

Analysis of narrow-domain short texts clustering

A research project report presented

by

David Eduardo Pinto Avendaño

to

The Department of Information Systems and Computation

in partial fulfillment of the requirements

for the obtention of

Diploma of Advanced Studies

(Diploma de Estudios Avanzados)

in the subject of

Pattern Recognition and Artificial Intelligence

Polytechnic University of Valencia

Valencia, Spain

September 2007

©2007 - David Eduardo Pinto Avendaño

All rights reserved.

Author
David Eduardo Pinto Avendaño
PhD Student

Thesis supervisors
Paolo Rosso
Héctor Jiménez Salazar

Analysis of narrow-domain short texts clustering

Abstract

Clustering short texts of narrow domain is considered to be the most difficult clustering problem because of the high term overlapping among the corpus texts and also the low frequencies of these terms. These characteristics lead current term-based techniques to fail when dealing with the above problem. The aim of this research work was to develop methods and techniques in order to tackle the described problem.

The experiments we have conducted have derived the following minor contributions:

1. A similarity measure based on the symmetric Kullback-Leibler distance.
2. One corpus compiled from MEDLINE in the medicine domain, specifically in the *Cancer* domain.
3. The Transition Point (TP) term selection technique.
4. A stabilisation model for the TP term selection technique.
5. A new unsupervised technique for threshold selection in vocabulary reduction

Up to now, the major contributions of the investigations carried out are enumerated as follows:

1. A relative hardness measure which uses the term overlapping among the categories of a supervised corpus
2. An enrichment technique for increasing the term frequencies named self-term expansion
3. A methodology for dealing with narrow-domain short text corpora which uses first self-term expansion and, thereafter, term selection

The new method based on self-term expansion, highly improves results of clustering narrow-domain short texts. Self term expansion means to obtain a thesaurus from the same dataset and then use it for expanding its own terms. Our study also investigates the performance of using the proposed self term expansion when different term selection techniques are employed. We have found that the best combination is to expand first the corpus and then apply a term selection technique. Particularly, when we experimented with a corpus of high energy particles domain (physics), we observed that by using only the term expansion method it is possible to improve the baseline of approximately 40%. Furthermore, by using term selection after expanding the corpus we can obtain a similar performance with a 90% reduction of the full vocabulary.

We have carried out several experiments observing that the clustering of narrow-domain short text corpora is a very challenging task. However, the contributions of this research work are evidence that it is possible to deal with this difficult problem improving the results obtained with typical techniques and methods.

Contents

Title page	i
Abstract	iii
Table of contents	v
Citations to previously published papers	vii
1 Introduction	1
1.1 Clustering	1
1.2 On the relative hardness of clustering corpora	2
1.3 Clustering narrow-domain short text corpora	2
1.4 Applications in other areas of NLP	3
1.5 Overview of the research report	3
2 Methods, techniques and datasets	4
2.1 Clustering methods	4
2.1.1 Hierarchical clustering methods	4
2.1.2 Agglomerative clustering methods	6
2.1.3 Density-based clustering methods	8
2.2 Similarity measures	8
2.2.1 The Jaccard index	9
2.2.2 The TF-IDF measure	9
2.2.3 The Kullback-Leibler Distance	10
2.3 Clustering measures	11
2.3.1 The F -Measure	12
2.3.2 The supervised evaluation measure	12
2.4 Term selection techniques	13
2.4.1 The Transition Point technique	13
2.4.2 The Document Frequency technique	14
2.4.3 The Term Strength technique	15
2.5 Co-occurrence terms	15
2.6 Datasets	16
2.6.1 Narrow-domain short text corpora	16
2.6.2 A new narrow-domain short text corpus	18
2.6.3 Other kind of corpora	22

3	On the relative hardness of clustering corpora	25
3.1	Description of the problem	25
3.2	Calculating the Relative Hardness of a corpus	26
3.3	Clustering the datasets	27
3.4	Correlation between Relative Hardness and F -Measure	27
3.5	Summary	30
3.6	Further work	30
4	Clustering narrow-domain short text corpora	32
4.1	Description of the problem	32
4.2	State of the art	33
4.3	Our research work	33
4.3.1	The role of the term selection process	33
4.3.2	A comparative study of clustering methods	38
4.3.3	A new clustering similarity measure	41
5	The Self-Expansion and Term Selection methodology	45
5.1	Introduction	45
5.2	state of the art	46
5.3	Experiments	46
5.4	Summary and further work	51
6	Applications in other areas of NLP	53
6.1	Introduction	53
6.2	Word sense induction	53
6.2.1	Applying the self-term expansion process	53
6.2.2	Experiments	54
6.2.3	Summary	55
6.2.4	Further work	56
7	Conclusions and further work	57
7.1	Conclusions	57
7.2	Further work	58
7.2.1	Summarization	58
7.2.2	Text clustering by using information retrieval and summarization	59
7.2.3	Fuzzy clustering: The FuzzyMajorClust algorithm	60
	Bibliography	61

Citations to previously published papers

Large portions of Chapters 3, 4, and 6 have appeared in the following papers:

D. Pinto, J. M. Benedí, and P. Rosso. Clustering narrow-domain short texts by using the kullback-leibler distance. In A. F. Gelbukh, editor, *CICLing 2007*, volume 4394 of *Lecture Notes in Computer Science*, pages 611–622. Springer-Verlag, 2007.

D. Pinto, H. Jiménez-Salazar, and P. Rosso. Clustering abstracts of scientific texts using the transition point technique. In A. F. Gelbukh, editor, *CICLing 2006*, volume 3878 of *Lecture Notes in Computer Science*, pages 536–546. Springer-Verlag, 2006 (*Best Student Paper Award*).

D. Pinto, A. Juan, P. Rosso, and H. Jiménez. A comparative study of clustering algorithms on narrow-domain abstracts. *Procesamiento del Lenguaje Natural*, 37(1):43–49, 2006.

D. Pinto and P. Rosso. KnCr: A short-text narrow-domain sub-corpus of medline. In *Proc. of TLH 2006, Advances in Computer Science*, pages 266–269, Colima, Mexico, 2006.

D. Pinto and P. Rosso. On the relative hardness of clustering corpora. In V. Matoušek and P. Mautner, editors, *TSD 2007*, volume 4629 of *Lecture Notes in Artificial Intelligence*, pages 155–161. Springer-Verlag, 2007.

D. Pinto, P. Rosso, and H. Jiménez-Salazar. UPV-SI: Word sense induction using self term expansion. In *Proc. of the 4th International Workshop on Semantic Evaluations - SemEval 2007*, pages 430–433. Association for Computational Linguistics, 2007.

E. Levner, D. Pinto, P. Rosso, D. Alcaide, and R.R.K. Sharma. Fuzzifying clustering algorithms: The case study of MajorClust. In A. F. Gelbukh and A. F. Kuri-Morales, editors, *MICAI 2007, Lecture Notes in Artificial Intelligence*. Springer-Verlag, 2007 (*Accepted to be published*).

Chapter 1

Introduction

1.1 Clustering

Clustering is a very important unsupervised learning method, due to its wide real possible applications. Informally, clustering may be defined as the problem of partitioning a set of elements from a dataset into “clusters”, such that given any element in a cluster, its similarity with each of the elements which belong to the same cluster is greater than the similarity with the elements of other clusters. In other words, clustering deals with finding a structure in a collection of unlabeled data [84].

Clustering is very common for statistical data analysis, and it is used in many research areas, including machine learning, data mining, pattern recognition, image analysis and others. The aim of this research work was to investigate clustering techniques for dealing with short texts from narrow-domain corpora. Clustering of short texts in narrow domains is one of the most difficult tasks due to the high overlapping of vocabularies among the texts and also to the specific terminology used in these corpora. Moreover, this particular task has not received too much attention by the computational linguistic community and, therefore, it is very important to investigate further in this area.

The following is a list of topics we have detected as meaningful to explore for the study of clustering narrow-domain short text corpora:

1. The determination of the hardness of clustering corpora
2. The study of methods and techniques for improving clustering of narrow-domain short text corpora
3. The application of the proposed methods and techniques in different areas of natural language processing

We have conducted a set of experiments and the aim is to accomplish the previous enumerated topics, and, therefore, we have structured this document according to

those. In the following sections we briefly describe the introduced topics and we refer to the corresponding chapter for a detailed explanation.

1.2 On the relative hardness of clustering corpora

The aim of this topic is to determine in some manner (even experimentally) whether a given corpus is difficult to be clustered or not. We expect at some stage to be able to “quantify” the hardness of a narrow-domain corpus from a clustering perspective. We have carried out a set of preliminary experiments by calculating the overlapping vocabulary degree [56, 55]. Our experimental results show that there exists a correlation between the relative hardness formula introduced and the F -Measure, at least with the MajorClust clustering algorithm. However, we must know if this correlation is also kept with other clustering algorithms and other datasets. Additionally, we should experiment by using density measures, such as Density Expected Measure (DEM), in order to determine whether a given corpus has a particular structure inside it or not. The definition of the above measures and a detailed description of the experiments carried out in this topic are given in Chapter 3.

1.3 Clustering narrow-domain short text corpora

We have carried out different experiments for determining the possible strategies that should be used in order to tackle the problem of both, the low frequencies of vocabulary terms and the vocabulary overlapping associated to this particular task.

One of our first research works was presented in [30] and in [29]. We introduced a new technique for keyword selection based on the used of mid-frequency terms; in addition we also used this new technique in the evaluation of a bigger size corpus [51]. The obtained results motivated a comparative study of clustering algorithms which showed that the used technique, named “transition point”, successfully obtains the best results in comparison with the “document frequency” and “term strength” techniques. Moreover, those results shown to be stable upon the use of different clustering algorithms. This suggests that there exists an independence between the feature selection techniques and the clustering methods [53].

Additionally, we have built a new corpus in the specific *Cancer* domain which we have named *KnCr* [54]. This corpus was evaluated in other experiments, such as the one presented in the CICLing conference (see [50]) in which the symmetric Kullback-Leibler distance was used in order to calculate the document similarity from a target clustering corpus. Besides, more recently, in [28], the *KnCr* corpus was used (together with the *CICLing-2002* and *hep-ex* corpora) to show the possible correlation among subjective and objective (i.e., external and internal) clustering measures.

We proposed a self-term expansion technique which expands (enriches) a target corpus by using a set of co-related terms. The experiments carried out demonstrate

how valuable this new technique is [58]. Although the document size use to increase significantly, a term selection technique can be used in order to decrease the size of the document vocabulary. Up to now, we have concluded that the use of term selection techniques is more useful when the enrichment step is carried out first.

These experimental results have also allowed us to introduce a methodology for clustering short-text narrow-domain corpora. To be clear, the methodology suggests to enrich the target clustering corpora expanding each vocabulary term by using its co-occurrence terms calculated over the same target corpora (self-term expansion). Thereafter, a term selection method may be used in order to downsize the expanded corpora vocabulary. Classical similarity measures then may be used in order to calculate the similarity matrix for some selected clustering algorithm. A detailed description of the self-term expansion technique as well as the experiments carried out in this topic are given in Chapter 4.

1.4 Applications in other areas of NLP

The self-term expansion technique, designed explicitly for short text narrow-domain corpora, has been applied to the Word Sense Induction (WSI) task which consists in discriminating from a given set of sentences (related with an ambiguous word), those that share the same sense. The third place obtained in the SemEval competition ([59]) highlights how valuable this simple technique can be in the clustering process. However, the complete evaluation of our methodology is not carried out yet. We should then perform the evaluation of the different approaches which may be derived from the application of different term selection and term expansion techniques. The same evaluation scripts used in the SemEval should be used in order to obtain values which can be easily compared with those results obtained by the other proposed approaches. A detailed description of the experiments carried out in this topic is given in Chapter 6.

1.5 Overview of the research report

The rest of this document is structured as follows. In Chapter 2 we give all the methods, techniques and datasets used in this research work, in order to give the reader a reference of them. Chapters 3, 4, 5 and 6 contain the description of the experiments briefly introduced in the previous sections of this chapter. Finally, in Chapter 7 we draw some conclusions of the research work carried out up to the date of the “Diploma de Estudios Avanzados” (DEA) examination and we discuss the future work for our PhD thesis.

Chapter 2

Methods, techniques and datasets

In this chapter we have brought together definitions of methods and techniques as well as datasets used in the experiments of this research work. Since these items are used around all the investigation, it is helpful to have a definition of each of them. In fact, we will refer to these sections quite often.

2.1 Clustering methods

We have used several clustering methods in the experiments we carried out. In this section we briefly explain how each of them works.

2.1.1 Hierarchical clustering methods

The Single Link Clustering method

Given a set of N documents to be clustered, and a $N \times N$ distance (or similarity) matrix, the basic process presented in [31] hierarchical clustering is this:

1. Start by assigning each item to its own cluster, so that if you have N documents, we now have N clusters, each containing just one item. Let the distances (similarities) between the clusters be equal to the distances (similarities) between the documents they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now we have one cluster less.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N .

Step 3 can be done in different ways, which is what distinguishes *single-link* from other similar approaches, such as *complete-link* and *average-link* clustering. In the *single-link* clustering (also called the connectedness or minimum method), we consider the distance between one cluster and another one to be equal to the shortest distance from any member of one cluster to any member of the other cluster.

The Complete Link Clustering method

In the *complete-link* clustering (also called the diameter or maximum method), the distance between one cluster and another one is considered to be equal to the longest distance from any member of one cluster to any member of the other cluster in Step 3 of the above algorithm [31].

The Lance & Williams recurrence

There exist a special recurrence formula useful in the computation of many hierarchical clustering methods (including the *average-link* one). This formula was proposed by Lance and Williams in 1971 [34]. By means of the Lance and Williams recurrence an infinite number of hierarchical clustering methods can be implemented by using only one generic and simple programme with quadratic spatial and cubic temporal costs.

Formally, let D be a matrix with the distance between clusters (for example, one cluster for each object) and let suppose that we decide to join the i and j clusters. The distance between the joined cluster, ij , and each other cluster, k , can be computed by using the following, named Lance and Williams, recurrence:

$$D_{ij,k} = \alpha D_{i,k} + \alpha D_{j,k} + \beta D_{i,j} + \gamma |D_{i,k} D_{j,k}|$$

where the α , β and γ coefficients depend on the specific selected method. For instance, Table 2.1 shows the coefficients for six hierarchical clustering methods.

Table 2.1: Six hierarchical clustering methods

Method	α	β	γ
Single link	0.5	0	-0.5
Complete link	0.5	0	0.5
Mean	0.5	-0.25	0
Average link	$\frac{n_i}{n_i+n_j}$	0	0
Center	$\frac{n_i}{n_i+n_j}$	$-\frac{n_i n_j}{(n_i+n_j)^2}$	0
Ward	$\frac{n_i+n_k}{n_i+n_j+n_k}$	$-\frac{n_k}{n_i+n_j+n_k}$	0

The Lance & Williams recurrence algorithm is given as follows:

1. Compute the distance matrix between all the clusters.
2. Determine the nearest clusters.
3. Update the distance matrix with the Lance and Williams recurrence.
4. If more than one cluster is left, then go to Step 2.

