Clustering Narrow-Domain Short Texts by using the Kullback-Leibler Distance *

David Pinto[(1,2)], José-Miguel Benedí[1], Paolo Rosso[1]

[1] Department of Information Systems and Computation,
UPV, Valencia 46022,
Camino de Vera s/n, SPAIN
{dpinto, jbenedi, prosso}@dsic.upv.es

[2] Faculty of Computer Science, BUAP, Puebla 72570,
Ciudad Universitaria, MEXICO
dpinto@cs.buap.mx

Abstract. Clustering short length texts is a difficult task itself, but adding the narrow domain characteristic poses an additional challenge for current clustering methods. We addressed this problem with the use of a new measure of distance between documents which is based on the symmetric Kullback-Leibler distance. Although this measure is commonly used to calculate a distance between two probability distributions, we have adapted it in order to obtain a distance value between two documents. We have carried out experiments over two different narrow-domain corpora and our findings indicates that it is possible to use this measure for the addressed problem obtaining comparable results than those which use the Jaccard similarity measure.

1 Introduction

The clustering of narrow-domain short texts is an emergent area that has been not attended into detail by the computational linguistic community and only few works can be found in literature [1] [11] [15] [19]. This behaviour may be 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. Therefore, it is difficult to deal with this kind of data: if a term selection method is applied, this has to be done very carefully because term frequencies in the texts are very low. Generally only 10% or 20% of the keywords from the complete keyword list occur in every document and their absolute frequency usually is one or two, and only sometimes three or four [1]. In this situation, changing a keyword frequency by one can significantly change the clustering results.

However, most current digital libraries and other web-based repositories of scientific and technical information provide free access only to abstracts and

* This work has been partially supported by the MCyT TIN2006-15265-C06-04 project, as well as by the BUAP-701 PROMEP/103.5/05/1536 grant

not to the full texts of the documents. Evenmore, some repositories such as the well known MEDLINE[1], and the Conseil Européen pour la Recherche Nucléaire (CERN)[2], receive hundreds of publications every day that must be categorized on some specific domain, sometimes with an unknown number of categories a priori. This led to construct novel methods for dealing with this real problem. Although sometimes, keywords are provided by authors for each scientific document, it has been seen that this information is insufficient for conforming a good clustering [21]; evenmore, some of these keywords can lead to more confusion on the clustering process.

We have carried out a set of experiments and our results have been compared with those published earlier in this field. We have used the two corpora presented in [19] and the one suggested in [21], which we consider the most appropiate for our investigation because of their intrinsic characteristics: narrow-domain, short texts and number of documents. The two best hierarchical clustering methods reported in [19] were also implemented. Finally, we have used, as refered by [11], three different feature selection techniques in order to improve the clustering task.

The comparison between documents is performed introducing a symmetric Kullback-Leibler (KL) divergence. As the texts may differ in the terms, the frequency of many compared terms in the document will be zero. This causes problems in the KL distance computation when probabilities are estimated by frequencies of occurrence. In order to avoid this issue, a special type of back-off scheme is introduced. The next section explains into detail the use of the Kullback and Leibler distance as a similarity measure in the clustering task. In Section 3 we present the characteristics of every corpus used in our experiments, describing the use of feature selection techniques for selecting only the most valuable terms from each corpus. The description and the results obtained in our executions are presented in Section 4 and, finally the conclusions of our experiments are given.

2 The Kullback-Leibler Distance

In 1951 Kullback and Leiber studied a measure of information from the statistical aspect viewpoint; this measure involved two probability distributions associated with the same experiment [13]. The Kullback-Leibler divergence is a measure of how different two probability distributions (over the same event space) are. The KL divergence of the probability distributions P, Q on a finite set X is defined as shown in Equation 1.

