Clustering Abstracts of Scientific Texts Using the Transition Point Technique*

David Pinto[1,2], Héctor Jiménez-Salazar[1], and Paolo Rosso[2]

[1] Faculty of Computer Science, BUAP, Puebla 72570,
Ciudad Universitaria, Mexico
{davideduardopinto, hgimenezs}@gmail.com
[2] Department of Information Systems and Computation,
UPV, Valencia 46022,
Camino de Vera s/n, Spain
{dpinto, prosso}@dsic.upv.es

Abstract. Free access to scientific papers in major digital libraries and other web repositories is limited to only their abstracts. Current keyword-based techniques fail on narrow domain-oriented libraries, e.g., those containing only documents on high energy physics like those of the *hep-ex* collection of CERN. We propose a simple procedure to cluster abstracts which consists in applying the transition point technique during the term selection process. This technique uses the mid-frequency terms to index the documents due to the fact that they have a high semantic content. In the experiments we have carried out, the transition point approach has been compared with well known unsupervised term selection techniques. Transition point technique shown that it is possible to obtain a better performance than traditional methods. Moreover, we propose an approach to analyse the stability of transition point term selection method.

1 Introduction

Nowadays, very short text clustering on narrow domains has not received too much attention by the computational linguistic community. This is derived from the high challenge that this problem implies, since the obtained results are very unstable or imprecise when clustering abstracts of scientific papers, technical reports, patents, etc. But, as we can see, most digital libraries and other web-based repositories of scientific and technical information nowadays provide free access only to abstracts and not to the full texts of the documents. Moreover, some institutions, like the well known CERN[1], receive hundreds of publications every day that must be categorized on some specific domain with an unknown number of categories. This led to construct novel methods for treating this real problem.

[*] This work was partially supported by BUAP-VIEP 3/G/ING/05, R2D2 (CICYT TIC2003-07158-C04-03), ICT EU-India (ALA/95/23/2003/077-054), and Generalitat Valenciana Grant (CTESIN/2005/012).

[1] Centre Européen pour la Recherche Nucléaire.

A. Gelbukh (Ed.): CICLing 2006, LNCS 3878, pp. 536–546, 2006.

Clustering of very short texts implies to deal with very low frequencies; moreover, if this kind of texts belong to scientific papers, the difficulty increases, due to the continue use of some words like, for instance: "in this paper we present...", etc.; as a matter of fact, in [1], it is said that:

> When we deal with documents from one given domain, the situation is cardinally different. All clusters to be revealed have strong intersections of their vocabularies and the difference between them consists not in the set of index keywords but in their proportion. This causes very unstable and thus very imprecise results when one works with short documents, because of very low absolute frequency of occurrence of the keywords in the texts. Usually only 10% or 20% of the keywords from the complete keyword list occur in every document and their absolute frequency usually is 1 or 2, sometimes 3 or 4. In this situation, changing a keywords frequency by 1 can significantly change the clustering results.

Some related work was presented in [9], where simple procedures in order to improve results by an adequate selection of keywords and a better evaluation of document similarity was proposed. The authors used as corpora two collections retrieved from the Web. The first collection was composed by a set of 48 abstracts (40 Kb) from the CICLing 2002 conference; the second collection was composed by 200 abstracts (215 Kb) from the IFCS-2000[2] conference. The main goal in this paper was to stabilize results in this kind of task; a 10% of differences among different clustering methods were obtained, taking into account different broadness of the domain and combined measures.

In [1] an approach for clustering abstracts in a narrow domain using Stein's MajorClust Method for clustering both keywords and documents was presented. Here, Alexandrov et al. used the criterion introduced in [8] in order to perform the word selection process. The authors based their experiments on the first CICLing collection used by Makagonov et al. [9], and they succeeded in improving those results. In the final discussion, Alexandrov et al. stated that abstracts cannot be clustered with the same quality as full texts, though the achieved quality is adequate for many applications; moreover, they suggested that, for an open access via Internet, digital libraries should provide document images of full texts for the papers and not only abstracts.

More recently, in [6] a third experiment with the CICLing collection was carried out. In this paper, a novel method for keyword selection was proposed, claiming improving results on clustering abstracts for that collection. Jiménez-Salazar et al. based their comparisons with different mechanisms of term selection by using the evaluation of feature selection employed in the text categorization task [7].