The EM clustering method

An Expectation-Maximization (EM) algorithm is used in statistics for finding the maximum likelihood estimate of parameters in a probabilistic model, where the model depends on unobserved latent variables. EM alternates between performing an expectation (E) step, which computes an expectation of the likelihood by including the latent variables as if they were observed, and a maximization (M) step, which computes the maximum likelihood estimate of the parameters by maximizing the expected likelihood found in the E step. The parameters found in the M step are then used to begin another E step, and the process is repeated.

The particular implementation used in the experiments is unknown since we executed the corresponding class of the Weka package [80].

2.1.2 Agglomerative clustering methods

The K-Star clustering method

The K-Star clustering method [72] starts by building the similarity matrix of the documents to be clustered (corpus). The algorithm follows as shown in the next steps:

1. It looks for the maximum similarity value in the matrix, $\text{Sim}(D_i, D_j)$, and constructs a cluster (C_i) made up by the two documents this similarity value refers to. It marks these documents (D_i and D_j) as assigned.
2. For each unassigned document (D_k)
 - If $\text{Sim}(D_k, D_i) < \tau$, where τ is a given threshold, then add D_k to cluster C_i and mark D_k as assigned.
3. Return to Step 1

In our particular case, we have used a canonic threshold defined as the average of the values in the similarity matrix.

The NN1 clustering method

The NN1 clustering algorithm [30] is a variation of the K-Star method. It differs in the manner it calculates the similarity of unassigned documents with the corresponding cluster. The NN1 algorithm uses the average of similarities and, therefore, it is more expensive in computational time than *K*-Star.

The *K*-NN clustering method

The *K*-Nearest Neighbour clustering algorithm, often simply known as *K*-NN [21], is among the simplest of all machine learning algorithms. An object is classified by a majority vote of its neighbours, with the object being assigned the most common class among its *k* nearest neighbours, where *k* is a positive integer, typically small. If *k* = 1, then the object is simply assigned the class of its nearest neighbour. In binary (two classes) classification problems, it is helpful to choose *k* to be an odd number as this avoids difficulties with tied votes.

The *K*-Means clustering method

The widely known *K*-Means algorithm assigns each object to the cluster whose center is nearest. The center is the average of all the points of the cluster. That is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster. The algorithm steps are ([39]):

1. Choose the number *K* of clusters.
2. Randomly generate *K* clusters and determine the cluster centers, or directly generate *K* random points as cluster centers.
3. Assign each point to the nearest cluster center.
4. Recompute the new cluster centers.
5. Repeat the two previous steps until some convergence criterion is met (usually that the assignment has not changed).

The main advantages of this algorithm are its simplicity and speed which allows it to run on large datasets. Its disadvantage is that it does not yield the same result with each run, since the resulting clusters depend on the initial random assignments. Therefore, it does not ensure that the result has a global minimum of variance.

2.1.3 Density-based clustering methods

The MajorClust clustering method

MajorClust executes iterative propagation of nodes into clusters according to the “maximum attraction wins” principle [75]. The algorithm starts by assigning each object in the initial set its own cluster. Within the following re-labelling steps, an object adopts the same cluster label as the “weighted majority of its neighbours”. If several such clusters exist, one of them is randomly chosen. The algorithm terminates if no object changes its cluster membership.

The MajorClust is a relatively new clustering algorithm with respect to other methods. However, its characteristic of automatically discovering the target number of clusters makes it very attractive [74, 4, 56, 48].

The MajorClust algorithm

Input: object set D , similarity measure $\varphi : D \times D \rightarrow [0; 1]$, similarity threshold τ .

Output: function $\delta : D \rightarrow N$, which assigns a cluster label to each point.

1. $i := 0$, ready := false
2. for all p from D do $i := i + 1$, $\delta(p) := i$ enddo
3. while ready = false do
 - (a) ready := true
 - (b) for all q from D do
 - i. $\delta^* := i$ if $\Sigma\{\varphi(p, q) | \varphi(p; q) \geq t \text{ and } \delta(p) = i\}$ is maximum.
 - ii. if $\delta(q) \neq \delta^*$ then $\delta(q) := \delta^*$, ready := false
 - (c) enddo
4. enddo

Remark. The similarity threshold τ is not a problem-specific parameter but a constant that serves for noise filtering purposes. Its typical value is 0.3.

2.2 Similarity measures

The clustering methods usually work with a similarity matrix calculated in advance. They do not care how this matrix is calculated, since they perform the clustering process assuming that the matrix has been calculated in some way. In the following subsections we explain a set of similarity measures used in the experiments carried out during our research work.

2.2.1 The Jaccard index

The Jaccard coefficient is a statistical measure used in natural language processing for comparing the similarity of a couple of documents. It is defined as the size of the intersection divided by the size of the union of the sample texts. Given two documents, X_1 and X_2 , the Jaccard coefficient is a useful measure of the overlap that X_1 and X_2 share with their words. Formally,

$$Jaccard(X_1, X_2) = \frac{|X_1 \cap X_2|}{|X_1 \cup X_2|} \quad (2.1)$$

2.2.2 The TF-IDF measure

The Term Frequency and Inverse Document Frequency *TF-IDF* is a statistical measure of weight often used in natural language processing to measure how important a word is to a document in a corpus. The importance of each word increases proportionally to the number of times a word appears in the document (frequency) but is offset by the frequency of the word in the corpus. In this document, we will refer to the *TF-IDF* as the complete similarity process of using the *TF-IDF* weight and a special similarity measure proposed by Salton [69] for the Vector Space Model, which is based on the use of the cosine among vectors representing the documents.

The TF component of the formula is calculated by the normalized frequency of the term, whereas the IDF is obtained by dividing the number of documents in the corpus by the number of documents which contain the term, and then taking the logarithm of that quotient. Given a corpus C and a document D_j ($D_j \in C$), the *TF-IDF* value for a term t_i in D_j is obtained by the product among the frequency of the term t_i in the document D_j (TF_{ij}) and the inverse document frequency of the term in the corpus (IDF_i) as follows.

$$TF_{ij} = \frac{Freq(t_i)}{\sum_{k=1}^{|D_j|} Freq(t_k)}$$

$$IDF_i = \log \left(\frac{|C|}{|D : t_i \in D, D \in C|} \right)$$

$$TF-IDF = TF_{ij} \times IDF_i$$

Each document can be represented by a vector where each entry corresponds to the *TF-IDF* value obtained by each vocabulary term of the given document. Thus, given two documents in vectorial representation, D_1 and D_2 , it is possible to calculate the cosine of the angle between these two vectors as follows:

$$\cos \theta = \frac{D_1 \times D_2}{\|D_1\| \times \|D_2\|}$$

2.2.3 The Kullback-Leibler Distance

In 1951 Kullback and Leiber studied a measure of information from the statistical viewpoint; this measure involved two probability distributions associated with the same experiment [33]. The Kullback-Leibler (KL) 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 2.2.

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)} \quad (2.2)$$

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 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 the experiments of this research work. Each KLD corresponds to the definition of Kullback and Leiber [33], Bigi [10], Jensen [24], and Bennet [7] [86], respectively.

$$D_{KLD1}(P||Q) = D_{KL}(P||Q) + D_{KL}(Q||P) \quad (2.3)$$

$$D_{KLD2}(P||Q) = \sum_{x \in X} (P(x) - Q(x)) \log \frac{P(x)}{Q(x)} \quad (2.4)$$

$$D_{KLD3}(P||Q) = \frac{1}{2} \left[D_{KL} \left(P || \frac{P+Q}{2} \right) + D_{KL} \left(Q || \frac{P+Q}{2} \right) \right] \quad (2.5)$$

$$D_{KLD4}(P||Q) = \max(D_{KL}(P||Q) + D_{KL}(Q||P)) \quad (2.6)$$

KL and KLD have been used in many natural language applications like query expansion [16], language models [12], and categorization [10]. They have also been used, for instance, in speech processing applications based on statistical language modeling [18], and in information retrieval, for topic identification [11].

We have considered to calculate the corpus document similarities in an inverse function with respect to the distance defined in Equations (2.3), (2.4), (2.5), or (2.6).

In the text clustering model proposed in this research work, a document D_j is represented by a term vector of probabilities \vec{D}_j and the distance measure is, therefore, the KLD (the symmetric Kullback-Leibler Divergence) between a pair of documents \vec{D}_i and \vec{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 a very small probability (*epsilon*- ϵ), 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 document represented as D_j . For the terms not in $V(D_j)$, 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:

$$P(t_k, D_j) = \begin{cases} \beta * P(t_k|D_j), & \text{if } t_k \text{ occurs in the document } D_j \\ \varepsilon, & \text{otherwise} \end{cases} \quad (2.7)$$

with:

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

where: $P(t_k|D_j)$ is the probability of the term t_k to be in the document D_j , β is a normalization 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 2.7 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:

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

2.3 Clustering measures

The quality of clustering results is often referred as “validity of document clustering” [43]. The role of the task of measuring the quality of the obtained clusters is to reflect the human idea of best classification. Basically, two are the validity indices took into account: internal and external (often also called objective and subjective). The former validity indices allow to decide whether the obtained clusters are well developed with respect to the structural properties of the target clustering corpora. Whereas the latter judge compares the obtained clusters with respect to the gold standard, i.e., a classification given by an expert. In the following subsections, we present a set of external clustering measures used in our research work.

2.3.1 The F -Measure

F -Measure is an external clustering measure which compares the clusters obtained by some clustering method with respect to the classification given by an expert. The latter classification is usually referred as the “set of classes”. Formally, given a set of clusters $\{G_1, \dots, G_m\}$ and a set of classes $\{C_1, \dots, C_n\}$, the F -Measure between a cluster i and a class j is given by the following formula.

$$F_{ij} = \frac{2 \cdot P_{ij} \cdot R_{ij}}{P_{ij} + R_{ij}}, \quad (2.8)$$

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

$$P_{ij} = \frac{\text{Number of texts from cluster } i \text{ in class } j}{\text{Number of texts from cluster } i} = \frac{|G_i \cap C_j|}{|G_i|}, \quad (2.9)$$

and

$$R_{ij} = \frac{\text{Number of texts from cluster } i \text{ in class } j}{\text{Number of texts in class } j} = \frac{|G_i \cap C_j|}{|C_j|} \quad (2.10)$$

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 (2.11).

$$F = \sum_{1 \leq i \leq m} \frac{|G_i|}{|D|} \max_{1 \leq j \leq n} F_{ij}. \quad (2.11)$$

2.3.2 The supervised evaluation measure

The supervised evaluation measure is performed as described in [1]. First, the corpus is splitted into a train/test part. Using the hand-annotated classes information in the train part, it is possible to compute a mapping matrix M that relates clusters and classes in the following way. Let suppose that there are m clusters and n classes for the target document. Then, $M = \{m_{ij}\}$ $1 \leq i \leq m$, $1 \leq j \leq n$, and each $m_{ij} = P(s_j|h_i)$, that is, m_{ij} is the probability of a document belonging to class j to be assigned to the cluster i . This probability can be computed by counting the times an occurrence with class s_j has been assigned to the cluster h_i in the train corpus.

The mapping matrix is used to transform any cluster score vector $\vec{h} = (h_1, \dots, h_m)$ returned by the clustering algorithm into a class score vector $\vec{s} = (s_1, \dots, s_n)$. It suffices to multiply the score vector by M , i.e., $\vec{s} = \vec{h}M$.

We use the M mapping matrix in order to convert the cluster score vector of each test corpus instance into a class score vector, and assign the class with maximum score to that instance. Finally, the resulting test corpus is evaluated according to the usual precision and recall measures for supervised clustering systems.

2.4 Term selection techniques

It is well-known that only those features which help to discriminate should be included in the clustering process. In fact, the addition of very few irrelevant features can lead to obtain bad results [22] [45]. Up to now, different Term Selection Techniques (TSTs) have been used in the clustering task; however, clustering short texts in a narrow-domain often 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 TST framework, a very interesting research work has been carried out from several decades by using testors [35]. A testor is a set of features which may be used to represent a dataset. A testor is named irreducible (typical) if none of its proper subsets is a testor. Although this theory may be adequate for selecting terms in a collection, it lacks of algorithms for efficient calculation of the testor set. In fact, in [70] the fastest algorithm, which is not polynomial in complexity, was presented. Some works such as the one presented by Pons-Porrata *et al.* [60] employed text mining by using testors as a term selection technique. In our research work, we have considered to use other TSTs which can be efficiently executed with large datasets. In the next subsections we will briefly describe each technique employed in our experiments: Transition Point (TP), Document Frequency (DF) and, Term Strength (TS). The first two unsupervised techniques have demonstrated their value in the clustering task [38], whereas the third TST has been especially used in text categorization [81] [49]. The TP technique is a simple calculation procedure that 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 [51, 66]). Therefore, we believe that there exists enough evidence to use it as a term selection technique. 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. We will use unsupervised term selection techniques because we want to avoid the use of hand-tagged external resources which are expensive (in time) to be created.

2.4.1 The Transition Point technique

The Transition Point is a frequency value that splits the vocabulary of a document into two sets of terms of low and high frequency. This technique is based on the Zipf law of word occurrences [85] and also on the refined studies of Booth [13], as well as Urbizagástegui [77]. These studies are meant to demonstrate that mid-frequency terms are closely related to the conceptual content of a document. Therefore, it is possible to assume that those terms whose frequencies are closer to the TP may be used as indexes of a document. A typical formula used to obtain this value is given

in Equation (2.12).

$$TP_V = \frac{\sqrt{8 * I_1 + 1} - 1}{2} \quad (2.12)$$

where I_1 represents the number of words with frequency equal to 1 in a given text T [47] [77]. 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 [13].

Let be t_i the i -th term of the document D and tf_i the term frequency of that term. Then consider the frequency-sorted vocabulary of D ; i.e.,

$$V_D = [(t_1, tf_1), \dots, (t_n, tf_n)],$$

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

$$V_{TP} = \{t_i | (t_i, tf_i) \in V_D, U_1 \leq tf_i \leq U_2\}, \quad (2.13)$$

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

For our experiments, the weight of each term t is calculated inversely proportional with respect to the distance between its frequency and the TP frequency of D , named TP_D . The following equation shows how to obtain this value:

$$IDTP(t, D) = \frac{1}{|TP_D - tf(t, D)| + 1}, \quad (2.14)$$

where $tf(t, D)$ is the frequency of the term t in the document D .

2.4.2 The Document Frequency technique

Document Frequency is an effective and simple technique which has shown to obtain comparable results to the classical supervised techniques such as χ^2 and Information Gain [82]. This technique assigns the value $DF(t)$ to each term t , where $DF(t)$ means the number of texts, in a collection, where t occurs. 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 of a text. This technique is based in the fact that rare terms are not valuable for determining the target cluster of some document. Therefore, by extracting those terms from the vocabulary we will obtain a dimensionality reduction of the vocabulary. DF is an easy TST that may be used in large-sized corpora, due to its complexity which is approximately linear in the number of the dataset documents.

