The Computer Journal Advance Access published September 13, 2010

doi:10.1093/comjnl/bxq069

A Self-enriching Methodology for Clustering Narrow Domain Short Texts

David Pinto[1,*], Paolo Rosso[2] and Héctor Jiménez-Salazar[3]

[1]*Faculty of Computer Science, Benemérita Universidad Autónoma de Puebla, Puebla, Mexico*
[2]*Natural Language Engineering Lab., ELiRF, Universidad Politécnica de Valencia, Valencia, Spain*
[3]*Information Technologies Dept., Universidad Autónoma Metropolitana, Mexico city, Mexico*
[*]*Corresponding author: dpinto@cs.buap.mx*

Clustering *narrow domain short texts* is considered to be a complex task because of the intrinsic features of the corpus to be clustered: (i) the low frequencies of vocabulary terms in short texts, and (ii) the high vocabulary overlapping associated to narrow domains. The aim of this paper is to introduce a self-term expansion methodology for improving the performance of clustering methods when dealing with corpora of this kind. This methodology allows raw textual data to be enriched by adding co-related terms from an automatically constructed lexical knowledge resource obtained from the same target data set (and not from an external resource). We also propose a set of supervised and unsupervised text assessment measures for evaluating different corpus features, such as *shortness*, *stylometry* and *domain broadness*. With the help of these measures, we may determine beforehand whether or not to use the methodology proposed in this paper. Finally, we integrate all these assessment measures in a freely available web-based system named Watermarking Corpora On-line System, which may be used by computer scientists in order to evaluate the different features associated with a given textual corpus.

Keywords: clustering and analysis of textual data; narrow domain short texts; natural language processing; Internet tools

Received 19 March 2010; revised 17 August 2010
Handling editor: Yannis Manolopoulos

1. INTRODUCTION

The huge volume of information available on the Internet is continuously growing. There is great interest in retrieving, clustering (when the categories are unknown a priori) or classifying (when the categories are known) this information in order to fulfil specific user needs.

We are particularly interested in the analysis of clustering and evaluation methods for textual data (corpora). Document clustering consists of the assignment of documents to a priori unknown categories. The task may be considered to be more difficult than supervised text categorization [1, 2] due to the fact that the information about the category name, number of categories and the correct structure of categorized documents are not provided in advance. Clustering of documents has been approached in different areas of text processing, such as text mining, summarization and information retrieval. In [3, 4], for instance, the way in which document clustering improves precision or recall in information retrieval systems has been studied. The grouping of all the documents that are conceptually similar and, the use of the similarity value between the centroid of each group and a target query have also been investigated in the literature [5–7]. However, the difficulty of finding clustering methods that perform well on different data collections is a problem that has existed for many years [8].

Furthermore, major document contents are written in natural language, and often without any specific helpful structure (title, subtitle, paragraphs, etc). Therefore, clustering of Internet documents has to deal with almost pure raw data, attempting to understand the meaning of those documents. One of the more recent successful approaches is the use of statistical natural language processing (S-NLP) techniques [9]. Some applications of clustering in different areas of S-NLP may include re-ranking of snippets in information retrieval, automatic clustering of scientific texts [10] and source code plagiarism detection [11].

The statistical approach, however, relies on the evidence of words (frequencies), which may be a big problem when dealing

with short texts. Despite the issue of low frequencies, there exist sufficient examples that justify the study of document clustering for the analysis of short documents that come from the Internet or any other source.

Let us suppose, for instance, that a user needs to find Internet information that is associated with the concept 'Cancer'. The results obtained by a web search engine, such as Yahoo or Google, may be ambiguous. In Wikipedia,[1] it is possible to find 11 different uses for this word (a group of malignant diseases, a constellation, an astrological sign, the major circle of latitude, etc.). Thus, the number of snippets obtained will be irremediably affected by the frequency of each possible sense of the word 'Cancer' on the Internet. Even if we are interested in the most frequent sense on the web (a group of malignant diseases), it would be desirable to provide an intuitive browsing capability for each one of the subcategories of the searched documents (prostate cancer, breast cancer, etc). Some web search engines have approached this idea with promising results (see Clusty, Mooter and KartOO);[2] however, as we mentioned above, the accuracy may be affected by the term frequency of the query submitted to the search engine and also by the possible ontologies used in the term clustering process.

