A Comparative Study of Clustering Algorithms on Narrow-Domain Abstracts *

David Pinto[1,2], **Paolo Rosso, Alfons Juan**
[1]DSIC, UPV, Spain
Camino de Vera s/n, 46022
{dpinto,prosso,ajuan}@dsic.upv.es

Héctor Jiménez-Salazar
[2]FCC, BUAP, Mexico
C.U. Edif. 135, 72050
hgimenezs@gmail.com

Resumen: El agrupamiento de resumenes de textos científicos de dominios sumamente restringidos implica un alto grado de complejidad, debido principalmente al alto grado de traslape de vocabularios entre los textos y la baja frecuencia de ocurrencia de los términos en dichos documentos. El uso de la técnica del punto de transición ha resultado de suma utilidad en esta tarea del Procesamiento del Lenguaje Natural (PLN). Su bondad se encuentra sustentada en el conjunto de palabras que extrae del vocabulario de un texto: los términos de frecuencia media. Si bien, la importancia del uso de este tipo términos en PLN es bastante conocida, la extracción de los mismos no lo es. En este trabajo se presentan resultados experimentales en el uso de dicha técnica como un mecanismo de selección de características en dos corpora de dominios sumamente restringidos. Los resultados experimentales muestran que la técnica elegida obtiene los mejores valores de medida-F bajo cinco diferentes métodos de agrupamiento.
Palabras clave: Agrupamiento de resúmenes, Técnica del punto de transición, Dominios restringidos

Abstract: Clustering abstracts of scientific texts of very narrow domain implies a big challenge. The first problem to attend is the high overlapping among the document's vocabularies, besides the low frequency of these terms. The transition point technique has been successfully used in this area of Natural Language Processing (NLP). Its best properties rely on the extraction of the mid-frequency terms. Although the importance of these terms on NLP has been known from time ago, the exact extraction of these terms is unknown. In this paper we present an application of this technique as a feature selection technique in two corpora of very narrow domain. The experimental results show that the transition point technique obtains the best results of F-measure with five different clustering methods.
Keywords: Clustering of abstracts, Transition Point technique, Narrow domain

1 Clustering on Narrow Domain

Free access to scientific papers in major digital libraries and other web repositories is limited to only their abstracts. Clustering abstracts of very narrow domains is a very challenging task that has been few treated by the computational linguistic community. The aim of this area is to classify scientific documents; moreover, this area proposes to detect emerging study fields by using unsupervised clustering methods. It is well known that clustering methods rely their performance upon the preprocessing step applied to the corpus. In this way, a good technique for selecting a subset of the terms that appear in each scientific paper is needed. However, current keyword-based techniques fail on narrow domain-oriented libraries; this fact is derived from the high terms overlapping in the abstracts and the high number of typical words used in those abstracts, like "In this paper we present...". Some approaches have been given for this new task; their proposals are mainly focused on the selection of a good technique for extracting terms from the vocabulary of each abstract. Makagonov, Alexandrov, and Gelbukh (2004), for instance, proposed simple procedures for improving results by an adequate selection of keywords and a better evaluation of document similarity. Another work in this context is presented in (Alexandrov, Gelbukh, and Rosso, 2005), where an

* This project was partially supported by the R2D2 (CICYT TIC2003-07158-C04-03) and ICT EU-India (ALA/95/23/2003/077-054) research projects, as well as by the BUAP-701 PROMEP/103.5/05/1536 grant.

approach for clustering abstracts in a narrow domain using Stein's MajorClust Method for clustering both, keywords and documents, was presented. Despite the small size of the collection, an interesting work was presented in (Jiménez, Pinto, and Rosso, 2005b), where a new technique for keyword selection was proposed; besides they also used this new technique in the evaluation of a bigger size corpus (Pinto, Jiménez-Salazar, and Rosso, 2006). Their results have motivated this comparative study. Therefore, we are interested in verifying whether this new technique could be capable of improving results obtained in feature selection environment, and to identify its scope. The remaining of this paper is distributed in the following way: first, we introduce the feature selection techniques used by Pinto, Jiménez-Salazar, and Rosso (2006). The next section describes the experiment we carried out, first by introducing the dataset used, and then we present a complete description of the comparative study. The section 3 shows the experiments carried out, and finally a discussion about findings is given.

2 Feature Selection Techniques

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. In the next subsection we describe briefly the transition point technique. In the final subsection we explain all the other techniques used in our experiments.