2.4.3 The Term Strength technique

This technique was first introduced in [78] in order to improve the performance of document retrieval by using TSTs. This method takes into account that the most valuable terms in a collection are those which are shared by related documents. Therefore, the weight of a term is calculated as the probability of finding it in the document T_i given that it has also appeared in the document T_j which is as similar to T_i as a given threshold ($\text{sim}(T_i, T_j) \geq \beta$). The weight given to each term t is then defined by the following equation:

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

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 of the term strength technique can be found in [81] and [49].

2.5 Co-occurrence terms

In literature, the calculation of co-occurrence terms is one of the most common technique used for the automatic construction of Lexical Data Bases (LDB) [25, 23]. On the one hand, a simple approach may use n -grams, which allow us to predict a word from previous words in a sample of text. The frequency of each n -gram is calculated and then filtered according to some threshold. The resulting n -grams constitute a LDB which may be used as an “expansion dictionary” for each term. On the other hand, an information theory-based co-occurrence measure is discussed in [42]. This measure is named pointwise Mutual Information (MI), and its applications for finding collocations are analysed by determining the co-occurrence degree among two terms. This may be done by calculating the ratio between the number of times that both terms appear together (in the same context and not necessarily in the same order) and the product of the number of times that each term occurs alone. Given two terms x_1 and x_2 , the pointwise mutual information between x_1 and x_2 can be calculated as follows:

$$MI(x_1, x_2) = \log_2 \frac{P(x_1 x_2)}{P(x_1) \times P(x_2)}$$

For instance, if we know the words “Computer” and “Science” occurs, respectively, with frequency 70 and 60, and the term “Computer Science” occurs with frequency 20 in a corpus of 15,000,000 tokens, then we may calculate the pointwise mutual information among these two words as:

$$MI(\text{Computer}, \text{Science}) = \log_2 \frac{\frac{20}{15,000,000}}{\frac{70}{15,000,000} \times \frac{60}{15,000,000}} \approx 16.12$$

The numerator could be modified in order to take into account only bigrams, as presented in [51], where an improvement of clustering short texts in narrow domains (i.e. domains with a high degree of overlapping between their vocabularies) has been obtained. We determined that the single occurrence of each term should be at least three as Manning did (see [42]), whereas the maximum separation among the two terms was selected as five.

2.6 Datasets

In the experiments we have carried out, corpora with different characteristics with respect to their size and their balance degree were used. We consider both, narrow and wide domain clustering corpora. The aim was to compare the performance of the algorithms in different kinds of corpora. We have preprocessed all these collections by eliminating stop words and by applying the Porter stemmer [61]. In the following subsections we describe each corpus into detail. The characteristics given in the below tables for each corpus were obtained after applying this preprocessing phase.

2.6.1 Narrow-domain short text corpora

The *CICLing-2002* corpus

This corpus is made up by 48 abstracts from the *Computational Linguistics* domain, which corresponds to the *CICLing 2002* conference. This collection was used first by Makagonov et al. [40] in their experiments on clustering abstracts of narrow domain. 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 (word sense disambiguation, anaphora, part of speech tagging, 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 is shown in Tables 2.2 and 2.3, respectively.

Table 2.2: Distribution of the *CICLing-2002* corpus

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

Table 2.3: 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

The *hep-ex* corpus of CERN

This corpus is based on the collection of abstracts compiled by the University of Jaén, Spain named *hep-ex* [46]. It is composed of 2,922 abstracts from the *Physics* domain originally stored in the data server of the CERN¹.

The distribution of the categories for each corpus is better described in Table 2.4; other characteristics are shown in Table 2.5. As can be seen, this corpus is totally unbalanced, which makes our task even more challenging.

Table 2.4: 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

The WSI-SemEval corpus

This corpus was provided by the organizers of the “Evaluating Word Sense Induction and Discrimination Systems” task of the SemEval 2007 workshop. The dataset consists of 100 ambiguous words (65 verbs and 35 nouns) borrowed from the “English lexical sample” task of the same workshop. The documents come from the Wall Street Journal corpus, and they were annotated manually with OntoNotes senses [3]. The name of the ambiguous words (verbs and nouns) along with the number of their

¹Centre Européen pour la Recherche Nucléaire

Table 2.5: 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

instances are presented in Table 2.6. A set of average values of the characteristics of this corpus is given in Table 2.7.

2.6.2 A new narrow-domain short text corpus

The absence of a specific forum for the evaluation of systems for the clustering narrow-domain short texts task, has not allowed to create a good corpus for using it as a standard evaluation. We made the effort of constructing a new narrow-domain short text corpus in the medicine domain, by downloading the last sample of documents provided by MEDLINE². This sample dataset contains approximately 30,000 abstracts, and we selected those related with the “Cancer” domain. We have named the new corpus as *KnCr* [54]. This corpus has been used in some experiments, such as the one presented at the CICLing 2007 conference (see [50]). More recently, in [28], the KnCr corpus was used (together with the *CICLing-2002* and *hep-ex* corpora) to show the possible correlation among subjective and objective (i.e. external and internal) clustering measures. This corpus was created for the specific task of clustering short texts of a medical narrow-domain [54]. It consists of 900 abstracts related with the “Cancer” domain.

Below we will explain how we have created the gold standard for this new corpus.

Automatic gold standard generation

In order to correctly evaluate the results of clustering, a corpus must be provided with a gold standard of the possible clustering classes distribution. Although the gold standard is normally constructed by humans, we tried to create it automatically.

Due to the fact that each retrieved abstract of our document set contains “keywords” provided by each author, we used them for constructing it. We selected three clustering methods for this experiment, two already implemented in the Weka machine learning software [80] (Expectation Maximization and K-Means), and *K-Star*.

²<ftp://ftp.nlm.nih.gov/nlmdata/sample/medline/>

Table 2.6: The ambiguous words of the *WSI-SemEval* corpus

Word	instances	Word	instances	Word	instances
president	1056	explain	103	turn	402
chance	106	announce	108	ask	406
authority	111	cause	120	complain	46
base	112	kill	127	improve	47
rate	1154	remember	134	propose	48
carrier	132	hope	136	attempt	50
defense	141	allow	143	purchase	50
condition	166	hold	153	contribute	53
source	187	end	156	regard	54
network	207	produce	159	express	57
effect	208	begin	162	complete	58
development	209	report	163	promise	58
job	227	build	165	replace	61
hour	235	raise	181	affect	64
drug	251	receive	184	recall	64
power	298	find	202	remove	64
share	3061	lead	204	approve	65
position	313	buy	210	claim	69
move	317	see	212	disclose	69
management	329	set	216	occur	69
capital	335	come	229	enjoy	70
area	363	grant	24	avoid	71
policy	370	need	251	maintain	71
value	394	start	252	prove	71
order	403	believe	257	prepare	72
plant	411	do	268	exist	74
exchange	424	say	2702	care	76
future	496	work	273	describe	76
bill	506	examine	29	join	86
system	520	go	305	estimate	90
part	552	fix	34		
point	619	negotiate	34		
state	689	keep	340		
space	81	rush	35		
people	869	feel	398		

(a) Nouns

(b) verbs

(c) verbs

A description of the clustering methods can be seen in Section 2.1. We used the F -Measure (see Section 2.11) for comparing each pair of clustering methods.

Table 2.7: Other features of the *WSI-SemEval* corpus

Feature	Value
Size of the corpus (bytes)	10,644,648
Number of ambiguous words	100
Number of sentences	27,132
Total number of terms	1,555,960
Vocabulary size (terms)	27,656
Average number of sentences (instances)	271.32
Average vocabulary size	47,65
Term average per sentence	57.34

The obtained results are given in Table 2.8. None pair combination of clustering methods obtained more than 0,51 of F-Measure and it was not possible to determine a more successful clustering method for constructing the gold standard. This experiment has shown that clustering narrow-domain corpora is really a difficult task, eventhought we have available the keywords of each abstract.

Table 2.8: Results obtained by clustering abstract keywords (evaluation without gold standard)

	EM	K-Means	K-Star
EM	–	0,51	0,45
K-Means	0,31	–	0,36
K-Star	0,36	0,33	–

Manual gold standard generation

Once obtained the previous results and in order to construct manually the gold standard, we had had manual inspection for classifying every document in its correct class. We used the ontology made available by the National Cancer Institute (NCI)³, for the construction of the gold standard categories. This ontology describes a hierarchy of cancer terms based in the anatomy kind and specifies the fine grain categories of this domain (the current owl version of the NCI thesaurus can be found in <http://www.mindswap.org/2003/CancerOntology/>). Tables 2.9 and 2.10 show the complete characteristics of this new cancer corpus. As can be seen, only 900 from 30,000 abstracts are related with the cancer topic, and the average length of each of them is about 126 words which makes it suitable for experiments in the narrow-domain short text corpora clustering task.

³<http://ncimeta.nci.nih.gov/>

Table 2.9: 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 2.10: 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

Once constructed the gold standard, we carried out some experiments to compare different methods of clustering against it, in order to investigate the hardness of clustering the texts of this corpus. We implemented two hierarchical clustering methods, namely Single and Complete Link Clustering, and three agglomerative clustering methods (*K*-NN, *K*-Star, NN1). A description of these clustering methods is also included in Section 2.1. The results obtained by clustering the abstracts instead of the keywords, and by using two well known vocabulary reduction techniques (Document Frequency and Term Strength), are presented in Table 2.11. We can observe low *F*-Measure values for each clustering method, which highlights again the hardness of this task.

Table 2.11: Results obtained by clustering abstracts (evaluation with the gold standard)

	DF	TS
K-Star	0,39	0,39
SLC	0,52	0,51
CLC	0,36	0,36
NN1	0,42	0,41
K-NN	0,38	0,37

In order to verify whether the clustering by keywords, provided by the abstract authors, behaves better than when using the vocabulary reduction techniques presented above, we carried out a third experiment. In this case we compared the results obtained by clustering those keywords with EM, KMeans and KStar methods with the gold standard manually built. The results are presented in Table 2.12. We may see that by using keywords instead of abstracts can lead to more confusion in the clustering narrow-domain short texts task. This may be due to the different viewpoints of scientific text author, and the few words added as keywords. That is, a little variation in the keyword set leads to classify similar documents in different classes.

Table 2.12: Comparison against the gold standard of clustering abstracts keywords

	<i>F</i> -Measure
EM	0,20
K-Means	0,22
K-Star	0,22

As a consequence of the few research works in clustering short text of narrow domains, there exist a lackness of this type of corpora that led us to compile scientific abstracts from high quality sources. We have selected MEDLINE as a repository source for the construction of a new corpus in the cancer domain. Our corpus is a moderate sized one, with 900 abstracts and 16 different balanced categories.

In order to investigate the possible hardness of clustering this corpus, we have carried out a set of experiments. First we tried to construct automatically the gold standard by comparing three different clustering methods upon the use of the keywords of each abstract. Due to the difficulty to evaluate the goodness of the automatically obtained gold standard, we decided to obtain it manually. Moreover, we compared the results of clustering keywords against clustering abstracts (using a vocabulary reduction), and in this particular case we found that author keywords may confuse the clustering process.

We have made free available this new corpus by email request to the authors. We consider that this corpus, together with its gold standard, will allow to test clustering algorithms on short texts of the cancer narrow domain.

2.6.3 Other kind of corpora

We have also used short texts corpora different than those which are narrow-domain. The goal is to analyse the performance obtained by employing both, narrow and wide domain corpora. Following, we describe each corpus into detail.

Reuters

Reuters-21578⁴ has been extensively used for categorization tests. The most recent version of Reuters is distributed as Reuters RCV1 and RCV2. In the experiments we have carried out, we have used clustering algorithms which assign each document to exactly one cluster and, therefore, we have used the R8 subcollection of the Reuters-21578 since it is a single-categorized dataset. The characteristics of this corpus are given in Tables 2.13 and 2.14. Each table shows the “Train” and “Set” subset features of this collection.

Table 2.13: Categories of the *R8-Reuters* corpus

Category	# of documents	Category	# of documents
trade	102	trade	319
grain	34	grain	78
monex-fx	130	monex-fx	366
crude	140	crude	314
interest	87	interest	202
acq	707	acq	1608
ship	43	ship	121
earn	1076	earn	2831

(a) Train

(b) Test

Table 2.14: Other features of the *R8-Reuters* corpus

Feature	Test	Train
Size of the corpus (Bytes)	912,553	2,567,683
Number of categories	8	8
Number of documents	2,319	5,839
Total number of terms	150,430	416,431
Vocabulary size (terms)	9,315	15,648
Term average per document	64.87	71.32

20 Newsgroups

20 Newsgroups⁵ is a well-known collection which has been used for benchmarking clustering algorithms. The corpus is made up by 20 different newsgroups (electronic mails), each corresponding to a different topic. Some of the newsgroups are very closely related to each other, whereas others are highly unrelated. In order to carry out preliminary experiments, we have used a small version of 20 Newsgroups which

⁴<http://www.daviddlewis.com/resources/testcollections/reuters21578/>

⁵<http://people.csail.mit.edu/jrennie/20Newsgroups/>

we have named *Mini20Newsgroups*. The characteristics of this corpus are given in Tables 2.15 and 2.16.

Table 2.15: Categories of the *Mini20Newsgroups* corpus

Category	documents	Category	documents
alt_atheism	100	misc_forsale	100
comp_graphics	100	rec_autos	100
comp_os_mswindows_misc	100	rec_motorcycles	100
comp_sys_ibm_pc_hardware	100	rec_sport_baseball	100
comp_sys_mac_hardware	100	rec_sport_hockey	100
comp_windows_x	100	soc_religion_christian	100
sci_crypt	100	talk_politics_guns	100
sci_electronics	100	talk_politics_mideast	100
sci_med	100	talk_politics_misc	100
sci_space	100	talk_religion_misc	100

Table 2.16: Other features of the *Mini20Newsgroups* corpus

Feature	Value
Size of the corpus (Bytes)	1,909,435
Number of categories	20
Number of documents	2,000
Total number of terms	290,067
Vocabulary size (terms)	23,509
Term average per document	145.03

Chapter 3

On the relative hardness of clustering corpora

3.1 Description of the problem

As described in the first chapter, clustering deals with finding a structure in a collection of unlabeled data [84]. When dealing with raw text corpora, the discovering of their most appropriate features can help on the selection of methods and techniques for determining the possible intrinsic grouping in those sets of unlabeled data. Therefore, the study of the characteristics of a given corpus should be of high benefit. As far as we know, research works in this field nearly have not been carried out in literature. We found just one attempt for determining the relative hardness of the Reuters-21578 clustering collection [19], but this research work neither derived formulae for determining the hardness of these corpora nor the possible set of features that are involved in the clustering hardness. A related work which could be considered in order to observe the hardness of a given corpus (with respect to a specific clustering algorithm) is partially presented in [44] and [43]. In these research works, the author discusses internal clustering quality measures, such as the one from the Dunn Index family, which showed to perform well in the experiments presented by Bezdek et al. in [9, 8].

Reuters-21578 (now Reuters RCV1 and RCV2) and 20 Newsgroups are well-known collections which have been used for benchmarking clustering algorithms. However, the fact that several clustering methods may obtain bad results over those corpora does not necessarily imply that they are difficult to be clustered. Further investigation needs to be done in order to determine whether the current clustering corpora are easy clustering instances or not.