Now, it is clearly seen that the Internet furnishes abundant proof of the inevitability and the necessity of analysing short texts. News, document titles, abstracts, FAQs, chats, etc., are some examples of the high volume of short texts available on the Internet. Therefore, there exists sufficient interest from the computational community to analyse the behaviour of categorization methods when using short text corpora [12–18]. If these short texts belong to the same domain (e.g. sports or physics), we say that they are narrow domain texts. It is already difficult to cluster short texts, but if those documents are also narrow domain, then the complexity of the task increases significantly. As an example, consider the performance of clustering narrow domain short texts (less than 0.6 of F_1-measure [19]) in comparison with clustering those corpora that are not of that kind (usually more than 0.6 of F_1-measure [20]).

Consequently, the aim of this research work is to investigate the problem of clustering a particular set of documents, namely *narrow domain short texts*. To achieve this goal, we have analysed different data sets and clustering methods. Moreover, we have introduced new corpus evaluation measures and a novel methodology in order to tackle the following two problems associated with corpora of this kind:

(i) the low frequencies of vocabulary terms in short texts, and
(ii) the high vocabulary overlapping associated to narrow domains.

[1] http://en.wikipedia.org/wiki/Cancer_(disambiguation).
[2] http://clusty.com/;http://mooter.com/;http://kartoo.com/.

The rest of this paper is structured as follows. Section 2 summarizes the challenges faced when dealing with short-length documents which belong to narrow domains.

In Section 3, we study different corpus features and we present some measures in order to evaluate them. In particular, we introduce and also revisit different formulae to evaluate the following three corpus features: (i) domain broadness, (ii) shortness and (iii) stylometry. The final goal is to easily evaluate a given corpus in order to determine in an unsupervised manner whether the corpus is narrow domain short text or not.

In Section 4, we present a novel methodology which alleviates the problem of low-term frequency derived from the short length of the documents. We take advantage of the a priori known corpus characteristic of being narrow domain in order to construct a list of co-occurrence terms which are useful to enrich the document representation.

The experimental results are given in Sections 5 and 6. We validate the theoretical hypothesis with different corpora, first by evaluating their features (Section 5) and later by clustering self-term expanded versions of narrow domain short text corpora (Section 6). The obtained results show that the combination of evaluation measures and the new methodology proposed is a valuable contribution when clustering narrow domain short texts. Finally, in Section 7 we discuss the conclusions, highlighting the contributions of this research work.

2. DEALING WITH NARROW DOMAIN SHORT TEXTS

Short text corpora are text collections made up of documents containing a few words. The principal characteristic of short texts is that the frequency of the terms is relatively low in comparison with their frequency in long documents. The ratio between the document vocabulary cardinality and the document size may give a clue about the low frequency of the document in short texts. The following discussion is motivated by the statements given in [21].

Formally, given a document d with cardinality $|d|$, vocabulary size $|V(d)|$ and the corresponding Short Representation of d (SR(d)), we may compute the Shortness Degree of d as $\mathrm{SD}(d) = \log|V(d)|/log|d|$. For instance, if we have both a full document d_{F} containing 1700 words with a vocabulary size of 530, and a short representation of the same document $\mathrm{SR}(d_{\mathrm{F}})$ (say an abstract) with cardinality 70 and vocabulary size equal to 48, the shortness of d_{F} and $\mathrm{SR}(d_{\mathrm{F}})$ will be 0.84 and 0.91, respectively. In other words, it is feasible to automatically determine whether or not a given document is a *short text*.

We may consider that the equality $|V(d)| = |d|^{\mathrm{SD}(d)}$ expresses in some way the shortness degree of d, and, therefore, the vocabulary size is assumed to be a simple power function of $|d|$. The closer $\mathrm{SD}(d)$ is to one, the more complex the document is. A *short text* (let us say 200–500 words) should

have SD(d) $\approx$ 1, whereas *very short texts*, such as a query input in a search engine (let us say 1–10 words) will usually have SD(d) = 1. A detailed description of how to determine whether or not a text is short is presented in Section 3.2.