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) log \frac{P(x)}{Q(x)} \tag{1}$$

Since this KL divergence is a non-symmetric information theoretical measure of distance of P from Q, then it is not strictly a distance metric. During the past

[1] http://www.nlm.nih.gov/

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

years, various measures have been introduced in the literature generalizing this measure. We therefore have used the following different symmetric Kullback-Leibler divergences i.e., Kullback-Leibler Distances (KLD) for our experiments. Each KLD corresponds to the definition of Kullback and Leibler [13], Bigi [4], Jensen [10], and Bennet [2] [27], respectively.

$$D_{KLD1}(P||Q) = D_{KL}(P||Q) + D_{KL}(Q||P) \tag{2}$$

$$D_{KLD2}(P||Q) = \sum_{x \in X} (P(x) - Q(x)) log \frac{P(x)}{Q(x)} \tag{3}$$

$$D_{KLD3}(P||Q) = \frac{1}{2} \left[D_{KL} \left(P|| \frac{P+Q}{2} \right) + D_{KL} \left(Q|| \frac{P+Q}{2} \right) \right] \tag{4}$$

$$D_{KLD4}(P||Q) = max \left(D_{KL}(P||Q) + D_{KL}(Q||P) \right) \tag{5}$$

KL and KLD have been used in many natural language applications like query expansion [8], language models [3], and categorization [4]. They have also been used, for instance, in natural language and speech processing applications based on statistical language modeling [9], and in information retrieval, for topic identification [5]. In this paper, we have considered to calculate the corpus document similarities in an inverse function with respect to the distance defined in Equations (2), (3), (4), or (5).

In the text clustering model proposed in this paper, a document j is represented by a term vector of probabilities $\overrightarrow{d_j}$ and the distance measure is, therefore, the KLD (the symmetric Kullbach-Leibler divergence) between a pair of documents $\overrightarrow{d_i}$ and $\overrightarrow{d_j}$.

A smoothing model based on back-off is proposed and, therefore, frequencies of the terms appearing in the document are discounted, whereas all the other terms which are not in the document are given an *epsilon* (ϵ) probability, which is equal to the probability of unknown words. The reason is that in practice, often not all the terms in the vocabulary (V) appear in the document d_j. Let $V(d_j) \subset V$ be the vocabulary of the terms which do appear in the documents represented in d_j. For the terms not in $V(dj)$, it is useful to introduce a back-off probability for $P(t_k, d_j)$ when t_k does not occur in $V(d_j)$, otherwise the distance measure will be infinite. The use of a back-off probability to overcome the data sparseness problem has been extensively studied in statistical language modelling (see, for instance [17]). The resulting definition of document probability $P(t_k, d_j)$ is:

$$P(t_k, d_j) = \begin{cases} \beta * P(t_k|d_j), & \text{if } t_k \text{ occurs in the document } d_j \\ \varepsilon, & \text{otherwise} \end{cases} \tag{6}$$

with:

$$P(t_k|d_j) = \frac{tf(t_k, d_j)}{\sum_{x \in d_j} tf(t_k, d_j)}$$

where: $P(t_k|d_j)$ is the probability of the term t_k in the document d_j, β is a normalisation coefficient which varies according to the size of the document; and ε is a threshold probability for all the terms not in d_j.

Equation 6 must respect the following property:

$$\sum_{k \in d_j} \beta * P(t_k|d_j) + \sum_{k \in V, k \notin d_j} \varepsilon = 1$$

and β can be easily estimated for a document with the following computation:

$$\beta = 1 - \sum_{k \in V, k \notin d_j} \varepsilon$$

3 Description of the corpora

In the experiments we have carried out, three corpora with different characteristics with respect to their size and their balance were used. We consider that all these very narrow domain corpora are suitable for our experiments because of their average size per abstract and their narrow domain. In the following subsections we describe each corpus into detail.

3.1 The *CICLing-2002* corpus

This corpus is made up by 48 abstracts from the *Computational Linguistics* domain, which corresponds to the conference *CICLing 2002*. This collection was used by Makagonov et al. [15] in their experiments on clustering short texts of narrow domains. We consider it a very small but a needed reference corpus, also for manually investigating the obtained results.

The topics of this corpus are the following ones: Linguistic (semantics, syntax, morphology, and parsing), Ambiguity (WSD, anaphora, POS, and spelling), Lexicon (lexics, corpus, and text generation), and Text Processing (information retrieval, summarization, and classification of texts). The distribution and the features of this corpus are shown in Tables 1 and 2, respectively.