After reviewing these works, we have observed that the feature selection process is the key of the clustering of abstracts task for narrow domains. Moreover, a bigger collection of abstracts is needed in order to confirm previously

[2] International Federation of Classification Societies; http://www.Classification-Society.org

obtained results. In the following Section we present a brief description of the Transition Point technique. The third Section describes the term selection methods used in the experiments we carried out. The fourth Section shows the data set and the performance measure formulas used. A comparison of the results obtained is presented in Section five. Finally, the conclusions of our experiments are given.

2 The Transition Point Technique

The Transition Point (TP) is a frequency value that splits the vocabulary of a document into two sets of terms (low and high frequency). This technique is based on the Zipf Law of Word Ocurrences [22] and also on the refined studies of Booth [2], as well as Urbizagástegui [20]. These studies are meant to demonstrate that terms of medium frequency are closely related to the conceptual content of a document. Therefore, it is possible to form the hypothesis that terms whose frequency is closer to TP can be used as indexes of a document. A typical formula used to obtain this value is given in equation 1:

$$TP_V = \frac{\sqrt{8 * I_1 + 1} - 1}{2}, \tag{1}$$

where I_1 represents the number of words with frequency equal to 1 in the text T [15] [20]. Alternatively, TP_V can be localized by identifying the lowest frequency (from the highest frequencies) that it is not repeated; this characteristic comes from the properties of Booth's law for low frequency words [2].

Let us consider a frequency-sorted vocabulary of a text T; i.e.,

$$V = [(t_1, f_1), ..., (t_n, f_n)],$$

with $f_i \geq f_{i-1}$, then $TP_V = f_{i-1}$, iif $f_i = f_{i+1}$. The most important words are those that obtain the closest frequency values to TP, i.e.,

$$V_{TP} = \{t_i | (t_i, f_i) \in V, U_1 \leq f_i \leq U_2\}, \tag{2}$$

where U_1 is a lower threshold obtained by a given neighbourhood value of the TP, thus, $U_1 = (1 - NTP) * TP_V$ ($NTP \in [0, 1]$). U_2 is the upper threshold and it is calculated in a similar way ($U_2 = (1 + NTP) * TP_V$).

The TP technique has been used in different areas of Natural Language Processing (NLP) like: clustering of short texts [5], categorization of texts [12] [13], keyphrases extraction [14] [19], summarization [3], and weighting models for information retrieval systems [4]. Thus, we believe that there exists enough evidence to use this technique as a term selection process.

3 Term Selection Methods

Up to now, different term selection methods have been used in the clustering task; however, as we mentioned in Section 1, clustering abstracts for a narrow domain implies the well known problem of the unidentified number of categories to be used in the clustering process. This led us to use unsupervised methods instead of supervised ones, as well as the identification of new categories, which is very usual in the domain of digital libraries. In this section we will describe the unsupervised term selection methods used in our experiments.

1. *Document Frequency (DF)*: This method assigns the value df_t to each term t, where df_t means the number of texts, in a collection, where t ocurrs. This method assumes that low frequency terms will rarely appear in other documents, and therefore, they will not have significance on the prediction of the class for this text.
2. *Term Strength (TS)*: The weight given to each term t is defined by the following equation:

   $$ts_t = Pr(t \in T_i | t \in T_j), \text{with } i \neq j,$$

   where $sim(T_i, T_j) \geq \beta$, and β is a threshold that must be tuned by reviewing the similarity matrix. A high value of ts_t means that the term t contributes to the texts T_i and T_j to be more similar than β. A more detailed description can be found in [21].
3. *Transition Point (TP)*: A higher value of weight is given to each term t, as its frequency is closer to the TP frequency, named TP_V. The following equation shows how to calculate this value:

   $$idtp(t, T) = \frac{1}{|TP_V - freq(t, T)| + 1},$$

   where $freq(t, T)$ is the frequency of the term t in the document T.

The unsupervised methods presented here are the most succesful in the clustering area. Particulary, DF is an effective and simple method, and it is known that this method obtains comparable results to the classical supervised methods like χ^2 (CHI) and Information Gain (IG) [17]. TP also has a simple calculation procedure, and as it was seen in Section 2, it can be used in different areas of NLP. The DF and TP methods have a temporal linear complexity with respect to the number of terms of the data set. On the other hand, TS is computationally more expensive than DF and TP, because it requires to calculate a similarity matrix of texts, which implies this method to be in $O(n^2)$, where n is the number of texts in the data set.