The problem of 'short text clustering' is quite relevant, given the current and future way people use 'small-language' (e.g. blogs, snippets, news and text-message generation such as email or chat). The difference between representing short texts and long documents is mainly 2-fold: high vocabulary dimensions and sparse data spaces. The average document similarity of short text collections is very low [22]. Therefore, it becomes a great drawback for clustering purposes because clustering methods have a very narrow gap to discriminate whether or not the documents are truly similar. In this case, it is very difficult to obtain an acceptable clustering accuracy [10].

Additionally, a corpus may be considered to be *narrow* or *wide* domain if the vocabulary overlapping level of the documents is high (let us say more than 60%) or low (let us say less than 40%), respectively. In the clustering task, it is very difficult to deal with narrow domain corpora such as scientific papers, technical reports, patents, etc [23]. In [24], the vocabulary overlapping is calculated for the documents of the most different groups of a corpus that is composed of scientific documents from the computational linguistics field, e.g. 'ambiguity' and 'text processing'. The authors obtained ∼70% of vocabulary overlapping between the two categories, which implies that the selected domain is rather *narrow*. Although it will be desirable to assign each document to only one of these two categories ('ambiguity' or 'text processing'), due to the high vocabulary overlapping most clustering methods would be confused, merging all the documents in only one cluster, a fact that highlights the complexity of clustering narrow domain corpora.

Therefore, the automatic detection of whether or not a given corpus is narrow domain could be of high benefit in the clustering task. However, until now there has not been an agreement about a simple formula to determine the degree of *domain broadness* for a given corpus, i.e. whether the corpus is *narrow* or *wide* domain. In Section 3, we introduce different approaches and formulae to calculate the degree of domain broadness of a corpus from a supervised and an unsupervised viewpoint.

In the literature, there exist some works that have studied the clustering of narrow domain short documents [19, 22–28]. We consider it important to discuss some of these approaches in order to give some insights on the previous works in the field of narrow domain short text corpora.

In [27], for instance, the tasks of categorization and clustering of narrow domain documents are investigated. The experiments for the classification task were carried out by using Bernoulli mixtures for binary data, and in the case of the clustering task, by means of the MajorClust clustering method [29]. The proposed method for clustering narrow domain short texts extracts sense clusters from abstracts, exploiting the WordNet [30] relationships existing between words in the same text. This work relies again on a hand-crafted external resource. It was claimed that the approach performed well for a particular narrow domain. However, it is not expected that this kind of approach based on domain-generic resources may be used in every domain with the same performance.

Makagonov *et al*. presented in [22] simple procedures to improve results by an adequate selection of keywords and a better evaluation of document similarity. These authors used as corpora two collections retrieved from the Web. The first collection was composed of a set of 48 abstracts (40 Kb) from the CICLing 2002 conference (described in Section 5.1.1); the second collection was composed of 200 abstracts (215 Kb) from the IFCS-2000[3] 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 the different broadness of the domain and combined measures. The authors propose two modifications to the traditional approach when clustering documents. Firstly, they suggest selecting keywords from the word frequency list taking into consideration objective criteria related to relative frequency of words with respect to general lexis and the expected number of clusters. Secondly, they propose to measure the document similarity by using a weighted combination of the cosine and polynomial measures. The problem from our particular viewpoint is that the filtering process relies on the existence of another balanced corpus of the same language. Moreover, the thresholds used are empirical which do not guarantee the same results in different environments.

In [26], a new method which uses latent semantic indexing is presented. This identifies conceptually related genes based on titles and abstracts in MEDLINE citations. They used small and well-defined gene-document collection containing 50 genes which was manually constructed by selecting three 'broad' categories (development, Alzheimer's disease and cancer biology). The authors claimed that they may obtain high precision despite the small number of documents. This result may be derived from the fact that the categories selected do not share high number of terms in their vocabulary.

In [24], an approach for clustering abstracts in a narrow domain using the MajorClust method for clustering both keywords and documents was presented. Here, Alexandrov and Colleaques used the criterion introduced in [25] in order to perform the word selection process. The idea was to use only those terms of the document collection that have greater frequency than double the frequency of those in a comparable corpus. They cluster the stemmed terms by using the same document vector space representation. Finally, they smooth the frequency of the final term index t_k in the document collection D with the following formula: $\log(1+tf(t_k, D))$, where $tf(t_k, D)$ is the frequency of t_k in D. The purpose of the latter formula

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

was to ameliorate the effects of dealing with the low frequencies of the terms. The authors based their experiments on the first collection (CICLing 2002) used by Makagonov *et al.* [22], 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 reinforced the statement given by Makagonov *et al.* in [22], suggesting that, for an open access via the Internet, digital libraries should provide document images of full texts for the papers and not only abstracts.