Table 1. Distribution of the *CICLing-2002* corpus

Category	**# of abstracts**
Linguistics	11
Ambiguity	15
Lexicon	11
Text Processing	11

Table 2. Other features of the *CICLing-2002* corpus

Feature	Value
Size of the corpus (bytes)	23,971
Number of categories	4
Number of abstracts	48
Total number of terms	3,382
Vocabulary size (terms)	953
Term average per abstract	70.45

3.2 The *hep-ex* corpus of CERN

This corpus is based on the collection of abstracts compiled by the University of Jaén, Spain [16], named *hep-ex*, and it is composed by 2,922 abstracts from the *Physics* domain originally stored in the data server of the CERN.

The distribution of the categories for each corpus is better described in Table 3; other characteristics are shown in Table 4. As can be seen, this corpus is totally unbalanced, which makes this task even more challenging.

Table 3. Categories of the *hep-ex* corpus

Category	# of abstracts
Particle physics (experimental results)	2,623
Detectors and experimental techniques	271
Accelerators and storage rings	18
Particle physics (phenomenology)	3
Astrophysics and astronomy	3
Information transfer and management	1
Nonlinear systems	1
Other fields of physics	1
XX	1

Table 4. Other features of the *hep-ex* corpus

Feature	Value
Size of the corpus (bytes)	962,802
Number of categories	9
Number of abstracts	2,922
Total number of terms	135,969
Vocabulary size (terms)	6,150
Term average per abstract	46.53

3.3 The *KnCr* corpus of MEDLINE

This corpus, named KnCr, was created for the specific task of clustering short texts of a medical narrow domain [21]. It consists of 900 abstracts related with the "Cancer" domain. Table 5 and 6, show the complete characteristics of this new corpus.

Table 5. Categories of the *KnCr* corpus

Category	# of abstracts	Category	# of abstracts
blood	64	lung	99
bone	8	lymphoma	30
brain	14	renal	6
breast	119	skin	31
colon	51	stomach	12
genetic studies	66	therapy	169
genitals	160	thyroid	20
liver	29	Other (XXX)	22

Table 6. Other features of the *KnCr* corpus

Feature	Value
Size of the corpus (bytes)	834,212
Number of categories	16
Number of abstracts	900
Total number of terms	113,822
Vocabulary size (terms)	11,958
Term average per abstract	126.47

3.4 Preprocessing

We have preprocessed all these collections by eliminating stop words and by applying the Porter stemmer [22]. The characteristics given in the above tables for each corpus were obtained after applying this preprocessing phase. The results reported in [19] show that better results can be obtained by using those terms which contribute to a better clustering (not noisy terms), instead of the complete vocabulary. This fact have led us to study this issue in order to apply it to our preprocessed corpora. Up to now, different Feature Selection Techniques (FSTs) have been used in the clustering task. However, clustering abstracts for a narrow domain implies the well known problem of the lackness of training corpora. This led us to use unsupervised term selection techniques instead of supervised ones. Following we describe briefly all the techniques employed in our experiments.

3.5 Description of the FSTs used

The first two unsupervised techniques we are presenting in this sub-section have shown their value in the clustering [14] and categorization area [25]. Particulary, the document frequency technique is an effective and simple technique, and it is known that it obtains comparable results to the classical supervised techniques like χ^2 and Information Gain [26]. With respect to the transition point technique, it has a simple calculation procedure, which has been used in other areas of computational linguistic besides clustering of short texts: categorization of texts, keyphrases extraction, summarization, and weighting models for information retrieval systems (see [19]). Therefore, we consider that there exists enough evidence to use this technique as a term selection process.

1. *Document Frequency (DF)*: This technique assigns the value df_t to each term t, where df_t means the number of texts, in a collection, where t ocurrs. This technique assumes that low frequency terms will rarely appear in other documents, 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:

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

   Besides, both texts, T_i and T_j must be as similar as a given threshold, i.e., $sim(T_i, T_j) \geq \beta$, where β must be tuned according to the values inside of 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 [25] and [18].
3. *Transition Point (TP)*: A higher value of weight is given to each term t, as its frequency is closer to a frequency named the transition point (TP_V) which can be found by an automatic inspection of the vocabulary frequencies of each text, identifying the lowest frequency (from the highest frequencies) that it is not repeated; this characteristic comes from the formulation of Booth's law for low frequency words [6] (see [19] for a complete explanation of this procedure). The following equation shows how to calculate the final 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 DF and TP techniques 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 technique to be in $O(n^2)$, where n is the number of texts in the data set.