We are interested in investigating two aspects:

1. A set of possible features hypothetically related with the hardness of the clustering task, and
2. the definition of a formula for the easy evaluation of the relative hardness of a given clustering corpus.

We empirically know that at least three components are involved:

1. The size of the clustering texts,
2. The broadness of the corpora domain, and
3. Whether the documents are single or multi categorized.

The preliminary experiments were carried out by using three different corpora: the R8 version of the Reuters collection (train and test) and, partially, a reduced version of the 20 Newsgroups, named *Mini20Newsgroups*. We have pre-processed each corpus eliminating punctuation symbols, stopwords and, thereafter, applying the Porter stemmer. The characteristics of each corpus after the pre-processing are given in Tables 2.13, 2.14, 2.15 and 2.16 of Section 2.6.

The rest of this chapter is structured as follows. In Section 3.2 we introduce the used formula and the employed approach to split the corpus in order to calculate the relative hardness for all the possible combinations of two or more categories. The Section 3.3 shows the manner we have evaluated the alternative clustering process used for obtaining the correlation with the relative hardness proposed formula. Section 3.4 shows the experimental results we obtained. Finally, some conclusions are drawn and the necessary further work to be done is discussed.

3.2 Calculating the Relative Hardness of a corpus

In order to determine the Relative Hardness (RH) of a given corpus, we have considered the vocabulary overlapping among the texts of the corpus. In our experiments, we have used the well-known Jaccard coefficient for calculating the overlapping (see Section 2.2.1). We considered all the possible combinations of more than two categories from the corpus and for each of them we calculated its RH. For instance, for a given corpus of n categories, $2^n - (n + 1)$ possible subcorpora will be obtained: e.g. for the *R8-Reuters* corpus (eight categories) we obtained 247 subsets.

Thereafter, we calculated their RHs as follows: given a corpus C_i made up of n categories (CAT), the RH of $C_i = \{CAT_1, CAT_2, \dots, CAT_n\}$ is:

$$RH(C_i) = \frac{1}{n(n-1)/2} \times \sum_{j,k=1;j < k}^n \text{Similarity}(CAT_j, CAT_k) \quad (3.1)$$

where the similarity among categories is obtained by using the Jaccard coefficient in order to determine their overlapping (see Eq. (3.2)). However, more sophisticated measures also could be used, such as the well-known *TF-IDF* or the one presented in [32] in the plagiarism degree calculation framework. The similarity measure used with the categories of the corpus can be seen as follows.

$$\text{Similarity}(CAT_j, CAT_k) = \frac{|CAT_j \cap CAT_k|}{|CAT_j \cup CAT_k|} \quad (3.2)$$

In the proposed formulae we have considered each category j as the “document” obtained by concatenating all the documents belonging to the category j .

3.3 Clustering the datasets

In order to evaluate the relative hardness formula used in the experiments, we have carried out an unsupervised clustering of all the documents of each subcorpus obtained for each dataset. We have chosen the MajorClust clustering algorithm [75] due to its peculiarity of taking into account both, the inside and outside similarities among the clusters obtained during its execution. In order to keep independent the validation with respect to RH, we have used the *TF-IDF* formula for calculating the input similarity matrix for MajorClust. A better explanation of the *TF-IDF* formula can be seen in Section 2.2.2, whereas the description of the MajorClust clustering algorithm is presented in Section 2.1.3. Each evaluation was performed with the *F*-Measure formula which is calculated as was shown in Section 2.3.1.

3.4 Correlation between Relative Hardness and *F*-Measure

Our preliminary experiments were carried out on the train and test version of the Reuters R8 collection and, partially, also on the *Mini20Newsgroups* dataset. In Figure 3.1 we may see the possible correlation between the relative hardness of the (i) train and (ii) test versions of the R8 collection of Reuters (*R8-Reuters*) with respect to the *F*-Measure obtained by using the MajorClust clustering algorithm. The smaller is the value of RH (x -axis) the higher is the obtained *F*-Measure (y -axis) and viceversa for both corpora. The relative hardness vs. *F*-Measure correlation was

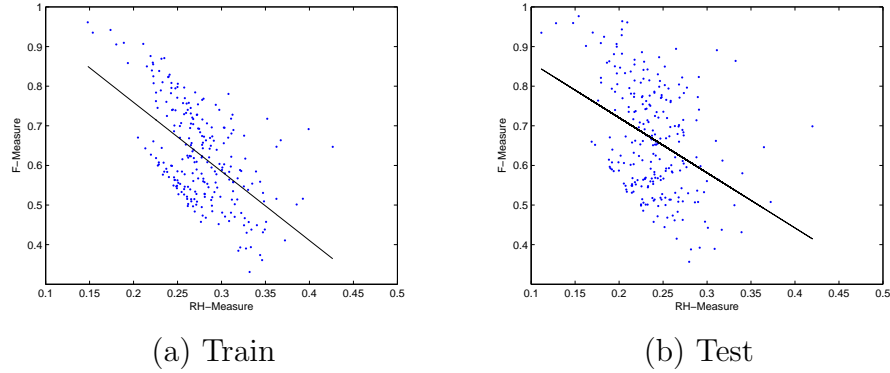
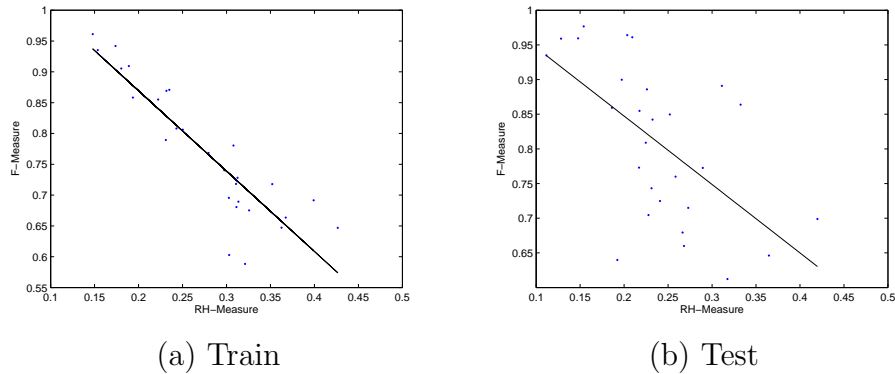


Figure 3.1: Evaluation of all R8 subcorpora (more than two categories per corpus)

Figure 3.2: Evaluation of single pairs of the *R8-Reuters* categories

calculated for all possible corpora variants of *R8-Reuters* (247). In order to easily visualize the correlation between RH and *F*-Measure, we have plotted the polynomial approximation of degree one.

In Figure 3.2 we may see the possible correlation between the relative hardness of each pair of categories of the *R8-Reuters* collection and the *F*-Measure obtained again by using the MajorClust clustering algorithm. The same conclusion is obtained: the smaller is the value of RH (x -axis) the higher is the obtained *F*-Measure (y -axis) and viceversa.

In order to fully appreciate the RH formula, the most and least related pairs of categories for the *R8-Reuters* dataset are presented in Tables 3.1 and 3.2, respectively. The RH value associated with each pair was calculated with the same formula presented in Section 3.2. Some preliminary experiments were carried out also with the *Mini20Newsgroups* dataset and the most and least related pairs of categories are shown in Tables 3.3 and 3.4, respectively.

Table 3.1: The most related categories of the *R8-Reuters* collection

RH value	Category	Category	RH value	Category	Category
0.426	trade	monex-fx	0.419	monex-fx	interest
0.399	monex-fx	interest	0.364	trade	monex-fx
0.367	trade	crude	0.332	trade	interest
0.362	monex-fx	crude	0.317	trade	crude
0.352	trade	interest	0.311	monex-fx	crude

(a) Train

(b) Test

Table 3.2: The least related categories of the *R8-Reuters* collection

RH value	Category	Category	RH value	Category	Category
0.188	interest	earn	0.186	interest	acq
0.180	acq	ship	0.154	ship	earn
0.173	ship	earn	0.147	acq	ship
0.153	grain	acq	0.128	grain	earn
0.147	grain	earn	0.111	grain	acq

(a) Train

(b) Test

Table 3.3: The most related categories of the *Mini20Newsgroups* collection

RH value	Category	Category
0.3412	talk politics guns	talk politics misc
0.3170	alt atheism	talk religion misc
0.3092	talk politics guns	talk religion misc
0.3052	talk politics misc	talk religion misc
0.3041	soc religion christian	talk religion misc
0.2988	sci crypt	talk politics guns
0.2985	soc religion christian	talk politics misc
0.2958	soc religion christian	talk politics guns
0.2932	talk politics mideast	talk politics misc
0.2905	sci electronics	sci space
0.2868	comp sys ibm pc hardware	comp sys mac hardware

Table 3.4: The least related categories of the *Mini20Newsgroups* collection

RH value	Category	Category
0.1814	comp os mswindows misc	rec sport hockey
0.1807	misc forsale	talk politics misc
0.1804	misc forsale	talk religion misc
0.1803	comp sys ibm pc hardware	talk politics mideast
0.1798	comp os mswindows misc	talk religion misc
0.1789	alt atheism	comp os mswindows misc
0.1767	alt atheism	misc forsale
0.1751	misc forsale	soc religion christian
0.1737	comp os mswindows misc	soc religion christian
0.1697	misc forsale	talk politics mideast
0.1670	comp os mswindows misc	talk politics mideast

3.5 Summary

In the preliminary experiments, we have investigated the possible relationship between the vocabulary overlapping of a given text corpus with the F -Measure obtained with the MajorClust clustering algorithm [75]. We have carried out a set of preliminary experiments by calculating the overlapping vocabulary degree [56, 55]. We have observed that it is possible to introduce a measure to determine the relative hardness of clustering corpora based on the vocabulary overlapping. The obtained results show that there exists a correlation between the F -Measure and the RH formula. With respect to the analysis carried out in [19], the introduced formula in our research work relies only on the vocabulary overlapping and it does not use any classifier. In fact, we use the MajorClust clustering algorithm only to evaluate the quality of the proposed formula by employing the F -Measure. Therefore, the introduced RH formula may be used efficiently in order to determine the relative hardness of clustering corpora.

3.6 Further work

As future work, we need to fully investigate the correlation between the relative hardness and the F -Measure also on the *Mini20Newsgroups* dataset. Moreover, we are interested in evaluating both, the vocabulary overlapping and the term frequencies. This will allow us to further investigate whether the use of the $TF-IDF$ formula in the same context improves the current results or not. Besides, we would like to investigate the possible relationship the RH-Measure could have with some cluster validity measures, such as the Density Expected Measure (DEM) which quantifies the similarity within clusters [43]. Some preliminary work has already been started in our research group [28]. The final aim of this research work is to determine the level of hardness of a short text narrow-domain corpus from a clustering task perspective.

To summarize, the tasks we are expecting to investigate are the follows:

1. To evaluate DEM by using MajorClust with R8-Reuters and Mini20Newsgroups.
2. To evaluate the X-Means clustering algorithm instead of MajorClust in the previous step in order to see if the results could depend on the employed algorithm.
3. To investigate the possible correlation of other internal clustering validity measures.
4. To directly evaluate Internal Clustering Validity Measures (ICVM) such as, “The Dunn index family”, DEM and/or Λ -Measure [43], by using the gold standard as the input clusters for those measures.
5. To investigate whether the broadness of a clustering corpora may be measured by internal clustering validity measures.

Chapter 4

Clustering narrow-domain short text corpora

4.1 Description of the problem

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 usually provide free access only to abstracts and not to the full texts of the documents. Moreover, some institutions, like the well known CERN, 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.

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. In [4], 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.

4.2 State of the art

Some related work was presented in [40], 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 CICALing 2002 conference; the second collection was composed by 200 abstracts (215 Kb) from the IFCS-2000¹ 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 [4] 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 [41] in order to perform the word selection process. The authors based their experiments on the first CICALing collection used by Makagonov et al. [40], 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.

The rest of the state of the art we know corresponds to the research work we have done. We will describe each experiment into detail in the following section.

4.3 Our research work

In our research work, we have investigated the possible strategies that should be used in order to tackle the problem of both:

1. The low frequencies of vocabulary terms, and
2. The vocabulary overlapping associated to this particular task.

In the following subsections, we present the experimental work we have conducted.

4.3.1 The role of the term selection process

In our first works presented in [30] and in [29], we introduced a new technique for term selection based on mid-frequency terms, named Transition Point. These works motivated us to use also this technique in the evaluation of a bigger size corpus [51] and to compare the results with other two term selection techniques used in

¹International Federation of Classification Societies; <http://www.Classification-Society.org>

literature [38], named Document Frequency and Term Strength. Our main concern was to evaluate the term selection methods described above, in the task of clustering of abstracts, specifically in a narrow-domain. We have used the *K*-Star clustering method [72] which is considered to be unsupervised and, therefore, it complies with our requirements. The similarities among the documents was calculated by means of the Jaccard similarity function.

Test over a subset of *hep-ex*

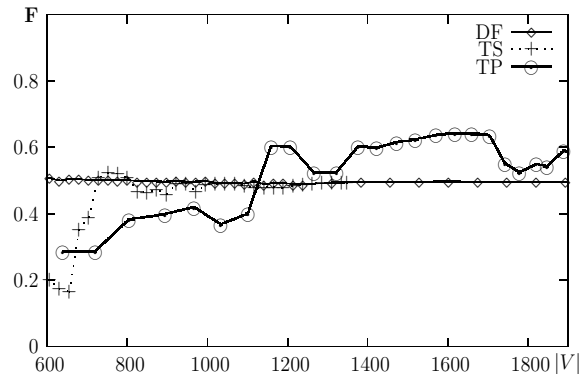
In order to have a first idea of the behaviour of each term selection method used in our experiments, we performed a first test over a subset of *hep-ex*, composed by 500 abstracts randomly selected from the original collection; in the case of those categories with only one instance, we randomly chose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

Figure 4.1 shows the F values for every term selection method executed over different percentages of the collection vocabulary (from 600 to 2,000 terms).

Given a percentage of the collection vocabulary, the DF and TS methods selected the higher score terms. The TP method selected terms in a local fashion; i.e., it took a given number of terms from each text. Therefore, a comparison among methods needed to be done through the vocabularies obtained in each selection of terms carried out by the methods. The 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 of the collection. Given a similar range of total terms, the TP selection method took from 5 to 30 terms from each text. In Fig. 4.1, the results of these three methods are shown; the horizontal axis represents the number of terms and the vertical axis the F values (Eq. 2.11). In order to apply the TS method, a similarity matrix was calculated as 3-tuples (T_i, T_j, sim_{ij}) and sorted according to sim_{ij} , then $TS(t)$ was computed for all terms. Since only 1,349 terms were obtained, the threshold β was fixed to 0.

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

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

Figure 4.1: Behaviour of DF, TS and TP methods in a subset of *hep-ex*.