All the authors of the research works previouly mentioned agree about the high level of difficulty that is faced when categorizing documents of this kind. The combination of both features, *narrow domain* and *short text* in a corpus will give it a higher level of complexity in order to obtain the desired accuracy of clustering. An explanation of this phenomenon can be given as follows: (i) on the one hand, even if a document set is made up of short texts, if the vocabulary overlapping is low, the clustering task may be carried out easily. The reason is that it is easy to distinguish among the categories of the given corpus. (ii) On the other hand, if the data collection is narrow domain but composed of long documents, the possibility of distinguishing the documents through terms other than those overlapping is still possible; that is the reason why combining both features, broadness and shortness, in a corpus increases the complexity of clustering it.

Collections of scientific documents (and, in consequence their abstracts) are an example of narrow domain short texts. However, clustering scientific abstracts implies a special level of difficulty. The reason is that texts belonging to scientific papers often share sequences of words such as 'in this paper we present', 'the aim is', 'the results', etc., which obviously increase the level of similarity among the short text collections. Therefore, the correct selection of terms when clustering texts is very important because the results may vary significantly.

The purpose of studying scientific abstracts is not only due to their specific high complexity, but also because most digital libraries and other web-based repositories of scientific and technical information provide free access only to abstracts and not to the full texts of the documents. Many scientific repositories such as MEDLINE, the CERN[4], the ACM[5], and others receive hundreds of publications that must be categorized in some specific domain, often with an unknown number of categories a priori.

Let us take, for instance, the PubMed portal[6] which contains an online search engine for the MEDLINE articles. It has indexed more than 16 million abstracts. This huge volume of information, which is practically impossible to manage using only human resources, requires the help of a computational-based system with the purpose of automatically categorizing those documents. Novel methods for clustering narrow domain short texts must be constructed to deal with this real problem.

Some approaches tackled this particular problem with successful results. However, the applications are domain-dependent since they made use of classifiers that were trained with data that were tagged with keywords extracted from domain-dependent thesauri [31]. However, in scientific domains, there rarely are linguistic resources to help in supervised categorization tasks due to the specific vocabulary (narrow) of the documents. Moreover, sometimes the use of scientific document keywords (which are seldom provided by authors) may be insufficient to perform a good clustering [32].

Due to the dynamic aspect of research, new interests could arise in a field and new subtopics need to be discovered through clustering in order to be introduced later as new categories. Therefore, the clustering of abstracts becomes a real necessity.

In the following section, we present the basis for the analysis and evaluation of textual data. In particular, we focus the analysis on narrow domain short texts.

3. ASSESSMENT OF TEXT CORPORA

Evaluation of textual resources is an important topic which needs to be addressed, for instance, in international evaluation forums. It is usually assumed that the corpora provided for experiments are of sufficient quality to be used as benchmarking in the competitions. However, the fact that a committee of experts agrees about the gold standard of a given corpus does not imply 100% usefulness or applicability of the resource for the specific purpose for which it was constructed. It could happen that some particular linguistic or structural feature, such as the cardinality of classes (class imbalance), may bias the expected results in a competition.

Moreover, when dealing with raw text corpora, it is possible to find a set of features that determine the hardness of the clustering task itself, ad hoc clustering methods may be used in order to improve the quality of the obtained clusters. Therefore, we believe that this study would be highly beneficial.

In [33], the authors attempted to determine the relative hardness of different Reuters-21578[7] subsets by executing various classifiers. However, in their research, no measure is defined for determining the hardness of these corpora, neither the possible set of features that could be involved in the process of calculating the relative corpus hardness.