4 Experimental results

Clustering very short narrow-domain texts, implies basically two steps: first it is necessary to perform the feature selection process and after the clustering itself. We have used the three unsupervised techniques described in Section 3.5 in order to sort the corpora vocabulary in non-increasing order, with respect to the score of each FST. Thereafter, we have selected different percentages of the vocabulary (from 20% to 90%) in order to determine the behaviour of each technique under different subsets of the vocabulary. The following step involves the use of clustering methods; three different clustering methods were employed for this comparison: Single Link Clustering (SLC) [12], Complete Link Clustering (CLC)[12], and KStar [23].

In order to obtain the best description of our experiments, we have carried out a v-fold cross validation evaluation [7]. This process implies to randomly split the original corpus in a predefined set of partitions, and then calculate the average F-measure (described in the next sub-section) among all the partitions results. The v-fold cross-validation allows to evaluate how well each cluster "performs" when is repeatedly cross-validated in different samples randomly drawn from the data. Consequently, our results will not be casual through the use of a specific clustering method and a specific data collection. In our case, we have used five partitions for the *CICLing-2002* corpus and, thirty for both, the *hep-ex* and the *KnCr* collections.

We have used the F-measure for determining the quality of clusters obtained, as it is described in the next sub-section. Thereafter the results are presented and discussed.

4.1 Performance measurement

We employed the F-measure, which is commonly used in information retrieval [24], 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.

$$F_{ij} = \frac{2 \cdot P_{ij} \cdot R_{ij}}{P_{ij} + R_{ij}}, \tag{7}$$

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

$$P_{ij} = \frac{\text{Number of texts from cluster } i \text{ in class } j}{\text{Number of texts from cluster } i}, \tag{8}$$

and

$$R_{ij} = \frac{\text{Number of texts from cluster } i \text{ in class } j}{\text{Number of texts in class } j}. \tag{9}$$

The global performance of a clustering method is calculated by using the values of F_{ij}, the cardinality of the set of clusters obtained, and normalizing by the total number of documents in the collection ($|D|$). The obtained measure is named F-measure and it is shown in equation 10.

$$F = \sum_{1 \leq i \leq m} \frac{|G_i|}{|D|} \max_{1 \leq j \leq n} F_{ij}. \qquad (10)$$

5 Results

In the experiments we have carried out, the DF and TS techniques do not improve the results obtained by the transition point technique, which reinforces the hypothesis suggested by [19]. Besides, we have observed that there is not a significant difference between any of the symmetric KL distances. Therefore, we consider that in other applications, the simplest one should be used. Tables 7, 8 and, 9 show our evaluation results for all Kullback-Leibler approaches implemented, by using the *CICLing-2002*, *hep-ex* and, *KnCr* corpus, respectively. In each table, we have defined three sections, named (a), (b) and, (c), each one corresponding to the use of the TP, DF and, TS feature selection technique, respectively. In the first column we have named as *KullbackOriginal*, *KullbackBigi*, *KullbackJensen* and, *KullbackMax*, the KLD defined by Kullback and Leibler [13], Bigi [4], Jensen [10], and Bennet [2] [27], respectively.

Table 7. Results obtained by using the *CICLing-2002* corpus

	(a)-TP			(b)-DF			(c)-TS		
	SLC	**CLC**	**KStar**	**SLC**	**CLC**	**KStar**	**SLC**	**CLC**	**KStar**
KullbackOriginal	0,6	0,7	0,7	0,6	0,6	0,6	0,5	0,6	0,6
KullbackBigi	0,6	0,7	0,7	0,6	0,7	0,6	0,5	0,5	0,6
KullbackJensen	0,6	0,6	0,7	0,6	0,6	0,6	0,5	0,6	0,6
KullbackMax	0,6	0,7	0,7	0,6	0,7	0,6	0,5	0,6	0,6

Table 8. Results obtained by using the *hep-ex* corpus

	(a)-TP			(b)-DF			(c)-TS		
	SLC	**CLC**	**KStar**	**SLC**	**CLC**	**KStar**	**SLC**	**CLC**	**KStar**
KullbackOriginal	0,86	0,83	0,68	0,60	0,83	0,68	0,80	0,84	0,67
KullbackBigi	0,86	0,82	0,69	0,60	0,82	0,67	0,80	0,85	0,67
KullbackJensen	0,85	0,83	0,68	0,61	0,83	0,69	0,80	0,83	0,66
KullbackMax	0,86	0,83	0,69	0,61	0,83	0,68	0,80	0,85	0,67