Analysis of the instability of TP

Although the TP method obtained the highest F value, it did not allowed to determine the correct (smallest) amount of terms to be used in the clustering task. It would be desirable to determine the best selection through an indicator based on the characteristics of the collection. First of all, the clustering method we have used has shown better performance when the number of clusters diminishes. This fact may be used in combination with $df_{V_i}^-$ (which we explain in the following paragraph), to find a possible formula which indicates the optimal number of terms to be selected by the TP method. This is explained in the following paragraph.

Let C_i be the text collection made up 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 (see 2.4.1). Let V_i be the vocabulary of C_i and $df_{V_i}^-$ the average of $DF(t)$ for terms t that belong to V_i but do not belong to V_{i-1} . The $df_{V_i}^-$ value is close-related with the similarity among the texts. Clearly, the lowest value of $df_{V_i}^-$ 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 $df_{V_i}^-$ value ($df_{V_i}^- < df_{V_{i-1}}^-$) contributed to change instances from an incorrect cluster to a correct one. Terms with low $df_{V_i}^-$ seem to help to distribute texts into the clusters, therefore we can use $df_{V_i}^-$ as an indicator of the goodness of a selection C_i .

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

$$df N_i = \frac{(N_i - N_{i-1}) \times (df_{V_i}^- - df_{V_{i-1}}^-)}{N_i}. \quad (4.1)$$

Table 4.1: 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

In Table 4.1 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 the unstability of the TP method, dfN_i can be used in order to determine what clustering set C_i must be used in the clustering task.

Test over the whole *hep-ex* collection

An experiment was performed using the entire collection and applying the three term selection techniques described in Section 2.4. In this case, the noisy words had a notably effect, mainly in the TP method. Since TP 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 and, therefore, the threshold used as parameter in the clustering method. 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 [20], since the terminology of texts is very specialized (see [30]). The problem was solved by using co-occurrence terms (see Section 2.5) as an approximation to related words.

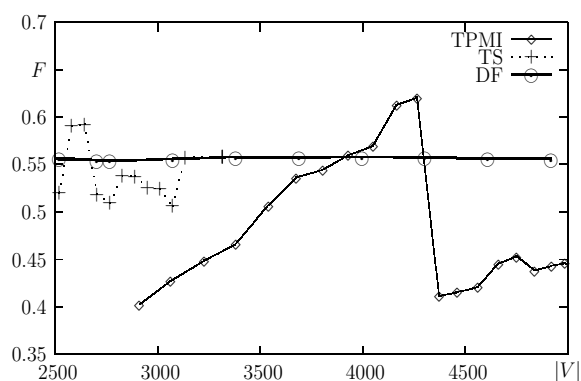


Figure 4.2: Behaviour of DF, TS and TPMI term selection methods

Improving the Transition Point approach

A refined method based on the transition point technique was proposed in order to improve the results obtained over the whole *hep-ex* collection. This novel method was named *Transition Point and pointwise Mutual Information* (TPMI), and basically uses $IDTP(t, D)$ (see Section 2.4.1) and mutual information (see Section 2.5). Therefore, TPMI is a refinement of the selection method provided by TP.

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

$$tpmi(t_i, T) = IDTP(t_i, T) * MI(TP_V, t_i) \quad (4.2)$$

The results obtained by using this refined method are shown in Figure 4.2. We may see that this approach obtains the best value of F -Measure. Very similar clustering results were obtained, for DF and TS methods, on the whole *hep-ex* collection. Anyway, the 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 are greater than the threshold β . The DF method showed to be very stable since it maintains its F values below of the baseline (0.5919). The TPMI method had a good high peak ($F = 0.6206$) selecting 20 terms from each document, obtaining a vocabulary size of 4,268 terms

Summary

We have proposed a new use for the transition point technique in the task of clustering of abstracts in a narrow-domain. We used as a corpus a set of documents (*hep-ex*) of the *High Energy Physics* domain, which led to experiment with real collections composed of very short texts. The findings of the execution of three unsupervised methods (DF, TS and TP) was that TP outperformed the other two methods over a subset of *hep-ex*. However, when the whole collection was used, a new term selection technique had to be developed in order to improve the previous results. This technique 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 very useful due to the very specialized vocabulary of this particular domain. After the calculation of a baseline in both corpora, the experiments we carried out allow us to verify that this technique (TPMI) outperformed the other 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 instability of TP, we carried out an analysis for explaining this behaviour and, therefore, to be able to determine the number of terms needed in the task. It is very important to continue with the study of the stability control for this term selection techniques. In fact, we consider that this should be the key for clustering very short texts.

As further work, we would like to investigate the use in the clustering task of other techniques of dimensionality reduction different than TF, DF, TS and TP such as fingerprinting [73].

4.3.2 A comparative study of clustering methods

In order to investigate a possible independence between the feature selection techniques and the clustering methods, we carried out a comparative study of clustering algorithms which are described as follows.

Description of the experiments

Clustering short texts of a narrow domain, implies basically two steps. For the first step, *the feature selection process*, we have used the three unsupervised techniques described in Section 2.4 in order to sort the vocabulary of each corpus 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. With respect to the second step, *the use of clustering methods*, five different clustering methods were applied for comparison: Single Link Clustering (SLC), Complete Link Clustering (CLC), K-Nearest Neighbour (KNN), K-Star and a modified version of the K-Star method (NN1). For a better description of the mentioned clustering methods, see Section 2.1.

In order to obtain the best description of our experiments, we have carried out a v -fold cross validation [14]. This process implies to randomly split the original corpus in a predefined set of partitions, and then calculate the average F -Measure among all the partitions results. The v -fold cross-validation allows to evaluate how well each cluster “performs” when it 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, respectively, four and thirty partitions for the *CICLing-2002* and *hep-ex* collections.

Experimental results

In Tables 4.2 and 4.3 we show the maximum F -Measure values obtained for each feature selection technique by using the five different clustering methods, for the two different corpora used in the experiments. 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 independent from the clustering method which is employed. 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 the clustering methods we could confirm the above hypothesis.

Table 4.2: Maximum F -Measure obtained over the *CICLing-2002* 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 4.3: Maximum F -Measure obtained over 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) over the *hep-ex* corpus by using the five clustering methods is shown in Figures 4.3, 4.4, and 4.5, respectively. It can be appreciated that the complete link clustering method obtains the best results for all the FSTs. The KNN method obtains instead very poor results. In Figure 4.6 is shown the standard deviation for different sizes of the vocabulary for the *hep-ex* corpus. By obtaining the average of the three FSTs, we can observe 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. This behaviour seems to verify our hypothesis, however, more experiments should be done in the future.

Summary

We have carried out a comparative study of the behaviour of five clustering methods which were applied to two corpora with very different characteristics. Each corpus belongs to a very narrow domain, doing the task even more difficult. We have observed that the transition point technique obtains the best results in comparison with the DF and TS techniques. The obtained results with the three TSTs 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 be desirable to repeat the experiments over other corpora of different domains

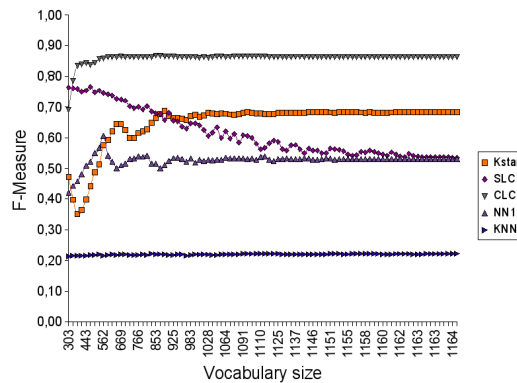


Figure 4.3: F -Measure of the TP technique as a function of the vocabulary size for the five clustering methods we considered (over the *hep-ex* corpus).

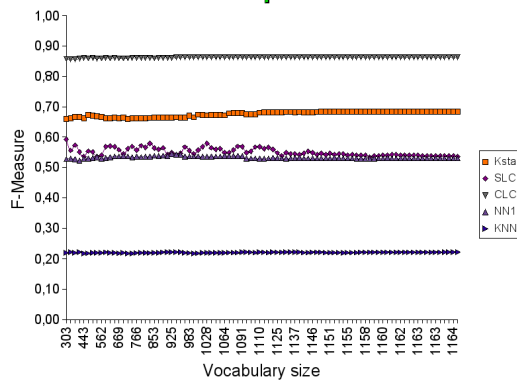


Figure 4.4: F -Measure of the DF technique as a function of the vocabulary size for the five clustering methods we considered (over the *hep-ex* corpus).

to confirm our hypothesis. Unfortunately, at the moment the lack of gold standards for clustering abstracts on narrow domains, makes this task even more difficult. We consider that more attention from the linguistic community is required on the clustering of narrow domain task, not only for the investigation of different feature selection techniques, but also for constructing new narrow domain corpora, with gold standards provided by experts in such domains.

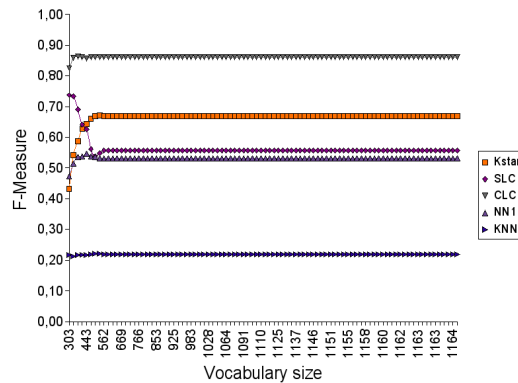


Figure 4.5: F -Measure of the TS technique as a function of the vocabulary size for the five clustering methods we considered (over the *hep-ex* corpus).

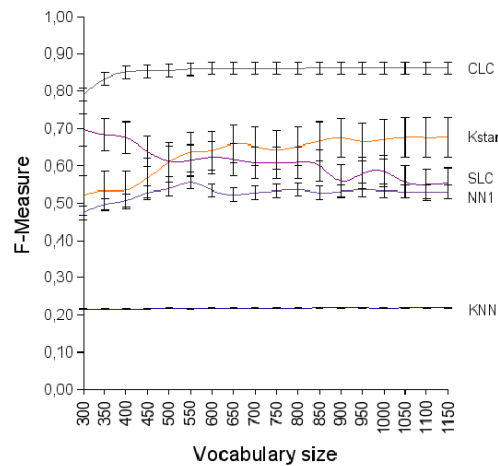


Figure 4.6: Average behaviour of all FSTs with each clustering method using the *hep-ex* corpus

4.3.3 A new clustering similarity measure

Clustering short texts is a difficult task itself, but 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 obtained using the Jaccard similarity measure. The complete

theoretical basis for the clustering similarity measure are described in Section 2.2.3.

Experimental results

We have used the three unsupervised techniques described in Section 2.4 in order to sort the corpora vocabulary in a 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, Complete Link Clustering, and K -Star (see, Section 2.1).

We have carried out a v -fold cross validation evaluation for the experiments. We have used five partitions for the *CICLing-2002* corpus and, thirty for both, the *hep-ex* and the *KnCr* collections. The quality of clusters obtained was determined by means of the F -Measure. The obtained results are presented and discussed below.

In the experiments we have carried out, the DF and TS techniques showed not to improve the results obtained by the transition point technique. These results reinforce the hypothesis made in [51]. 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 4.4, 4.5 and, 4.6 show our evaluation results for all Kullback-Leibler approaches implemented, by using the *CICLing-2002*, *hep-ex* and, *KnCr* corpora, 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 [33], Bigi [10], Jensen [24], and Bennet [7] [86], respectively.

Table 4.4: Results obtained over 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 4.5: Results obtained over 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

We have made a comparison among these results and those reported in [53]. This evaluation is presented in Tables 4.7 and 4.8, where our best approach, which we have

Table 4.6: Results obtained over 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

named *PintoetAl*, is compared with the results presented in [53]. 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 suggest to use a smoothing procedure; unfortunately, the number document terms that does not appear in the corpus vocabulary can be extremely high. Further analysis will investigate this issue.

Table 4.7: Comparison over 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 4.8: Comparison over 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

Summary

We have addressed the problem of clustering short texts of a narrow domain with the use of a new distance measure between documents, which is based on the symmetric Kullback-Leibler distance. We observed that there are few differences in the use of any of the symmetric KL distances analysed. We have evaluated our approach with three different narrow-domain short text 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. The fact that the KLD distance measure is computationally more expensive than the Jaccard one, let us to consider that the simplest implementation should be used in future investigations, i.e., the Jaccard similarity measure.

We have implemented the KLD for using it for the narrow-domain short text clustering task, however we consider that this approach could be successfully 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 is more similar to the corpus vocabulary. Therefore, we consider that a performance improvement could be obtained by using a term expansion method before calculating the similarity matrix with the analysed KLD. Further analysis will investigate this issue.

Chapter 5

The Self-Expansion and Term Selection methodology

5.1 Introduction

The self-term expansion method consists in replacing terms of a document with a set of co-related terms.

Formally, given a corpus of n documents: $C = \{D_1, D_2, \dots, D_n\}$ with vocabulary \mathcal{V} , a document $D_k \in C$: $D_k = \{W_1, W_2, \dots, W_{|D_k|}\}$, and a set of terms (or words) Co-Related (\mathcal{CR}) to each vocabulary word of C obtained by using the same target dataset: $\mathcal{CR} = \{W_i \overset{\circ}{=} W_j | W_i, W_j \in \mathcal{V}\}^1$, the complete self-term expanded version of C (C') is obtained by replacing each term of the corpus by its co-related terms, that is: $C' = \{D'_1, D'_2, \dots, D'_n\}$ with $D'_k = \{W'_1, W'_2, \dots, W'_{|D_k|}\}$, where $W'_j = \{\bigcup W_i | W_i \overset{\circ}{=} W_j\}$.

Although the self-term expansion process may be carried out by mean of different ways, often just by using a knowledge database, we particularly consider important to use first the intrinsic information of the target dataset before using external resources. Thus, we have used the pointwise MI for obtaining a co-occurrence list from the same target dataset (see Section 2.5). This list is then used to expand every term of the original corpus. Since the co-occurrence formula captures relations between related terms, it is possible to see that the self-term expansion magnifies the noisy in a lower degree than it does for the meaningful information. Therefore, we believed that the execution of the clustering algorithm in the expanded corpus should outperform the one executed over the non-expanded data.

¹We will use the symbol $\overset{\circ}{=}$ to mathematically represent the Co-Relation operator.

5.2 state of the art

The expansion of short sentences is not new. In information retrieval, for instance, the expansion of query terms is a very investigated topic which has shown to improve results with respect to when query expansion is not employed [63, 68, 5, 25, 65].

The availability of Machine Readable Resources (MRR) like *Dictionaries*, *Thesauri* and *Lexicons* has allowed to apply term expansion to other fields of natural language processing like WSD. In [6] we may see the typical example of using an external knowledge database for determining the correct sense of a word given in some context. In this approach, every word close to the one we would like to determine its correct sense, is expanded with its different senses by using the WordNet ontology. Then, an overlapping factor is calculated in order to determine the correct sense of the ambiguous word. Different other approaches have made use of a similar procedure. By using dictionaries, the proposals presented in [36, 79, 26] are the most successful in WSD. Yarowsky [83] used instead thesauri for his experiments. Finally, in [76, 64, 6] the use of lexicons in WSD has been investigated. Although in some cases the knowledge resource seems not to be used strictly for term expansion, the application of co-occurrence terms is included in their algorithms.