4 Clustering of Abstracts in a Narrow Domain

As was mentioned in Section 1, previous works for clustering abstracts in a narrow domain (see [9], [1], and [6]) used a very small collection (only 48 abstracts and 6 categories). Therefore, there exists a need of a bigger sized real corpus in order to verify the results obtained. Following, we introduce *hep-ex* collection, a real corpus obtained from the CERN.

4.1 Data Set

In our experiments we used two corpora based on the collection of abstracts compiled and provided to us by the University of Jaén, Spain [11], named *hep-ex*. The first corpus was built by extracting a subset of documents from the full collection. We used the full collection as a second corpus, which is composed by 2,922 abstracts from the *Physics* domain originally stored in CERN[3]. The distribution obtained for both corpora is shown in Table 1. The distribution of the categories for each corpus is better described in Table 2.

We have preprocessed these collections by eliminating stopwords and by applying the Porter stemmer. Due to their average size per abstract (aprox. 47 words), the preprocessed collections are suitable for our experiments.

Table 1. Collections (preprocessed) features

Feature	**Subset of** *hep-ex*	**Full collection** *hep-ex*
Size of the corpus (bytes)	165,349	962,802
Number of categories	7	9
Number of abstracts	500	2,922
Total number of terms	23,500	135,969
Vocabulary size (terms)	2,430	6,150
Term average per abstract	47	46.53

Table 2. Categories in corpora

	Number	**Subset of**	**Full**
Category	**of texts**	*hep-ex*	**collection**
Information Transfer and Management	1	NO	YES
Particle Physics - Phenomenology	3	YES	YES
Particle Physics - Experimental Results	2,623	YES	YES
XX	1	YES	YES
Nonlinear Systems	1	YES	YES
Accelerators and Storage Rings	18	YES	YES
Astrophysics and Astronomy	3	YES	YES
Other Fields of Physics	1	NO	YES
Detectors and Experimental Techniques	271	YES	YES

4.2 Performance Measurement

We used F-measure (commonly used in information retrieval [16]) in order to determine which method obtains the best performance. Given a set of clusters $\{G_1, \ldots, G_m\}$ and a set of classes $\{C_1, \ldots, C_n\}$, the F-measure between a cluster i and a class j is given by the following formula.

$$F_{ij} = \frac{2 \cdot P_{ij} \cdot R_{ij}}{P_{ij} + R_{ij}}, \tag{3}$$

[3] http://library.cern.ch

where $1 \leq i \leq m$, $1 \leq j \leq n$. P_{ij} and R_{ij} are defined as follows:

$$P_{ij} = \frac{\text{Number of texts from cluster } i \text{ in class } j}{\text{Number of texts in cluster } i}, \tag{4}$$

and

$$R_{ij} = \frac{\text{Number of texts from cluster } i \text{ in class } j}{\text{Number of texts in class } j}. \tag{5}$$

The global performance of the clustering is calculated using the values of F_{ij}. This measure is named F measure and it is shown as follows:

$$F = \sum_{1 \leq i \leq m} \frac{|G_i|}{|D|} \max_{1 \leq j \leq n} F_{ij}, \tag{6}$$

where $|D|$ is the number of documents in the collection.

5 Experimental Results

Our main concern was to evaluate the term selection methods described above, in the clustering of abstracts task, specifically in a narrow domain. Thus, we have used only one clustering method, k-NN based on Jaccard similarity function [18], which is consider as unsupervised, and therefore it complies with our requirements.

5.1 Test over a Subset of *hep-ex*

In order to obtain a first glance of the behaviour of each term selection method used in our experiments, we performed a first test over a subset of *hep-ex*, composed by 500 abstracts taken randomly from the original collection; in the case of those categories with only one instance, we randomly choose two categories. The threshold used as the minimum similarity accepted in the k-NN clustering method was tuned over this collection. The average of similarities was used as a threshold.

Figure 1 shows F values for every term selection method executed over different percentages of the collection's vocabulary (from 600 to 2,000 terms).