Table 9. Results obtained by using the *KnCr* corpus

	(a)-TP			(b)-DF			(c)-TS		
	SLC	**CLC**	**KStar**	**SLC**	**CLC**	**KStar**	**SLC**	**CLC**	**KStar**
KullbackOriginal	0,52	0,38	0,39	0,51	0,37	0,38	0,49	0,36	0,38
KullbackBigi	0,52	0,38	0,39	0,51	0,37	0,38	0,49	0,36	0,38
KullbackJensen	0,52	0,36	0,40	0,52	0,36	0,39	0,48	0,34	0,38
KullbackMax	0,51	0,37	0,40	0,51	0,37	0,39	0,50	0,37	0,38

We have made a comparison among our results and those reported by Pinto et al. [20]. This evaluation is presented in Tables 10 and 11, where our best approach is compared with the results presented in [20], which we have named *PintoetAl*. The comparison could be done only by using both, the *CICLing-2002* and the *hep-ex* corpora, because up to now, there are not published results with the characteristics needed for the *KnCr* corpus. We have observed that the use of KLD obtains comparable results, and we consider that this behaviour is derived from the size of each text. We are suggesting to use a smooth procedure, but the number document terms that does not appear in the corpus vocabulary can be extremely high. Further analysis will investigate this issue.

Table 10. Comparison by using the *CICLing-2002* corpus

	(a)-TP			(b)-DF			(c)-TS		
	SLC	**CLC**	**KStar**	**SLC**	**CLC**	**KStar**	**SLC**	**CLC**	**KStar**
KullbackMax	0,6	0,7	0,7	0,6	0,7	0,6	0,5	0,6	0,6
PintoetAl	0,6	0,7	0,7	0,6	0,7	0,6	0,5	0,7	0,6

Table 11. Comparison by using the *hep-ex* corpus

	(a)-TP			(b)-DF			(c)-TS		
	SLC	**CLC**	**KStar**	**SLC**	**CLC**	**KStar**	**SLC**	**CLC**	**KStar**
KullbackMax	0,86	0,83	0,69	0,61	0,83	0,68	0,80	0,85	0,67
PintoetAl	0,77	0,87	0,69	0,59	0,86	0,68	0,74	0,86	0,67

6 Conclusions

We have addressed the problem of clustering short texts of a very narrow domain with the use of a new measure of distance between documents, which is based on the symmetric Kullback-Leibler distance. We observed that there are very little differences in the use of any of the symmetric KL distances analysed. This fact led us to consider that in case of using this approach, the simplest implementation should be used.

Moreover, we have evaluated our approach with three different short-text narrow-domain corpora and, our findings indicates that it is possible to use this measure to tackle this problem, obtaining comparable results than those that uses the Jaccard similarity measure.

Despite we have implemented the KLD for using it in the short-text narrow-domain clustering task, we consider that this approach could be sucessfully implemented in other clustering tasks which involve the use of a more general domain and big size text corpora.

The use of a smooth procedure should be of more benefit as far as the vocabulary of each document would be more similar to the corpus vocabulary. Therefore, we consider that a performance improving could be obtained by using a term expansion method before calculating the similarity matrix with the analysed KLD. Further analysis will investigate this issue.

References

1. M. Alexandrov, A. Gelbukh, and P. Rosso: *An Approach to Clustering Abstracts*, In Proceedings of the 10th International Conference NLDB-05, volume 3513 of Lecture Notes in Computer Science, pages 275-285, Springer-Verlag, 2005.
2. C.H. Bennett, P. Gács, M. Li, P. Vitányi, and W. Zurek: *Information Distance*, IEEE Trans. Inform. Theory, 44:4, pages 1407–1423, 1998.
3. B. Bigi, Y. Huang, R. d. Mori: *Vocabulary and Language Model Adaptation using Information Retrieval*, In Proceedings of the ECIR-2003, volume 2633 of Lecture Notes in Computer Science, pages 305-319, Springer-Verlag, 2003.
4. B. Bigi: *Using Kullback-Leibler Distance for Text Categorization*, In Proceedings of the ECIR-2003, volume 2633 of Lecture Notes in Computer Science, pages 305-319, Springer-Verlag, 2003.
5. B. Bigi, R. d. Mori, M. El-Bèze, T. Spriet: *A fuzzy decision strategy for topic identication and dynamic selection of language models*, Special Issue on Fuzzy Logic in Signal Processing, Signal Processing Journal, 80(6):1085–1097, 2000.
6. A. D. Booth: *A Law of Occurrences for Words of Low Frequency*, Information and control, 10(4):386-393, 1967.
7. P. Burman, *A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods*, Biometrika 76(3):503-514, 1989.
8. C. Carpineto, R. d. Mori, G. Romano, B. Bigi: *An information-theoretic approach to automatic query expansion*, ACM Transactions on Information Systems, 19(1):1-27, 2001.
9. I. Dagan, L. Lee, F. Pereira: *Similarity-based models of word cooccurrence probabilities*, Machine Learning, 34(1–3):43-69, 1999.