Like in information retrieval, the application of term expansion in WSD by using co-related terms has shown to improve the baseline results if we carefully select the external resource to use (i.e., with a priori knowledge of the domain). Even more, we have to be sure that the Lexical Data Base (LDB) has been suitably constructed. Due to the last facts, we consider that the use of a self automatically constructed LDB (using the same test corpora), may be of high benefit. This assumption is based on the intrinsic properties extracted from the corpus itself. Our proposal is somehow related with the approaches presented in [71] and [62], where words are also expanded with co-occurrence terms for word sense discrimination. The main difference consists in the use of *the same corpus* for constructing the co-occurrence list and not of an external resource.

5.3 Experiments

The aim of our experiments was to investigate the use of a self-term expansion method in conjunction with a term selection technique for clustering short texts of the very narrow domain corpus. We first observed the behaviour of the application of each TST to the complete collection (named as *baseline* results) before the clustering process is performed. Thereafter, we conducted a set of tests for verifying how the self-term expansion method may improve these baseline results. In our particular case, we have focused on using the unsupervised *K*-Star clustering method, to keep the number of variables as small as possible and make easy the analysis of the main concern of this investigation: the boosting of the performance of clustering narrow-

domain short texts employing the self-term expansion method.

The three unsupervised techniques described in Section 2.4 were used to sort the corpus vocabulary in non-increasing order, with respect to the score of each TST ($IDTP(t, D)$, $DF(t)$ and $TS(t)$). Thereafter, we have selected different percentages of the vocabulary for determining each technique behaviour, under different subsets of the *baseline* corpus. In the experiments we carried out, the v -fold cross validation evaluation was used with ten partitions for the *hep-ex* corpus. For the evaluation of the results, we just created the gold standard of the *hep-ex* collection taking into account the categories that each document has in CERN, and then we have applied the F -Measure for determining the quality of the obtained clusters.

First, we have evaluated the performance of each term selection technique with respect to the preprocessed original corpus. Figure 5.1 shows that the v -fold cross-validated execution over the baseline does not obtain successful results on improving the F -Measure with almost all the term selection techniques (TP, DF and, TS). The best values of each TST are of little significance; in fact, in some of the v executions, none of the TST obtains better F -Measure values than the *baseline*, which in this case is exactly the result of clustering the complete collection (without selection of terms). This experiment will be our general baseline and it will be used for comparison with the following experiments we have carried out. The results show the behaviour of the K-Star clustering method applied to both, the complete corpus (baseline) and, a subset of the *hep-ex* corpus; the latter were extracted by means of three different term selection techniques: transition point, document frequency and, term strength. For each TST, we have reduced from 10% to 90% the corpus vocabulary by selecting the m most relevant terms according to each TST.

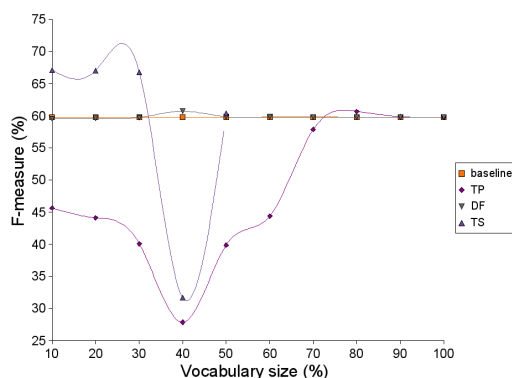


Figure 5.1: Behaviour of TP, DF and TS with respect to the baseline on the *hep-ex* corpus (no expansion)

In order to determine the correct method for calculating the list of co-occurrence used in the self-term expansion process (thesaurus), we have tested two different co-occurrence methods, both with different thresholds: n -grams and pointwise mutual

information. After a set of experiments, we have observed that it is possible to obtain a considerable improvement by using bigrams of frequency 4 and, by using the pointwise mutual information with some restrictions. We have established that each term considered in the MI formula must have frequency equal or greater than 3; besides, we have investigated that the best behaviour of this formula is by considering only bigrams in the numerator of the MI formula. In Figure 5.2 we can see how the baseline results are highly improved by just using the self-term expansion method. We consider that this behaviour is derived from the hypothesis that by adding correlated terms to the original dataset we are increasing both, noise and information to the corpus, however, the valuable information added to the expanded corpus is considerably bigger than the noise introduced and, therefore, it is possible to improve the original results.

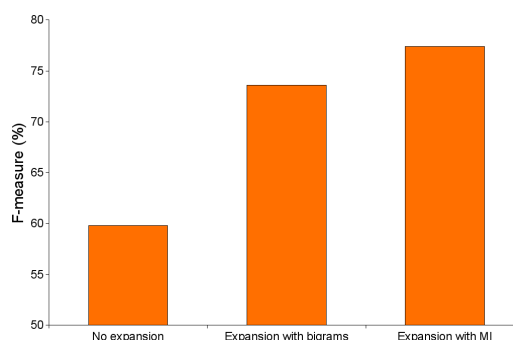


Figure 5.2: Comparison of two different thesauri constructed by using MI and bigrams

From the last observation, we have considered to study the behaviour of each term selection technique with respect to the use of the two self-term expansion methods discussed above. The automatically constructed thesauri allowed to perform term expansion over all the subsets of the *hep-ex* corpus, included the baseline. Firstly, we investigated the effect of expanding after selecting terms. In Figure 5.3 we can observe this experiment by using the two co-occurrence techniques based, respectively, on: (a) bigrams and, (b) pointwise mutual information. With the exception of the TS technique, for every other TST the F -Measure is improved, which means that the self-term expansion works quite well for clustering narrow-domain short texts which is the purpose of this research. However, the TS behaviour lead us to experiment in order to understand whether is better to do first the term selection or the self-term expansion process.

Therefore, we carried out another experiment by expanding the baseline corpus with two different thesauri constructed by using the same two co-occurrence techniques. Thereafter, we have selected different subsets of the expanded corpora by using the three term selection techniques. As we expected, the best results were obtained by using this second approach, that is, to expand first and to reduce the vocabulary

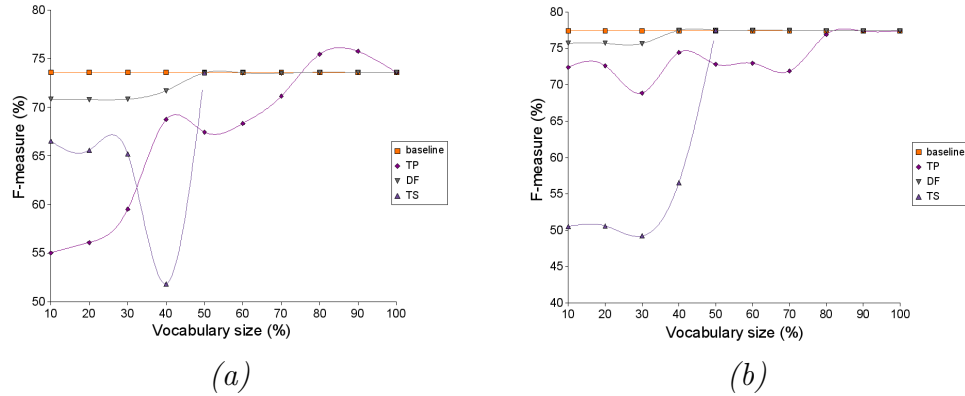


Figure 5.3: Self term expansion (by using: (a) bigrams, (b) MI) *after* the selection of terms (with TP, DF and, TS)

dimensionality later. In Figure 5.4 we show the F -Measure obtained for each subset which was calculated by expanding the original corpus with the thesauri constructed by using (a) bigrams and (b) pointwise mutual information. It is remarkable that, when using the self-term expansion method before applying the term selection technique the best results are obtained with a very small size of the vocabulary. The discrimination of noisy terms is well-executed by each TST. For the *hep-ex* corpus, in particular, we have seen that the DF technique is the one that performs better in comparison with the other two TSTs. Besides that the DF technique obtains the best F -Measure results, it also reduces the corpus vocabulary of approximately 90%.

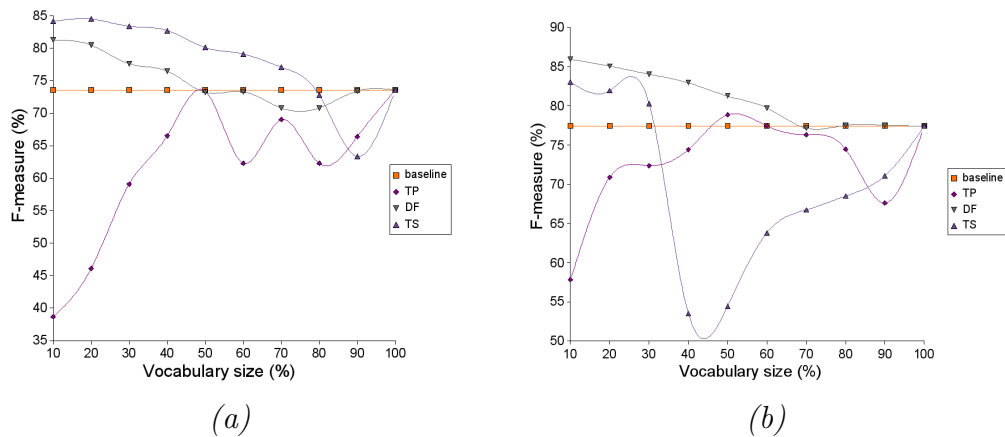


Figure 5.4: Self-term expansion (by using: (a) bigrams, (b) MI) *before* the selection of terms (with TP, DF and, TS)

In order to further observe the effect of applying a self-term expansion method to the clustering short texts of narrow-domain corpora task, a discussion of the best results for each TST is important. In Figure 5.5(a) we may see a comparison of the

best self-term expansion approach for each term selection technique with respect to its correspondig baseline (non-expanded version). Figures 5.5(b), (c) and, (d) show the behaviour of every TST. Following, we will discuss these results in order to clearly see the effect of applying the self-term expansion method proposed before the addition of the TP, DF and, TS techniques.

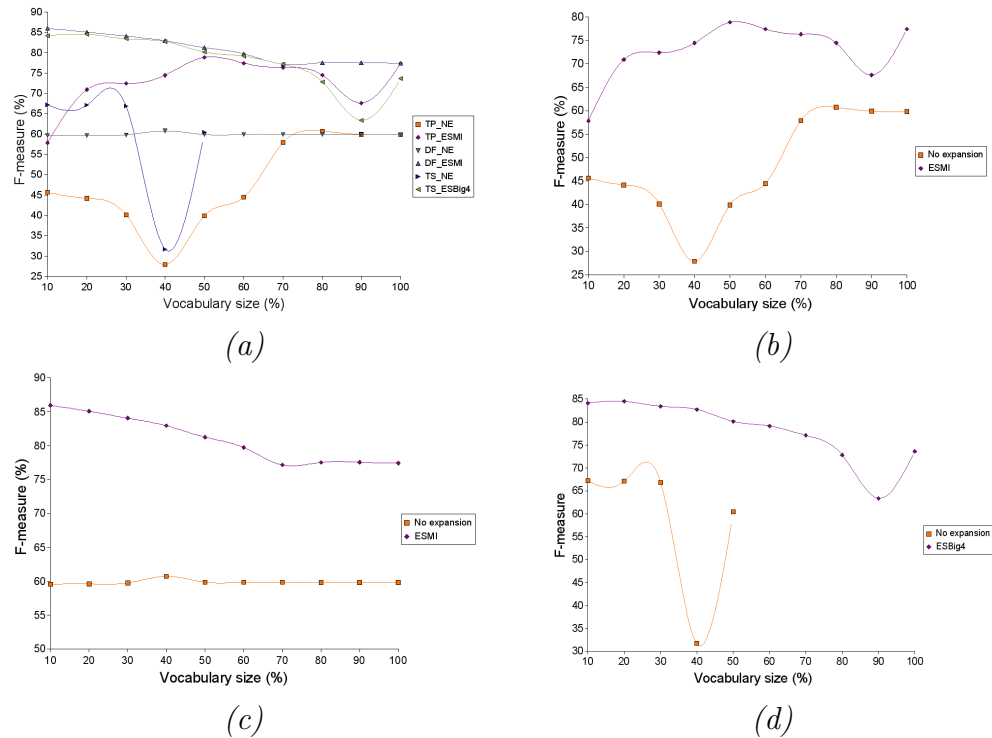


Figure 5.5: Best behaviour of the TSTs with the self-term expansion method: (a) All the TSTs, (b) TP, (c) DF and, (d) TS

Figure 5.5(b) shows the TP behaviour before (no expansion) and after applying the self-term expansion method based on the use of pointwise mutual information (Expansion Selection Mutual Information - ESMI). All the different vocabulary thresholds selected with the TP technique have obtained a high improvement in its correspondat F -Measure. Evenmore, we have observed that the classical erratical behaviour of TP, discussed in [51], become to be less with the use of the self-term expansion process. The experiments carried out allows us to conclude that the TS term selection technique obtains the best performance by using bigrams (Expansion Selection Bigrams - ESBig4) in its self-term expansion process (see Figure 5.5(d)). Another interesting aspect is to observe that a better performance with respect to the non-expanded version is obtained when the expansion process is applied before the term selection. However, the time needed for calculating all the terms is huge and, therefore, the use of the TS term selection technique still not viable for practical

problems.

Finally, we would like to discuss the DF term selection technique. In Figure 5.6 we may see a comparison of the baseline (No expansion approach) with all the other results after evaluating the DF technique with: selection and then expansion by using bigrams (Selection Expansion Bigrams - SEBig4), expansion and then selection by using bigrams (Expansion Selection Bigrams - ESBig4), selection and then expansion by using pointwise mutual information (Selection Expansion Mutual Information - SEMI) and, finally expansion and then selection by using pointwise mutual information (Expansion Selection Mutual Information - ESMI). Here we may see that by expanding first we obtain better results than by first selecting terms, which is quite obvious because the self-term expansion method allows to enrich the baseline corpus and then the TST obtains the best valuable terms for applying the clustering process. The behaviour shown by the ESMI approach with the DF term selection technique is the one which should be expected by a good term selection technique, because we would find more noisy terms as far as we would introduce more terms in the vocabulary.

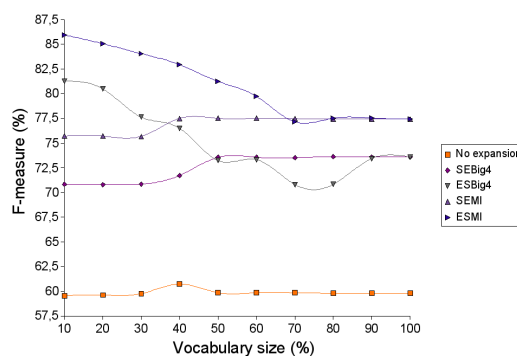


Figure 5.6: Self term expansion and term selection by using DF

5.4 Summary and further work

Clustering short texts of narrow-domain is a very challenging task, because of the high overlapping which exists among all the documents and the low frequencies that the corpora terms have. Therefore, the correct selection of terms for this kind of documents is a quite difficult task. We have introduced a self-term expansion method which allows to enrich the baseline corpus by adding co-related terms from an automatically thesaurus constructed from the *same* target clustering dataset (and not from an external resource). This was done by using two different co-occurrence techniques based, respectively, on: bigrams and pointwise mutual information. Our empirical analysis has shown that it is possible to highly improve clustering results

by doing first the self-term expansion and, thereafter, the term selection process. Moreover, by just doing the self-term expansion, it is possible to improve the target dataset.

