1
Classic	0.28	0.61	0.38	0.21	0.74	0.33	0.33	0.93	0.48
TR	0.24	0.17	**0.2**	0.17	0.2	**0.18**	0.34	0.78	**0.47**
TP	0.34	0.06	0.1	0.13	0.29	0.18	0.44	0.31	0.33
Full text	0.16	0.47	0.24	0.17	0.68	0.27	0.18	0.69	0.28
TR	0.16	0.22	**0.19**	0.19	0.39	**0.26**	0.18	0.48	**0.26**
TP	0.37	0.07	0.11	0.19	0.33	0.24	0.19	0.19	0.19

6 Conclusions

We presented an approach that allows for detection of the transition range, i.e., the range of terms with medium frequencies in a text. This range has the properties that correspond to the expected behavior of the terms, which are in the transition from terms with low frequency to terms with high frequency. It is supposed that terms in this range are the most representative terms of a text. The advantage of the approach is that it does not require choosing manually any thresholds. Certainly, the results are not as good as in the classic approach that uses $tf_{ij} \cdot ifd_j$ or as in the case of usage of the complete documents. We showed that the transition range gives better results than the transition point; however, this claim must be tested in a larger collection. It is necessary to take into account that transition range has similar behavior and inherits practically all numerous applications of the transition point. So, we can recommend the usage of the transition range in natural language processing applications instead of the transition point because of its extensive advantages.

References

1. Salton, G., Wong, A. & Yang, C.: A Vector Space Model for Automatic Indexing, *Communications of the ACM*, 18(11) pp 613-620, 1975.
2. Urbizagástegui, A.R.: Las Posibilidades de la Ley de Zipf en la Indización Automática, `http://www.geocities.com/ResearchTriangle /2851/RUBEN2.htm`, 1999.
3. van Rijsbergen, C.J.: *Information Retrieval.* London, Butterworths, 1999.
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. Baeza-Yates, R.: *Modern Information Retrieval*, Addison Wesley, 1999.
7. Rubí Cabrera, David Pinto, Darnes Vilariño & Héctor Jiménez: Una nueva ponderación para el modelo de espacio vectorial de recuperación de información, *Research on Computing Science* 13, pp 75-81, 2005.
8. Claudia Bueno, David Pinto & Héctor Jiménez: El párrofo virtual en la generación de extractos, *Research on Computing Science* 13, pp 83-90, 2005.