2.1 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 (Zipf, 1949) and also on the refined studies of Booth (Booth, 1967), as well as Urbizagástegui (Urbizagástegui, 1999). 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 hypothesize 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:

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

where I_1 represents the number of words with frequency equal to 1 in the text T (Moyotl and Jiménez, 2004b) (Urbizagástegui, 1999). 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 formulation of Booth's law for low frequency words (Booth, 1967).

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.,

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

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 like: clustering of short texts (Jiménez, Pinto, and Rosso, 2005a), categorization of texts (Moyotl and Jiménez, 2004a) (Moyotl-Hernández and Jiménez-Salazar, 2005), keyphrases extraction (Pinto and Pérez, 2004) (Tovar et al., 2005), summarization (Bueno, Pinto, and Jiménez-Salazar, 2005), and weighting models for information retrieval systems (Cabrera, Pinto, and H. Jiménez, 2005). Therefore, we believe that there exists enough evidence to use this technique as a term selection process.

2.2 Description of the FSTs used

The first two unsupervised techniques we are presenting in this subsection have demonstrated their value in the clustering area (Liu et al., 2003). 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 (Sebastiani, 2002). With respect to the transi-

tion point technique, it has a simple calculation procedure, and as it was seen in Subsection 2.1, it had also been used in text clustering.

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 in a pair of texts (T_i, T_j) 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 (Yang, 1995).

3. *Transition Point*: 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 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.

3 Experimental results

3.1 Dataset

In our tests we have used two corpora with quite different characteristics with respect to the size and the balance of each one. Following we describe each corpus in detail.

3.1.1 The *CICLing* corpus

This corpus is balanced and it is composed by 48 abstracts from the "Computational Linguistics and Text Processing" domain, which were extracted from the *CICLing 2002* conference[1]. The topics of this corpus are the following: 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 *CICLing*

Category	**# of abstracts**
Linguistics	11
Ambiguity	15
Lexicon	11
Text processing	11
Total	**48**

Table 2: Other features of *CICLing*

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

3.1.2 The *hep-ex* corpus

This corpus is based on the collection of abstracts compiled by the University of Jaén, Spain (Montejo-Ráez, Urena-López, and Steinberger, 2005), named *hep-ex*, and it is composed by 2.922 abstracts from the *Physics* domain originally stored in the data server of the "Conseil Européen pour la Recherche Nucléaire" (CERN)[2].

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

[1] http://www.cicling.org
[2] http://library.cern.ch

Table 3: Categories of *hep-ex*

Category	# of texts
- Experimental results	2.623
- Detectors and experimental techniques	271
- Accelerators and storage rings	18
- Phenomenology	3
- Astrophysics and astronomy	3
- Information transfer and management	1
- Nonlinear systems	1
- Other fields of physics	1
- XX	1
Total	**2.922**

Table 4: Other features of *hep-ex*

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

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

3.2 Description of the experiments

Clustering short-texts of narrow domain, implies basically two steps: first it is necessary to perform the feature selection process. We have used the three unsupervised techniques described in Section 2 in order to sort the vocabulary of the corpora in non-increasing order according to the score of each FST. We have selected different percentages of the sorted vocabulary (from 20% to 90%) in order to determine the behaviour of each technique under different subsets of the vocabulary. The second step involves the use of clustering methods; five different clustering methods were applied for this comparison: Single Link Clustering (SLC), Complete Link Clustering (CLC), K-Nearest Neighbour (KNN), KStar (Shin and Han, 2003) and a modified version of the KStar method (NN1). The aim of the comparative study of the above clustering algorithms was to investigate whether exist a close relationship between a specific clustering method and a specific feature selection technique.

In order to obtain the best description of our experiments, we have carried out a v-fold cross validation. 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 subsection) 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 four partitions for the *CICLing* collection and thirty partitions for the *hep-ex* collection.

3.3 Performance measurement

We employed the F-measure, which is commonly used in information retrieval (Rijsbergen, 1979), 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.

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

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

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

and

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

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 6.

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

3.4 Results

We show in Tables 5 and 6, the maximum F-measure values obtained for each feature se-

lection technique by using five different clustering methods, for the *CICLing* and *hep-ex* corpus, respectively. As may be seen, the transition point technique obtains better or equal results than DF and TS for all the clustering methods for both corpora. Having obtained these results on two different corpora (in size and balance), we believe that the transition point technique could be clustering method independent. In order to further investigate this hypothesis, we have carried out an analysis of each selection technique on the five different clustering methods. By observing a stable behaviour of almost all clustering methods we could confirm the above hypothesis.