The experiments were carried out on a real collection extracted from the CERN, which contains abstracts of papers related to the high energy particles narrow-domain. The main goal of this investigation was the boosting of the performance of clustering narrow-domain short texts employing the self-term expansion method which successfully allowed to improve the baseline F -Measure of approximately 40%. Furthermore, by using term selection after expanding the corpus we have obtained a similar performance with a 90% of reduction of the full vocabulary.

Up to now, we have observed that the above behaviour is true for the narrow-domain short texts corpus used in the experiments, however the application of this methodology to other corpora needs to be investigated. We conclude remarking that the use of term selection techniques is more useful when a prior step of enrichment is carried out. Our experiments with the proposed self-term expansion technique demonstrate how valuable this new technique could be [58]. These experimental results have also allowed us to introduce a methodology for clustering narrow-domain short text corpora. The first step consists in enriching the target clustering corpora by expanding each vocabulary term with its co-occurrence terms calculated over the same target corpora (self-term expansion). The next step will select some term selection method in order to downsize the expanded corpora vocabulary. Once the corpus is expanded and the vocabulary is reduced, classical similarity measures may be used in order to calculate the similarity matrix for some selected clustering algorithm.

The pending tasks to be further investigated are the following ones:

1. To experimentally evaluate the suggested methodology by using the self-term expansion, different kinds of clustering algorithms and term selection techniques as well as different similarity measures when possible.
2. To determine which other areas of natural language processing may benefit from the proposed methodology.

Chapter 6

Applications in other areas of NLP

6.1 Introduction

Theoretical research in Natural Language Processing (NLP) is continuously carried out. However, it is expected to apply the obtained achievements in real situations. Thus, in this chapter we have included the possible applications we have considered up to now for the methods and techniques developed in this research work.

6.2 Word sense induction

Word Sense Induction (WSI) is a particular task of computational linguistics which consists in automatically induce the correct sense for each instance of a given ambiguous word [3]. This problem is closely-related to Word Sense Disambiguation (WSD) [27]. However, whereas in WSD the aim is to tag each ambiguous word in a text with one of the senses known a priori, in WSI the aim is to induce the different senses of that word. Typically, the major systems for WSD tackle this task by using two different approaches: corpus-based and knowledge-based. The accuracy of the corpus-based algorithms for WSD is usually proportional to the amount of hand-tagged data available, but the construction of that kind of training data is often difficult for real applications. The knowledge-based approach uses the ambiguous word context and the information extracted from ontologies (such as WordNet) in order to disambiguate the different senses of a word. For instance, in [15] a knowledge-based approach which uses the conceptual density technique is presented.

6.2.1 Applying the self-term expansion process

For the experiments we have carried out, we used the self-term expansion method described in Chapter 5. We replaced each term of the target corpus with a set of co-related terms calculated by using the pointwise mutual information (see Section

2.5). In order to fully appreciate the self-term expansion method, in Table 6.1 we show the co-occurrence list for some words related with the verb “kill” calculated from the English language corpus used in the “Evaluating Word Sense Induction and Discrimination Systems” task¹ of the SemEval 2007 workshop [3, 59]. Since the MI is calculated after preprocessing the corpus, we present the stemmed version of the terms.

Table 6.1: An example of co-occurrence terms

Word	Co-occurrence terms
soldier	kill
rape	women think shoot peopl old man kill death beat
grenad	today live guerrilla fight explod kill
death	shoot run rape person peopl outsid murder life lebanon kill convict...
temblor	tuesday peopl least kill earthquak

6.2.2 Experiments

For the task #2 of SemEval 2007, a set of 100 ambiguous words (35 nouns and 65 verbs) were provided. We preprocessed this original dataset by eliminating stopwords and then applying the Porter stemmer. Thereafter, when we used the pointwise MI, we determined that the single occurrence of each term should be at least three (see [42]), whereas the maximum separation among the two terms was five. Finally, we selected the unsupervised *K*-Star clustering method for our experiments, defining the average of similarities among all the sentences for a given ambiguous word as the stop criterion for this clustering method. The input similarity matrix for the clustering method was calculated by using the Jaccard coefficient.

The task organizers decided to use two different measures for evaluating the runs submitted to the task. The first measure is called unsupervised one, and it is based on the Fscore measure (*F*-Measure), whereas the second measure is called supervised recall. For further information on how these measures are calculated refer to [1, 2], or see Section 2.3.2 of this document. Since these measures give conflicting information, two different evaluation results are reported in this chapter.

In Table 6.2 we may see our ranking and the Fscore measure obtained (UPV-SI) as well as the best and worst team Fscores, the total average and two baselines proposed by the task organizers. The first baseline (Baseline1) assumes that each ambiguous word has only one sense, whereas the second baseline (Baseline2) is a random assignation of senses. We are ranked as third place and our results are better

¹<http://nlp.cs.swarthmore.edu/semeval/tasks/task02/summary.shtml>

scored than the other teams except for the best team score. However, given the similar values with the “Baseline1”, we may assume that the best team presented one cluster per ambiguous word as the Baseline1 did. Our UPV-SI system obtained instead 9.03 senses per ambiguous word on average.

Name	Rank	All	Nouns	Verbs
Baseline1	1	78.9	80.7	76.8
Best Team	2	78.7	80.8	76.3
UPV-SI	3	66.3	69.9	62.2
Average	-	63.6	66.5	60.3
Worst Team	7	56.1	65.8	45.1
Baseline2	8	37.8	38.0	37.6

Table 6.2: Unsupervised evaluation (F -Measure performance).

In Table 6.3 we show our ranking and the supervised recall obtained (UPV-SI). Once more, we show also the best and worst team recalls. The total average and one baseline are also presented (the other baseline obtained the same Fscore). In this case, the baseline tags each test instance with the most frequent sense obtained in a train split. We are ranked again in the third place and our score is slightly above the baseline.

Name	Rank	All	Nouns	Verbs
Best Team	1	81.6	86.8	76.2
UPV-SI	3	79.1	82.5	75.3
Average	-	79.1	82.8	75.0
Baseline	4	78.7	80.9	76.2
Worst Team	6a	78.5	81.8	74.9
Worst Team	6b	78.5	81.4	75.2

Table 6.3: Supervised evaluation.

6.2.3 Summary

The self-term expansion technique, designed explicitly for narrow-domain short text corpora, has been applied to the word sense induction task which consists in discriminating from a given set of sentences (related with an ambiguous word), those that share the same sense. The results show that the technique employed was able to learn, since our simple approach obtained a better performance than the baselines, especially the one that have chosen the most frequent sense as baseline.

6.2.4 Further work

The third place obtained at the SemEval competition [59] highlights how valuable this simple technique can be in the clustering process. However, the complete evaluation of our methodology has not completely carried out yet. We should perform the evaluation of the different approaches which may be derived from the application of different term selection techniques and term expansion methods. The same evaluation scripts used at SemEval should be used in order to obtain values which can be easily compared with those results obtained by the other proposed approaches. The following are the suggested experiments we plan to carry out:

1. To evaluate the approach which enriches the target dataset by expanding only ambiguous words: Just Ambiguous Word Expanded (JAWE) approach.
2. To evaluate the approach which does not enrich the target dataset: No Expansion (NE) approach.
3. To evaluate the approach which does not enrich the target dataset, using different thresholds of vocabulary reduction: No Expansion with Term Selection (NETS) approach.
4. To evaluate the JAWE approach with different thresholds of vocabulary reduction: Just Ambiguous Word Expanded with Term Selection (JAWETS) approach.
5. To evaluate the approach which enriches the target dataset by expanding all the vocabulary words: All terms Expanded with Term Selection (AETS) approach.
6. To evaluate word by word the results for the best obtained approach.
7. To execute the best approach in other dataset with a language other different than English. We suggest to use the Arabic language, since there exists a dataset already prepared with these characteristics by the organizers of the “Arabic Semantic Labeling”² task of the SemEval competition.

²<http://nlp.cs.swarthmore.edu/semeval/tasks/task18/description.shtml>

Chapter 7

Conclusions and further work

In this last chapter, we draw some conclusions of the investigations we have carried out. Finally, some future works we are interested in investigate are presented.

7.1 Conclusions

Clustering of narrow-domain short text corpora is one of the most difficult tasks of unsupervised data analysis. Dealing with the two features: the high overlapping of vocabularies among the texts, and the specific terminology used in narrow-domain corpora, leads to investigate novel techniques to tackle both problems.

We have addressed the above problems by investigating in three directions:

1. The determination of the hardness of clustering corpora.
2. The study of methods and techniques for improving clustering of narrow-domain short text corpora.
3. The applications of the proposed methods and techniques in different areas of natural language processing.

The minor and major contributions of this research work are enumerated below:
Minor contributions:

1. A similarity measure based on the symmetric Kullback-Leibler distance.
2. One corpus compiled from MEDLINE in the medicine domain, specifically in the *Cancer* domain.
3. The Transition Point (TP) term selection technique.
4. A stabilisation model for the TP term selection technique.
5. A new unsupervised technique for threshold selection in vocabulary reduction

Major contributions:

1. A relative hardness measure which uses the term overlapping among the categories of a supervised corpus
2. An enrichment technique for increasing the term frequencies named self-term expansion
3. A methodology for dealing with narrow-domain short text corpora which uses first self-term expansion and, thereafter, term selection

The new method based on self-term expansion, highly improves results of clustering narrow-domain short texts. Self term expansion means to obtain a thesaurus from the same dataset and then use it for expanding its own terms. Our study also investigates the performance of using the proposed self term expansion when different term selection techniques are employed. We have found that the best combination is to expand first the corpus and then apply a term selection technique. Particularly, when we experimented with a corpus of high energy particles domain (physics), we observed that by using only the term expansion method it is possible to improve the baseline of approximately 40%. Furthermore, by using term selection after expanding the corpus we can obtain a similar performance with a 90% reduction of the full vocabulary.

We have carried out several experiments observing that the clustering of narrow-domain short text corpora is a very challenging task. However, the contributions of this research work are evidence that it is possible to deal with this difficult problem improving the results obtained with typical techniques and methods.

7.2 Further work

There are further experiments we would like to perform in the next future. The description of the future work is given in the following subsections.

7.2.1 Summarization

We are interested in observing the possible relationship the clustering of narrow-domain short text corpora may have with summarization and viceversa. Our plan is to integrate the proposed self-term expansion technique in the summarization task and determine whether the added methodology may improve the classical summarization approach or not. When we talk about the classical approach, we refer to a summarization system that does not use the self-term expansion nor any term selection techniques. Up to now, the summarization task has fully moved its focus to question-answering, since people are interested in obtain a summary from the global content from a data collection. Therefore, we also would like to experiment with

a simple technique which integrates at least the following areas of natural language processing: information retrieval, clustering and summarization. An enumeration of the planned tasks follows:

1. To integrate our self-term expansion methodology in a competitive summarization system.
2. To evaluate the use of information retrieval, clustering and summarization in order to perform the “focused summarization” task.

7.2.2 Text clustering by using information retrieval and summarization

We have carried out different experiments in the Information Retrieval (IR) area. For instance, in [52] we introduced a dimensional reduction technique (also named “vocabulary reduction”) for the evaluation of a Cross-Lingual Information Retrieval (CLIR). In our CLIR system, we took into account different percentages of mid-frequency terms for constructing the representation of the document collection. Those terms were selected by using the transition point technique. Moreover, in the WebCLEF 2006 competition [67], we presented a variant of the earlier approach, enriching the document representation with high co-occurrence terms (we used bigrams). Although the obtained results were not well ranked we observed that the enrichment technique would have worked well by using other co-occurrence term techniques such as the pointwise mutual information. In the same edition of WebCLEF, we introduced a new ranking formula for information retrieval which adds a penalisation factor to the Jaccard coefficient [57]. The obtained rank in the WebCLEF competition, encourages to continue using the proposed formula. Finally, in [66] we introduced a new weighting schema for the vector space model that improves the traditional weighting schema proposed by Salton around the 70s. The proposed technique uses a re-ranking formula for term frequencies based on both, transition point and entropy. Our research work in IR and CLIR areas made us to understand that the use of integrated techniques can be of high benefit to different areas of NLP.

We have observed that a possible clustering method (quite similar to K-Means) may be developed by merging information retrieval, clustering and summarization. Given a set of documents, the proposed clustering method will consist in the following steps:

1. To obtain a single summary for each document of the target collection.
2. To feed an information retrieval system with each summary (query) obtained.
3. To select the best ranked elements of the obtained ranking lists.

4. To assign a similarity weight to each pair of documents by using the ranking lists as a similarity criterion.
5. To use the similarity matrix obtained in order to find out the clusters.
6. To obtain a multi-document summary for each found cluster in order to create a new query.
7. To calculate an internal clustering validity measure, such as Λ -Measure or DEM, on the obtained clusters in order to verify whether the implicit structure of the dataset has been completely detected. If the difference between the previous and the current internal clustering validity measure is less than a given threshold ϵ then finish, otherwise go to the Step 2.

The tasks associated to this topic are:

1. To program the proposed clustering algorithm.
2. To modify Step 2 of the proposed algorithm, using term selection techniques instead of summaries.
3. To compare both approaches on narrow-domain short text clustering corpora.

7.2.3 Fuzzy clustering: The FuzzyMajorClust algorithm

Among various document clustering algorithms that have been proposed so far, the most interesting are those that automatically reveal the number of clusters and assign each target document to exactly one cluster. However, in many real situations, there not exists an exact boundary between different clusters. In this topic we are planning to introduce a fuzzy version of the MajorClust algorithm [37]. The clustering method will assign documents to more than one category by taking into account a membership function for both, edges and nodes of the input similarity matrix for the mentioned clustering algorithm. Thus, the clustering problem will be formulated in terms of weighted fuzzy graphs. The fuzzy approach will permit to decrease some negative effects which appear for clustering of large-sized corpora with noisy data.

After implementing the FuzzyMajorClust algorithm we would test the performance of it against other fuzzy clustering algorithm in the specific problem we have investigated up to now: the clustering of narrow-domain short text corpora.

The tasks planned with respect to this topic are the following ones:

1. To introduce membership functions that will allow to distinguish to which cluster each node belongs to.
2. To test the approach with the Reuters collection and to perform a comparison with the fuzzy version of K-Means, named C-Means.