For the purpose of our investigation, we took into account three different corpus features: *domain broadness*, *shortness* and *stylometry*. We consider that these features could be partially used to evaluate the relative hardness of a document collection in order to agree on, for instance, whether or not there is a narrow gap between the gold standard of a corpus and the

[4]Conseil Européen pour la Recherche Nucléaire; http://www.cern.ch/.
[5]Association for Computing Machinery.
[6]http://www.ncbi.nlm.nih.gov.
[7]http://www.daviddlewis.com/resources/testcollections/reuters21578/.

categories obtained through the execution of some clustering method. As far as we know, research in this field nearly has been carried out in the literature. The description of the features investigated is given as follows.

3.1. Broadness detection

The domain broadness of a given corpus is a very important categorizer-independent feature that should be considered when evaluating a data collection to be clustered, in order to determine its possible relative hardness.

When we evaluate the broadness of a given corpus, we assume (see, for instance, [34]) that it is easier to cluster documents belonging to very different categories, for instance, 'sports' and 'seeds', than those belonging to very similar ones, e.g. 'barley' and 'corn'. A binary 'broadness' classifier would assign, respectively, the tags *wide* to the former 'sports-seeds' collection and *narrow* to the latter 'barley-corn' one.

However, it is not clear the manner in which this evaluation should be carried out. In the rest of this section, we introduce different measures to attempt to evaluate the corpus domain broadness degree from a vocabulary-based perspective. We present the supervised and unsupervised version of two approaches: one based on statistical language modelling (SLM) and another based on vocabulary dimensionality.

3.1.1. Using SLM for domain broadness evaluation

SLM is commonly used in different natural language application areas such as machine translation, part-of-speech tagging, information retrieval, etc. (e.g. [35–37]). However, it has been originally known for its use in speech recognition (see, for instance, [38]) which is still its most important application area.

Informally speaking, the goal of SLM consists of building an SLM in order to estimate the distribution of words/strings of natural language. The calculated probability distribution over strings S of length n, also called n-grams, attempts to reflect the relative frequency in which S occurs as a sentence. In this way, from a text-based perspective, such a model tries to capture the writing features of a language in order to predict the next word given a sequence of them.

We propose to use SLM in order to calculate probabilities of sequences of words (n-grams) and, thereafter, to determine the domain broadness degree of a given corpus by using two different supervised and unsupervised variants.

We have assumed that every hand-tagged category of a given corpus to be clustered has a particular/specific language model. Therefore, if this model is very similar to the models of the other categories, then we could affirm that the corpus is narrow domain. The similarity between two models is calculated by means of perplexity which is a measurement in information theory defined as b raised to the power of cross-entropy in base b. In summary, the perplexity of a discrete probability distribution p is defined as $b^{H_b(p,q)}$, where $H(p,q)$ is the cross-entropy of the distribution p with respect to q.

The degree of broadness may be approximated by evaluating this proposed *supervised* approach over the target corpora. We also approached in an unsupervised way the problem of determining the domain broadness of a given corpus. For this purpose, we calculate language models for v random exclusive partitions of the corpus, without any knowledge about the expert document categorization (gold standard).

Due to the fact that the perplexity is by definition dependent on the text itself, we should make sure that the chosen text is representative of the entire corpus [39]. In fact, in [40] it is said that 'The perplexity of a language model depends on its application domain. There is generally higher precision (and less ambiguity) in specialized fields than in general English'.

On the basis of the previous assumptions, we propose a supervised evaluation measure for the relative broadness of corpora to be clustered as follows.

Given a corpus D with a gold standard made up of k categories (also named as classes) $\mathcal{C}^* = \{C_1^*, C_2^*, \ldots, C_k^*\}$, we obtain the language model of all the categories except C_i^* ($\bar{C}_i^*$) and, afterwards, we compute the perplexity of the obtained language model with respect to the model of C_i^*, that is, we use the category C_i^* as a test corpus and the remaining ones as a training corpus in a leave-one-out process. Formally, the *Supervised* Language Modelling Based (SLMB) approach for determining the domain broadness degree of the corpus D may be obtained as shown in the following equation:

$$\mathrm{SLMB}(D) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(2^{H_2(C_i^*,\bar{C}_i^*)} - \mu(2^{H_2(\mathcal{C}^*)})\right)^2}, \qquad (1)$$

where

$$\mu(2^{H_2(\mathcal{C}^*)}) = \frac{\sum_{i=1}^{k} 2^{H_2(C_i^*,\bar{C}_i^*)}}{k}. \qquad (2)$$

Introducing an unsupervised measure for evaluating the domain broadness of corpora would be beneficial, for instance, for clustering algorithms. In fact, we could select ad hoc techniques in order to enrich the documents (Section 4) in a previous step of the clustering process itself or to approximate the exact point to cut-off in hierarchical clustering methods.