Given a percentage of the collection vocabulary, DF and TS methods selected the higher score terms. TP method selected terms in a local fashion; i.e. it took a given number of terms from each text. Therefore, comparison among methods must be done through the vocabularies obtained in each selection of terms carried out by the methods. DF and TS methods used from 2% to 70% of the vocabulary terms. This range corresponds from 21 to 1,700 of the total terms in the collection. The TP selection method took from 5 to 30 terms from each text, given a similar range of total terms. In Fig. 1, the results of these three methods are shown; the horizontal axis represents the number of terms and the vertical axis the F values (eq. 6). In order to apply TS method, similarity matrix was calculated as 3-tuples (T_i, T_j, sim_{ij}) and sorted according sim_{ij}, then ts_t

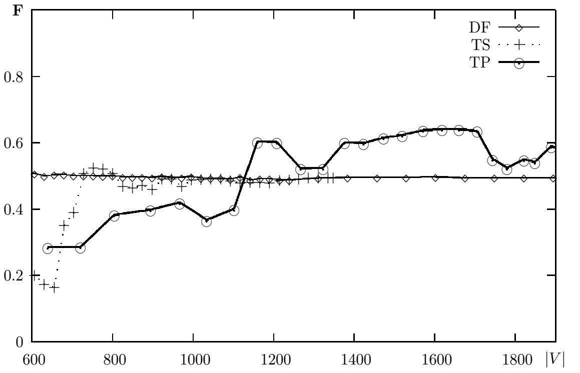

Fig. 1. Behaviour of DF, TS and TP methods in a subset of *hep-ex*

was computed for all terms. Since only 1,349 terms were obtained, threshold β was fixed to 0.

DF method was very stable but it did not help to the clustering task. From the beginning, DF included the most frequent terms in the texts, and this contributed to mantain a minimum level of similarity during the clustering task. Baseline, i.e. the clustering done without term selection ($F = 0.5004$), indicates that DF selects terms to represent texts that mantain resemblance with the original ones. On the other hand, TS method reached the maximum F value after 700 terms, and after 900 terms it obtained stability as well as the DF method did.

TP method outperformed the other two methods. The maximum F value for TP method was 0.6415. This value was reached with a vocabulary size of 1,661 terms which corresponds to only 22 terms per text. The unstability of TP method is derived from noisy words that are difficult to detect because of their low frequencies. Next subsection presents an analysis of the TP selection process, in order to control the unstability.

Analysis of the Unstability of TP: Although the TP method obtained the high F values, it did not allowed to decide the best quantity of terms to be used in the clustering task. It would be desirable to determine the best selection through an indicator based on characteristics of the collection. First of all, clustering method we have used has shown better performance when the number of clusters diminishes. This fact may be used in combination with $\bar{df}_{V_i}$, which is explained in the following paragraph.

Let C_i be the text collection composed by the texts whose terms have been obtained by applying the TP method and by including the i terms with frequency value closer to TP_V from each original text. Let V_i be the vocabulary of C_i and $\bar{df}_{V_i}$ the average of df_t for terms t that belong to V_i but do not belong to V_{i-1}. $\bar{df}_{V_i}$ value is linked to the similarity among the texts. Clearly, the lowest value of $\bar{df}_{V_i}$ is 1, and it means that the new terms added to V_{i-1} are not shared by the texts of C_i. In our experiments it was observed that a decreasing in the $\bar{df}_{V_i}$ value ($\bar{df}_{V_i} < \bar{df}_{V_{i-1}}$) contributed to change instances from an incorrect cluster to a correct one. Therefore, terms with low $\bar{df}_{V_i}$ help to distribute texts into the clusters. Now, we can define an indicator of the goodness of a selection C_i.

Whenever the number of clusters (N_i) decreases after applying clustering to C_i, a lower $\bar{df}_{V_i}$ value means that new terms added to vocabulary V_i will provide a rising of similarity between texts in C_i. In such conditions $\bar{df}_{V_i}$ indicates a good selection. A way to express the above description is by saying that a good clustering supposes that $\bar{df}_{V_i}$ should be greater than $\bar{df}_{V_{i-1}}$ and N_i should be greater than N_{i-1}. We define the goodness of selection C_i as:

$$dfN_i = \frac{(N_i - N_{i-1}) \times (\bar{df}_{V_i} - \bar{df}_{V_{i-1}})}{N_i}. \qquad (7)$$

In Table 3 a neighbour of the maximum value of dfN_i is shown. Row 1 shows the i number of terms selected by the TP method; row 2, the size of the vocabulary of C_i; row 3, the normalized values of dfN_i; and row 4, the F measure. As we can see, dfN_i obtains the maximum value at $i = 22$, as also F does. Thus, independently of unstability of TP method, dfN_i can be used in order to determine what collection C_i must be used in the clustering task.