Bibliography

- [1] E. Agirre, O. Lopez de Lacalle Lekuona, D. Martinez, and A. Soroa. Evaluating and optimizing the parameters of an unsupervised graph-based WSD algorithm. In *Proc. of the Textgraphs 2006 workshop - NAACL06*, pages 89–96, 2006.
- [2] E. Agirre, O. Lopez de Lacalle Lekuona, D. Martinez, and A. Soroa. Two graph-based algorithms for state-of-the-art WSD. In *Proc. of the EMNLP Conference*, pages 585–593. Association for Computational Linguistics, 2006.
- [3] E. Agirre and A. Soroa. SemEval-2007 Task 2: Evaluating Word Sense Induction and Discrimination Systems. In *Proc. of the 4th International Workshop on Semantic Evaluations - SemEval 2007*, pages 7–12. Association for Computational Linguistics, 2007.
- [4] M. Alexandrov, A. Gelbukh, and P. Rosso. An Approach to Clustering Abstracts. In *Proceedings of the 10th International NLDB-05 Conference*, volume 3513 of *Lecture Notes in Computer Science*, pages 8–13. Springer-Verlag, 2005.
- [5] R. Baeza-Yates and B. Ribeiro-Neto. *Modern information retrieval*. New York: ACM Press; Addison-Wesley, 1999.
- [6] S. Banerjee and T. Pedersen. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In *Proc. of the CICLing 2002 Conference*, volume 3878 of *Lecture Notes in Computer Science*, pages 136–145. Springer-Verlag, 2002.
- [7] C. H. Bennett, P. Gács, M. Li, P. Vitányi, and W. Zurek. Information distance. *IEEE Transactions on Information Theory*, 44(4):1407–1423, 1998.
- [8] J. C. Bezdek, W. Q. Li, Y. Attikiouzel, and M. Windham. Geometric approach to cluster validity for normal mixtures. *Soft Computing*, 1(4):166–179, 1997.
- [9] J. C. Bezdek and N. R. Pal. Cluster validation with generalized dunn’s indices. In *Proc. of the 2nd International two-stream conference on ANNES*, pages 190–193, 1995.

-
- [10] B. Bigi. Using kullback-leibler distance for text categorization. In *Proc. of the ECIR 2003 Conference*, volume 2633 of *Lecture Notes in Computer Science*, pages 305–319. Springer-Verlag, 2003.
- [11] B. Bigi, R. d. Mori, M. El-Bèze, and T. Spriet. A fuzzy decision strategy for topic identification and dynamic selection of language models. *Special Issue on Fuzzy Logic in Signal Processing, Signal Processing Journal*, 80(6):1085–1097, 2000.
- [12] B. Bigi, Y. Huang, and R. d. Mori. Vocabulary and language model adaptation using information retrieval. In *Proc. of the International Conference on Spoken Language Processing - INTERSPEECH04*, pages 1361–1364, 2004.
- [13] A. D. Booth. A Law of Occurrences for Words of Low Frequency. *Information and control*, 10(4):386–393, 1967.
- [14] 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.
- [15] D. Buscaldi, P. Rosso, and F. Masulli. The upv-unige ciao-senso wsd system. In *Proc. of the Senseval-3 Workshop*, pages 77–82. Association for Computational Linguistics, 2004.
- [16] C. Carpineto, R. d. Mori, G. Romano, and B. Bigi. An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems*, 19(1):1–27, 2001.
- [17] R. d. Mori. *Spoken Dialogues with Computers*. Academic Press, 1998.
- [18] I. Dagan, L. Lee, and F. Pereira. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1–3):43–69, 1999.
- [19] F. Debole and F. Sebastiani. An analysis of the relative hardness of reuters-21578 subsets. *Journal of the American Society for Information Science and Technology*, 56(6):584–596, 2005.
- [20] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
- [21] E. Fix and J. L. Hodges. Discriminatory analysis: nonparametric discrimination: small sample performance. Technical Report 11, USAF School of Aviation Medicine, Randolph Field, Texas, 1952. Project No. 21-49-004.
- [22] E. B. Fowlkes, R. Gnanadesikan, and J. R. Kettenring. Variable selection in clustering. *Journal of Classification*, 5:205–228, 1988.

-
- [23] W. B. Frakes and R. A. Baeza-Yates. *Information Retrieval: Data Structures & Algorithms*. Prentice-Hall, 1992.
- [24] B. Fuglede and F. Topse. Jensen-shannon divergence and hilbert space embedding. In *Proc. of the International Symposium on Information Theory*, pages 31–40, 2004.
- [25] G. Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic, 1994.
- [26] N. Ide and J. Véronis. Mapping dictionaries: A spreading activation approach. In *Proc. of the 6th Annual Conference of the Centre for the New Oxford English Dictionary*, pages 52–64, 1990.
- [27] N. Ide and J. Véronis. Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, 24(1):2–40, 1998.
- [28] D. A. Ingaramo, Marcelo L. Errecalde, and Paolo Rosso. Medidas subjetivas y objetivas en el agrupamiento de resúmenes científicos de dominios reducidos. *Procesamiento del Lenguaje Natural*, 39, 2007. In press.
- [29] H. Jiménez, D. Pinto, and P. Rosso. Selección de términos no supervisada para agrupamiento de resúmenes. In *Proc. of the Human Language Workshop - ENC05*, pages 86–91, 2005.
- [30] 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.
- [31] S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32:241–254, 1967.
- [32] N. O. Kang, A. Gelbukh, and S. Y. Han. Ppchecker: Plagiarism pattern checker in document copy detection. In *Proc. of Text, Speech and Dialogue 2006 Conference - TSD06*, volume 4188 of *Lecture Notes in Artificial Intelligence*, pages 661–667. Springer-Verlag, 2006.
- [33] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- [34] G. N. Lance and W. T. Williams. A note on a new divisive classificatory program for mixed data. *The Computer Journal*, 14(2):154–155, 1971.
- [35] M. Lazo-Cortes, J. Ruiz-Shulcloper, and E. Alba-Cabrera. An overview of the evolution of the concept of testor. *Pattern Recognition*, 34(4):753–762, 2001.

-
- [36] M. Lesk. Automatic sense disambiguation: How to tell a pine cone from an ice cream cone. In *Proc. of the ACM SIGDOC Conference*, pages 24–26. ACM Press, 1986.
- [37] E. Levner, D. Pinto, P. Rosso, D. Alcaide, and R.R.K. Sharma. Fuzzifying clustering algorithms: The case study of MajorClust. In *Proc. of Advances in Artificial Intelligence - MICAI 2007*, Lecture Notes in Artificial Intelligence. Springer-Verlag, 2007. In press.
- [38] T. Liu, S. Liu, Z. Chen, and W. Ma. An evaluation on feature selection for text clustering. In *Proc. of the 20th International Conference on Machine Learning - ICML 2003*, pages 488–495. AAAI Press, 2003.
- [39] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297. Berkeley, University of California Press, 1967.
- [40] P. Makagonov, M. Alexandrov, and A. Gelbukh. Clustering Abstracts instead of Full Texts. In *Proc. of the Text, Speech and Dialogue 2004 Conference - TSD04*, volume 3206 of *Lecture Notes in Artificial Intelligence*, pages 129–135. Springer-Verlag, 2004.
- [41] P. Makagonov, M. Alexandrov, and K. Sboyshakov. Keyword-based technology for clustering short documents. *Selected Papers. Computing Research*, 2:105–114, 2000.
- [42] D. C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 2003. Revised version May 1999.
- [43] S. Meyer zu Eissen. *On information need and categorizing search*. Phd thesis, University of Paderborn, Germany, Feb 2007.
- [44] S. Meyer zu Eissen and B. Stein. Analysis of clustering algorithms for web-based search. In *Proc. of the 4th International Conference on Practical Aspects of Knowledge Management*, volume 2569 of *Lecture Notes in Artificial Intelligence*, pages 168–178. Springer-Verlag, 2002.
- [45] G. W. Milligan. A validation study of a variable weighting algorithm for cluster analysis. *Journal of Classification*, 6:53–71, 1989.
- [46] 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.

-
- [47] E. Moyotl and H. Jiménez. Experiments in text categorization using term selection by distance to transition point. *Advances in Computing Science*, 10:139–146, 2004.
- [48] J. Neville, M. Adler, and D. Jensen. Clustering relational data using attribute and link information. In *Proc. of the Text Mining and Link Analysis Workshop - IJCAI03*, 2003.
- [49] V. Pekar, M. Krkoska, and S. Staab. Feature weighting for co-occurrence-based classification of words. In *Proc. of the 20th Conference on Computational Linguistics - COLING04*, page 799, 2004.
- [50] D. Pinto, J. M. Benedí, and P. Rosso. Clustering narrow-domain short texts by using the kullback-leibler distance. In *Proc. of the CICLing 2007 Conference*, volume 4394 of *Lecture Notes in Computer Science*, pages 611–622. Springer-Verlag, 2007.
- [51] D. Pinto, H. Jiménez-Salazar, and P. Rosso. Clustering abstracts of scientific texts using the transition point technique. In *Proc. of the CICLing 2006 Conference*, volume 3878 of *Lecture Notes in Computer Science*, pages 536–546. Springer-Verlag, 2006.
- [52] D. Pinto, H. Jiménez-Salazar, P. Rosso, and E. Sanchis. BUAP-UPV TPIRS: A System for Document Indexing Reduction at WebCLEF. In *CLEF 2005, Accessing Multilingual Information Repositories*, volume 4022 of *Lecture Notes in Computer Science*, pages 873–879. Springer-Verlag, 2006.
- [53] D. Pinto, A. Juan, P. Rosso, and H. Jiménez. A comparative study of clustering algorithms on narrow-domain abstracts. *Procesamiento del Lenguaje Natural*, 37(1):43–49, 2006.
- [54] D. Pinto and P. Rosso. KnCr: A short-text narrow-domain sub-corpus of medline. In *Proc. of TLH 2006 Conference*, *Advances in Computer Science*, pages 266–269, 2006.
- [55] D. Pinto and P. Rosso. Easy and hard clustering corpora. In *Proc. of NooJ Conference*, Barcelona, Spain, 2007.
- [56] D. Pinto and P. Rosso. On the relative hardness of clustering corpora. In *Proc. of the Text, Speech and Dialogue 2007 Conference - TSD07*, volume 4629 of *Lecture Notes in Artificial Intelligence*, pages 155–161. Springer-Verlag, 2007.
- [57] D. Pinto, P. Rosso, and E. Jiménez. A Penalisation-Based Ranking Approach for the Mixed Monolingual Task of WebCLEF 2006. In *Cross Language Evaluation Forum - CLEF 2006*, volume 4730 of *Lecture Notes in Computer Science*, pages 826–829. Springer-Verlag, 2007.

-
- [58] D. Pinto, P. Rosso, and H. Jiménez-Salazar. Boosting the clustering of narrow-domain short texts with self term expansion, 2007. In reviewing process.
- [59] D. Pinto, P. Rosso, and H. Jiménez-Salazar. UPV-SI: Word sense induction using self term expansion. In *Proc. of the 4th International Workshop on Semantic Evaluations - SemEval 2007*, pages 430–433. Association for Computational Linguistics, 2007.
- [60] A. Pons-Porrata, R. Berlanga-Llavori, and J. Ruiz-Shulchloper. Topic discovery based on text mining techniques. *Information Processing and Management*, 43(3):752–768, 2007.
- [61] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3), 1980.
- [62] A. Purandare and T. Pedersen. Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proc. of the Conference on Computational Natural Language Learning*, pages 41–48, 2004.
- [63] Y. Qiu and H. P. Frei. Concept based Query Expansion. In *Proc. of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 160–169. ACM Press, 1993.
- [64] P. Resnik. Disambiguating Noun Groupings with Respect to WordNet Senses. In *Proc. of the 3rd Workshop on Very Large Corpora*, pages 54–68. Association for Computational Linguistics, 1995.
- [65] C. J. Van Rijsbergen. *Information Retrieval, 2nd edition*. Dept. of Computer Science, University of Glasgow, 1979.
- [66] F. Rojas, H. Jiménez-Salazar, and D. Pinto. A Competitive Term Selection Method for Information Retrieval. In *Proc. of the CICLing 2007 Conference*, volume 4394 of *Lecture Notes in Computer Science*, Springer-Verlag, pages 468–475, 2007.
- [67] F. Rojas, H. Jiménez-Salazar, and D. Pinto. Vocabulary Reduction and Text Enrichment at WebCLEF. In *Cross Language Evaluation Forum - CLEF 2006*, volume 4730 of *Lecture Notes in Computer Science*, pages 838–843. Springer-Verlag, 2007.
- [68] G. Ruge. Experiments on linguistically-based term associations. *Information Processing & Management*, 28(3):317–332, 1992.
- [69] G. Salton, A. Wong, and C. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

-
- [70] Y. Santiesteban and A. Pons-Porrata. Lex: a new algorithm for the calculus of typical testors. *Mathematics Sciences Journal*, 21(1):85–95, 2003.
- [71] H. Schütze. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123, 1998.
- [72] K. Shin and S. Y. Han. Fast clustering algorithm for information organization. In *Proc. of the CICLing 2003 Conference*, volume 2588 of *Lecture Notes in Computer Science*, pages 619–622. Springer-Verlag, 2003.
- [73] B. Stein. Fuzzy-fingerprints for text-based information retrieval. In *Proc. of the 5th International Conference on Knowledge Management - I-KNOW 05*, pages 572–579, 2005.
- [74] B. Stein and S. Meyer zu Eissen. Automatic document categorization. In *Proc. of Advances in Artificial Intelligence - KI 2003*, pages 254–266, 2003.
- [75] B. Stein and O. Nigemman. On the nature of structure and its identification. In *Proc. of the 25th International Workshop on Graph-Theoretic Concepts in Computer Science*, volume 1665 of *Lecture Notes in Computer Science*, pages 122–134. Springer-Verlag, 1999.
- [76] M. Sussna. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proc. of the 2nd International Conference on Information and Knowledge Management*, pages 67–74, 1993.
- [77] A. R. Urbizagástegui. Las posibilidades de la ley de zipf en la indización automática. Technical report, Universidad de California, Riverside, 1999.
- [78] J. W. Wilbur and K. Sirotkin. The automatic identification of stopwords. *Journal of Information Science*, 18:45–55, 1997.
- [79] Y. Wilks, D. Fass, C. Guo, J. McDonald, T. Plate, and B. Slator. Providing machine tractable dictionary tools. *Machine Translation*, 5(2):99–154, 1990.
- [80] I. H. Witten and E. Frank. *Data mining: Practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann, 2000.
- [81] Y. Yang. Noise reduction in a statistical approach to text categorization. In *Proc. of the 18th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR-ACM*, pages 256–263, 1995.
- [82] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proc. of the 14th. International Conference on Machine Learning - ICML 97*, pages 412–420, 1997.

-
- [83] D. Yarowsky. Word-sense disambiguation using statistical models of Rogets categories trained on large corpora. In *Proc. of the 14th Conference on Computational Linguistics*, pages 454–460. Association for Computational Linguistics, 1992.
- [84] O. R. Zaïane. Principles of knowledge discovery in databases - chapter 8: Data clustering, online-textbook, 1999. <http://www.cs.ualberta.ca/~zaiane/courses/cmput690/slides/Chapter8/>.
- [85] G. K. Zipf. *Human behaviour and the principle of least effort*. Addison-Wesley, 1949.
- [86] B. 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.