The main problem consists of finding the correct way of splitting the corpus in order to allow that the evaluation by using SLMs could make sense. An immediate solution would be splitting the document collection in percentages (e.g. 10) and using, for instance, 10 or 20% for test and the remaining for training purposes. This approach should work well for the evaluation of narrow domain corpora, but it would not be useful for wide domain corpora, since the expected language model for the training and test partitions would hold high similarity. We then propose to use a static number of documents for the test split which should work well with narrow and wide domain evaluation.

The *Unsupervised* Language Modelling Based (ULMB) approach for assessing the domain broadness of a text corpus is formally described as follows.

Given a corpus D split into subsets C_i $(1 \leq i \leq k)$ of l documents, we calculate the perplexity of the language model of C_i with respect to the model of a training corpus composed of all the documents which are not contained in C_i $(\bar{C}_i)$. If $\bar{C}_i \cup C_i = D$, such as $\bar{C}_i \cap C_i = \emptyset$ and k =Integer($|D|/|C_i|$) with $|C_i| \approx l$, then the *unsupervised* broadness degree of a text corpus D may be obtained as shown in the following equation:

$$\mathrm{ULMB}(D) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(2^{H_2(C_i,\bar{C}_i)} - \mu(2^{H_2(\mathcal{D})})\right)^2}, \quad (3)$$

where

$$\mu(2^{H_2(\mathcal{D})}) = \frac{\sum_{i=1}^{k} 2^{H_2(C_i,\bar{C}_i)}}{k}. \quad (4)$$

3.1.2. *Using vocabulary dimensionality for domain broadness evaluation*

This measure of domain broadness assumes that corpora subsets that belong to a narrow domain share the maximum number of vocabulary terms compared with those subsets which do not. In case of a wide domain corpus, it is expected (at least with short texts) that the standard deviation of vocabulary sizes obtained from subsets of this corpus (with respect to the full corpus vocabulary) is greater than the one of a narrow domain corpus. The formal description of the above presented hypothesis is given as follows.

Given a corpus D with a gold standard made up of k categories $\mathcal{C}^* = \{C_1^*, C_2^*, \ldots, C_k^*\}$, if $|V(D)|$ is the cardinality of the complete document set vocabulary and $|V(C_i^*)|$ the vocabulary size of the category C_i^*, the *Supervised* Vocabulary-Based (SVB) measure for calculating the domain broadness of D may be written as shown in the following equation:

$$\mathrm{SVB}(D) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(\frac{|V(C_i^*)| - |(V(D))|}{|D|}\right)^2}. \quad (5)$$

An Unsupervised version of the Vocabulary-Based (UVB) domain broadness evaluation measure may be also proposed. This approach would be useful when the gold standard is not available. Since the categories are unknown, we could then use each document instead of the corpus categories. Formally, given a corpus made up of n documents $D = \{d_1, d_2, \ldots, d_n\}$, if $|V(D)|$ is the cardinality of its vocabulary and $|V(d_i)|$ the vocabulary size of the document d_i, then the *unsupervised* broadness evaluation measure of D (based on the vocabulary dimensionality) may be written as shown in the following equation:

$$\mathrm{UVB}(D) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{|V(d_i)| - |(V(D))|}{|D|}\right)^2}. \quad (6)$$

3.2. Shortness evaluation

The evaluation measures presented in this section calculate features derived from the length of a text, such as the ratio between the document vocabulary size and the document length. The term frequency, for instance, is crucial for the majority of similarity measures used in text categorization. When dealing with very short texts, we expect the frequency of their vocabulary terms to be very low. Therefore, clustering algorithms, for instance, will have problems detecting the correct cluster assignment since the similarity matrix will have very low values. This is derived from the fact that many clustering algorithms assume that the expected average of normalized similarities (between 0 and 1) in a corpus is greater than the average (in this case 0.5), which is not true when dealing with short texts.