Table 3. Some normalized values of dfN_i

i	20	21	22	23	24
$\|V_i\|$	1,572	1,619	1,661	1,706	1,744
dfN_i	0.573	0.621	1.027	0.584	0.990
F	0.637	0.6411	0.6415	0.636	0.551

5.2 Test over the Whole *hep-ex* Collection

An experiment was performed using the entire collection and applying the three methods described in Section 3. In this case, the noisy words had a notably effect, mainly in the TP method. Since TP method selects one term per time for each text, a wrong selection may be crucial in the clustering task. In some cases, this iterative process includes words that change dramatically the composition of texts. Thus, a term with very low DF value changes threshold used in the clustering task. We tried to face this problem with an enrichment of terms selected by TP. It is not possible to solve this task using related terms dictionaries like WordNet, since the terminology of texts is very specialized (see [6]). The problem was solved using n-grams as an approximation to related words.

Improving Transition Point Approach: A refined method based on the Transition Point technique was proposed in order to improve the results obtained over the whole collection of *hep-ex*. This method was named *Transition Point and Mutual Information* (TPMI), and basically uses $idtp(t, T)$ and mutual information. Thus TPMI is a refinement of the selection method provided by TP.

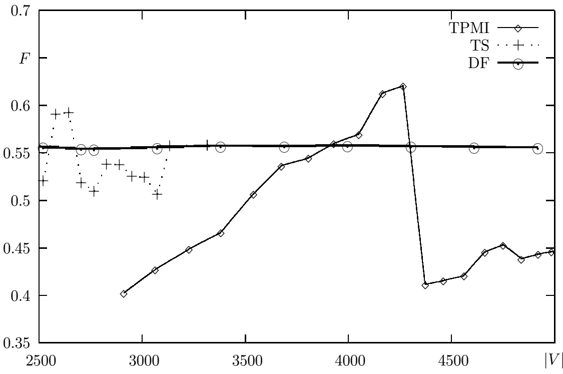

Fig. 2. Behaviour of DF, TS and TPMI term selection methods

Let TP_V be the transition point of the text $T = [t_1, \ldots, t_k]$. We can calculate MI score of each term t_i as $MI(TP_V, t_i)$. The TPMI will assigns the final score:

$$tpmi(t_i, T) = idtp(t_i, T) * MI(TP_V, t_i) \tag{8}$$

$MI(x, y)$ was computed considering n-grams of x, where y appears at a distance of 2 words from x, and the frequency of both x and y was greater than 2.

The results obtained by using this refined method are shown in Figure 2. There we can see that this approach obtains the best value of F measure. Very similar results of clustering on the whole collection were obtained for DF and TS methods, with respect to the subset of *hep-ex*. Anyway, TS method reached the maximum F value (0.5925) with 43% of terms, which corresponds to a collection vocabulary size of 2,644 terms, and only 3,318 terms hold the threshold β. Whereas the DF method is very stable, it mantains its F values below of the baseline (0.5919). TPMI method had a good high peak ($F = 0.6206$) taking 20 terms, and giving a vocabulary size of 4,268 terms

6 Conclusions

In this paper we have proposed a new use of the Transition Point technique in the task of clustering of abstracts in a narrow domain. We used as a corpus a set of documents originally stored at CERN, in the *High Energy Physics* domain, which led to experiment with real collections conformed by very short texts (*hep-ex*). Findings after the execution of three unsupervised methods (DF, TS and TP) were that TP outperforms the other two methods over a subset of *hep-ex*. However, when the whole collection was used, a new filtering method had to be developed in order to improve the previous results. This method was named TPMI, and it used a dictionary of related terms, constructed over the same collection by using mutual information, since common dictionaries are not able to solve this case due to the very specialized vocabulary of this particular domain. After the calculation of a baseline in both experiments was carried out, we could verify that this value was outperformed by our approaches.

We observed that there are not methods to determine the number of terms that a term selection method must obtain, in order to carry out the clustering task. Due to the unstability of TP, we carried out an analysis for explaining this behaviour and therefore to be able to determine the number of terms needed in such task. It is very important to continue with the study of the stability control for this methods, since, this is in fact the key in the clustering of very short texts.