Table 5: Maximum F-measure obtained using the *CICLing* corpus

	TP	**DF**	**TS**
KStar	**0,7**	0,6	0,6
SLC	**0,6**	0,6	0,5
CLC	**0,7**	0,7	0,7
NN1	**0,7**	0,7	0,7
KNN	**0,7**	0,6	0,6

Table 6: Maximum F-measure obtained using the *hep-ex* corpus

	TP	**DF**	**TS**
KStar	**0,69**	0,68	0,67
SLC	**0,77**	0,59	0,74
CLC	**0,87**	0,86	0,86
NN1	**0,61**	0,54	0,55
KNN	**0,22**	0,22	0,22

The performance of each feature selection technique, TP, DF, and TS, upon the use of the *hep-ex* corpus and by using the five clustering methods are shown in Figures 1, 2, and 3, respectively. For this corpus, it can be seen that the complete link clustering method obtains the best results in all the FSTs. On the other hand, the KNN method obtains very poor results. By obtaining the average of the three FSTs, we can observe (Figure 4) that there exist some independence (with exception of the SLC method) on the behaviour of each clustering method, which suggests that the feature selection process is independent from the clustering method. In Figure 4 is shown the standard deviation for different sizes of the vocabulary for the *hep-ex* corpus. The behaviour seems to verify our hypothesis, however, more experiments need to be done in the future.

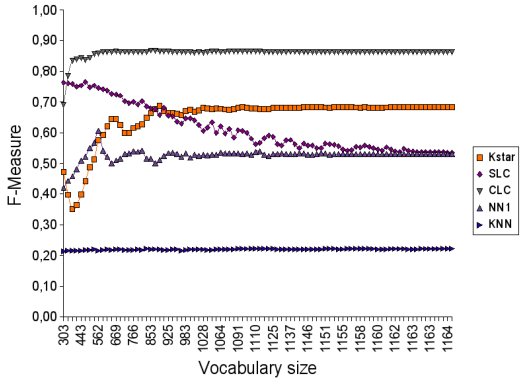

Figure 1: *F*-measure of the TP technique as a function of the vocabulary size for the five clustering methods considered (over the *hep-ex* corpus).

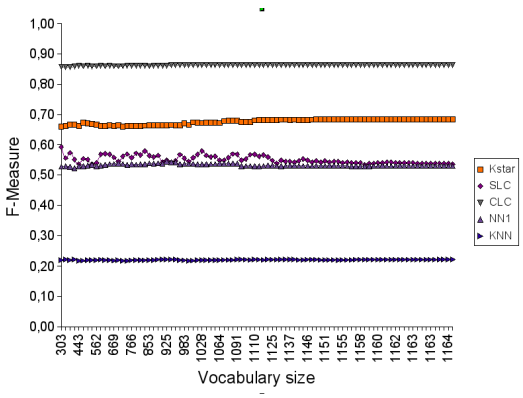

Figure 2: *F*-measure of the DF technique as a function of the vocabulary size for the five clustering methods considered (over the *hep-ex* corpus).

4 *Discussion*

We have carried out a comparative study of the behaviour of five clustering methods applied to two corpora with very different characteristics. Each corpus belongs to a very narrow domain, doing our task even more difficult. The use of the transition point technique has been successful and we have observed that this technique obtains best results in comparison with the DF and TS techniques. Moreover, those results are stable upon the use of different clustering algorithms. This suggests that there exists an independence between the feature selection techniques and the clustering methods. Despite we have used a very strong measure for the clustering process (F-measure), it would

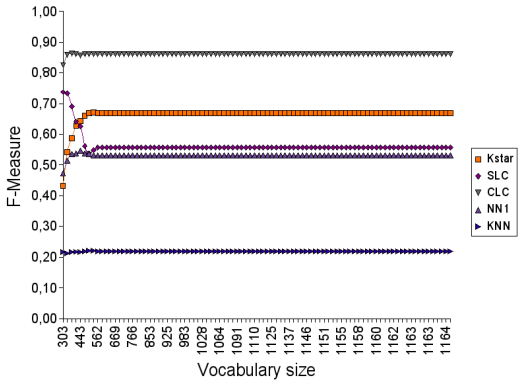

Figure 3: *F*-measure of the TS technique as a function of the vocabulary size for the five clustering methods considered (over the *hep-ex* corpus).