10. B. Fuglede, F. Topse: *Jensen-Shannon Divergence and Hilbert space embedding*, IEEE Int Sym. Information Theory, 2004.
11. H. Jiménez, D. Pinto, and P. Rosso: *Uso del punto de transición en la selección de términos índice para agrupamiento de textos cortos*, Procesamiento del Lenguaje Natural, 35(1):114-118, 2005 (*in Spanish*).
12. S. C. Johnson: *Hierarchical Clustering Schemes*, Psychometrika, 2:241–254, 1967.
13. S. Kullback, R. A. Leibler: *On information and sufficiency*, Annals of Mathematical Statistics, 22(1):79–86, 1951.
14. T. Liu, S. Liu, Z. Chen, and W. Ma: *An evaluation on feature selection for text clustering*, In T. Fawcett and N. Mishra, editors, ICML, pages 488-495, AAAI Press, 2003.
15. P. Makagonov, M. Alexandrov, and A. Gelbukh: *Clustering Abstracts instead of Full Texts*, In Proceedings of the Seventh International Conference on Text, Speech and Dialogue (TSD 2004), volume 3206 of Lecture Notes in Artificial Intelligence, pages 129-135, Springer-Verlag, 2004.
16. A. Montejo-Ráez, L. A. Ureña-López, and R. Steinberger: *Categorization using bibliographic records: beyond document content*, Procesamiento del Lenguaje Natural, 35(1):119-126, 2005.
17. R. d. Mori: Spoken Dialogues with Computers, Academic Press, 1998.
18. V. Pekar, M. Krkoska, S. Staab. Feature Weighting for Co-occurrence-based Classification of Words, In Proceedings of the 20th Conference on Computational Linguistics, COLING-2004, 2004.
19. D. Pinto, H. Jiménez-Salazar, and P. Rosso: *Clustering abstracts of scientific texts using the transition point technique*, In Alexander F. Gelbukh, editor, CICLing, volume 3878 of Lecture Notes in Computer Science, pages 536-546. Springer-Verlang, 2006.
20. D. Pinto, P. Rosso, A. Juan, and H. Jiménez, : *A Comparative Study of Clustering Algorithms on Narrow-Domain Abstracts*, Procesamiento del Lenguaje Natural, 37(1):43–49, 2006.
21. D. Pinto, and P. Rosso: *KnCr: A Short-Text Narrow-Domain Sub-Corpus of Medline*, In Proceedings of TLH-ENC06, pages 266–269, 2006.
22. M. F. Porter: *An algorithm for suffix stripping*, In Program, 14(3), 1980.
23. K. Shin and S. Y. Han: *Fast clustering algorithm for information organization*, In A. F. Gelbukh, editor, CICLing, volume 2588 of Lecture Notes in Computer Science, pages 619-622, Springer-Verlang, 2003.
24. C. J. Van Rijsbergen: Information Retrieval, 2nd edition, Dept. of Computer Science, University of Glasgow, 1979.
25. Y. Yang: *Noise reduction in a statistical approach to text categorization*, In Proceedings of SIGIR-ACM, pages 256-263, 1995.
26. Y. Yang , J. O. Pedersen. A comparative study on feature selection in text categorization. In Proc. ICML, pages 412–420, 1997.
27. J. Ziv and N. Merhav: *A measure of relative entropy between individual sequences with application to universal classification*, IEEE Transactions on Information Theory, 39(4):1270–1279, 1993.