Clustering abstracts in a narrow domain has received not too much attention by the computational linguistic community, and therefore it is very important to continue with the experiments in this area.

Acknowledgments

We thank to Davide Buscaldi by his useful comments on the present work. Also to Arturo Ráez and Alfonso Ureña who kindly provided the *hep-ex* collection.

References

1. M. Alexandrov, A. Gelbukh, P. Rosso: *An Approach to Clustering Abstracts*, A. Montoyo et al. (Eds.): NLDB 2005, LNCS 3513, pp. 275–285, 2005.
2. A. Booth: *A Law of Ocurrences for Words of Low Frequency*, Information and control, 1967.
3. C. Bueno, D. Pinto, H. Jimenez: *El párrafo virtual en la generación de extractos*, Research on Computing Science Journal, 2005.
4. R. Cabrera, D. Pinto, H. Jimenez, D. Vilariño: *Una nueva ponderación para el modelo de espacio vectorial de recuperación de información*, Research on Computing Science Journal, 2005.
5. H. Jimenez, D. Pinto, P. Rosso, *Selección de Términos No Supervisada para Agrupamiento de Resúmenes*, In proceedings of Workshop on Human Language, ENC05, 2005.
6. H. Jiménez-Salazar, D. Pinto & P. Rosso: *Uso del punto de transición en la selección de términos índice para agrupamiento de textos cortos*, Journal: Procesamiento del Lenguaje Natural, Num. 35, pp. 114–118, 2005.
7. T. Liu, S. Liu, Z. Chen, W.-Y. Ma: *An Evaluation on Feature Selection for Text Clustering*, In Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.
8. P. Makagonov , M. Alexandrov, K. Sboychakov: *A toolkit for development of the domain oriented dictionaries for structuring document flows*, In: Data Analysis, Classification, and Related Methods, Studies in classification, data analysis, and knowledge organization, Springer, pp. 83–88, 2000.
9. P. Makagonov, M. Alexandrov, A. Gelbukh: *Clustering Abstracts instead of Full Texts*, Text, Speech and Dialogue (TSD-2004). Lecture Notes in Artificial Intelligence, N 3206, Springer-Verlag, pp. 129–135, 2004.
10. C. Manning, H. Schütze: *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridge, MA, May 1999.

11. A. Montejo-Ráez, L. A. Ureña-López, R. Steinberger: *Text Categorization using bibliographic records: beyond document content*, Journal: Procesamiento del Lenguaje Natural, Num. 35, pp. 119–16, 2005.
12. E. Moyotl, H. Jiménez: *An Analysis on Frequency of Terms for Text Categorization*, Proceedings of XX Conference of Spanish Natural Language Processing Society (SEPLN-04), 2004.
13. E. Moyotl-Hernández, H. Jiménez-Salazar: *Enhancement of dtp feature selection method for text categorization.* Lecture Notes in Computer Science 3406, Gelbukh (Ed.), pp. 719–722, 2005.
14. D. Pinto, F. Pérez:*Una Técnica para la Identificación de Términos Multipalabra*, Proceedings of 2nd. National Conference on Computer Science, México, 2004.
15. Edgar Moyotl Hernández: *DTP, un metodo de selección de términos para agrupamiento de textos*, Tesis de maestría: Facultad de Ciencias de la Computación, BUAP, 2005.
16. C. J. van Rijsbergen: *Information Retrieval*, London, Butterworths, 1999.
17. F. Sebastiani: *Machine Learning in Automated Text Categorization*, ACM Computing Surveys, Vol. 34(1), pp 1–47, 2002.
18. K. Shin, S. Y. Han: *Fast clustering algorithm for information organization*, In A. F. Gelbukh, editor, CICLing, Lecture Notes in Computer Science, Volume 2588, pp. 619–622, 2003.
19. M. Tovar, M. Carrillo, D. Pinto, H. Jimenez, *Combining Keyword Identification Techniques*, Journal: Research on Computing Science, 2005.
20. R. Urbizagástegui: *Las posibilidades de la Ley de Zipf en la indización automática*, Research report of the California Riverside University, 1999.
21. Y. Yang: *Noise Reduction in a Statistical Approach to Text Categorization*, in *Proc. of SIGIR-ACM*, pages 256–263, 1995.
22. G. K. Zipf: *Human Behavior and the Principle of Least-Effort*, Addison-Wesley, Cambridge MA, 1949.