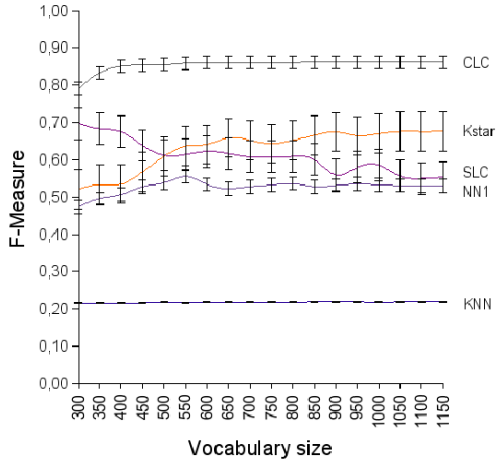

Figure 4: Average behaviour of all FSTs with each clustering method using the *hep-ex* corpus

be desirable to repeat the experiments over other corpora of different domains to confirm our hypothesis. Unfortunately, at the moment there exist a lackness of gold standards for clustering abstracts on narrow domains, doing this task even more difficult. We consider that more attention from the linguistic community is required for the clustering of narrow domain task, not only for experimenting on different feature selection techniques, but also for constructing new narrow domain corpora, with gold standards provided by experts in such domains.

References

Alexandrov, M., A. Gelbukh, and P. Rosso. 2005. An Approach to Clustering Abstracts. In *Proceedings of the 10th International Conference NLDB-05*, Lecture Notes in Computer Science, pages 8–13, Alicante, Spain. Springer-Verlag.

Booth, A. D. 1967. A Law of Occurrences for Words of Low Frequency. *Information and control*, 10(4):386–393.

Bueno, C., D. Pinto, and H. Jiménez-Salazar. 2005. El párrafo virtual en la generación de extractos. *Research on Computing Science*, 13:83–90.

Cabrera, R., D. Pinto, and D. Vilariño H. Jiménez. 2005. Una nueva ponderación para el modelo de espacio vectorial de recuperación de información. *Research on Computing Science*, 13:75–81.

Jiménez, H., D. Pinto, and P. Rosso. 2005a. Selección de términos no supervisada para agrupamiento de resúmenes. In *proceedings of Workshop on Human Language, ENC05*, pages 86–91.

Jiménez, H., D. Pinto, and P. Rosso. 2005b. Uso del punto de transición en la selección de términos índice para agrupamiento de textos cortos. *Procesamiento del Lenguaje Natural*, 35(1):114–118.

Liu, T., S. Liu, Z. Chen, and W. Ma. 2003. An evaluation on feature selection for text clustering. In T. Fawcett and N. Mishra, editors, *ICML*, pages 488–495. AAAI Press.

Makagonov, P., M. Alexandrov, and A. Gelbukh. 2004. 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, Brno, Czech Republic. Springer-Verlag.

Montejo-Ráez, A., L. A. Urena-López, and R. Steinberger. 2005. Categorization using bibliographic records: beyond document content. *Procesamiento del Lenguaje Natural*, 35(1):119–126.

Moyotl, E. and H. Jiménez. 2004a. An analysis on frequency of terms for text categorization. In SEPLN, editor, *Memorias del XX Congreso de la Sociedad Espaola para el Procesamiento del Lenguaje Natural*, pages 141–146. SEPLN.

Moyotl, E. and H. Jiménez. 2004b. Experiments in text categorization using term

selection by distance to transition point. *Advances in Computing Science*, 10:139–146.

Moyotl-Hernández, E. and H. Jiménez-Salazar. 2005. Enhancement of dtp feature selection method for text categorization. In Alexander F. Gelbukh, editor, *CICLing*, volume 3406 of *Lecture Notes in Computer Science*, pages 719–722. Springer.

Pinto, D., H. Jiménez-Salazar, and P. Rosso. 2006. 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.

Pinto, D. and F. Pérez. 2004. Una técnica para la identificación de términos multipalabra. In L. Sandoval, editor, *Proceedings of the 2nd National Conference on Computer Science*, pages 257–259. BUAP Press.

Rijsbergen, C. J. Van. 1979. *Information Retrieval, 2nd edition.* Dept. of Computer Science, University of Glasgow.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

Shin, K. and S. Y. Han. 2003. Fast clustering algorithm for information organization. In A. F. Gelbukh, editor, *CICLing*, volume 2588 of *Lecture Notes in Computer Science*, pages 619–622. Springer.

Tovar, M., M. Carrillo, D. Pinto, and H. Jiménez. 2005. Combining keyword identification techniques. *Research on Computing Science*, 14:157–162.

Urbizagástegui, A. R. 1999. Las posibilidades de la ley de zipf en la indización automática. Technical report, Universidad de California, Riverside.

Yang, Y. 1995. Noise reduction in a statistical approach to text categorization. In *Proceedings of SIGIR-ACM*, pages 256–263.

Zipf, G. K. 1949. *Human behaviour and the principle of least effort.* Addison-Wesley.