Given a corpus made up of n documents $D = \{d_1, d_2, \ldots, d_n\}$, we revisited three *unsupervised* text length-based evaluation measures which take into account the level of shortness [21]. In the first and second approaches, we directly calculate the arithmetic mean of Document Lengths (DL) and Vocabulary Lengths (VL) as shown in Equations (7) and (8), respectively. In Equation (9) we show the third measure, introduced in [41], which obtains the average of Vocabulary vs. Document cardinality Ratios (VDR).

$$\mathrm{DL}(D) = \frac{1}{n}\sum_{i=1}^{n}|d_i|, \quad (7)$$

$$\mathrm{VL}(D) = \frac{1}{n}\sum_{i=1}^{n}|V(d_i)|, \quad (8)$$

$$\mathrm{VDR}(D) = \frac{\log(\mathrm{VL}(D))}{\log(\mathrm{DL}(D))}. \quad (9)$$

3.3. Stylometry analysis

Stylometry studies the linguistic style of a human writer. One of the practical applications of this field consists of determining the authorship of documents. In our case, the aim is not to attribute the authorship but to distinguish between scientific and other kind of texts.

It has been observed that when the collection to be clustered is composed of scientific texts, then a new level of difficulty arises [23]. This observation may have its basis on domain-dependent vocabulary sentences or terms that are not considered in the pre-processing step, such as 'in this paper', 'the obtained results', 'in table', etc.

There have been carried out several works on the statistical study of the writing style (stylometry) field [42], which is still an active research area [43, 44]. For the analysis of stylometry introduced in this section, we make use of the Zipf law. This empirical law was formulated by using mathematical statistics. In the context of text analysis, the Zipf law refers to the fact that the term frequency distribution may be described by a particular

distribution named 'Zipfian'. This is a particularization of a more general fact depicted in [45] which establishes the many types of data that could be described by the Zipfian distribution.

Formally, given a corpus D with vocabulary $V(D)$, we may calculate the probability of each term $t_i \in V(D)$ as shown in Equation (10), where $tf(t_i, D)$ is the frequency of t_i in D. The expected Zipfian distribution of terms is obtained as shown in Equation (11). We used the classic version of the Zipf's law and, therefore, s was set to 1.

$$P(t_i, D) = \frac{tf(t_i, D)}{\sum_{t_i \in V(D)} tf(t_i, D)}, \tag{10}$$

$$Q(t_i, D) = \frac{1/i^s}{\sum_{r=1}^{|V(D)|} 1/r^s}. \tag{11}$$

the *unsupervised* Stylometric Evaluation Measure (SEM) of D is obtained by calculating the asymmetrical Kullback–Leibler distance of the term frequency distribution of D with respect to its Zipfian distribution, as shown in the following equation:

$$\text{SEM}(D) = \sum_{t_i \in V(D)} P(t_i, D) log \frac{P(t_i, D)}{Q(t_i, D)}. \tag{12}$$

4. THE SELF-TERM EXPANSION METHODOLOGY

The previously presented assessment measures allow us to analyse a given corpus in order to determine whether or not it is composed of narrow domain short texts. If the correct corpus features are discovered, then we may take advantage of ad hoc techniques in order to further improve the corpus clustering task.

In this section, we present a novel methodology for dealing with the two particular problems derived from clustering narrow domain short texts. Depending on the analysis done over the corpus (Section 3), and before starting the clustering phase, we propose to improve the representation of its short-length documents by using a term enrichment procedure (Section 4.1), often called *term expansion*. We consider that a proper enrichment of the target documents in the clustering task will improve the 'semantic' similarity hidden behind the lexical structure. Thereafter, we consider the use of term selection techniques (TSTs) (Section 4.2) in order to filter the most discriminant terms and to decrease the number of terms being used in the clustering process. Figure 1 graphically depicts the proposed Self-Term Expansion methodology.

4.1. Self-term expansion

The term expansion process consists of replacing terms of a document with a set of co-related terms. This procedure may be carried out in different ways, often by using an external knowledge resource which usually helps in obtaining successful results [46–48].

However, we consider particularly important the attempt to use firstly the intrinsic information of the target data set

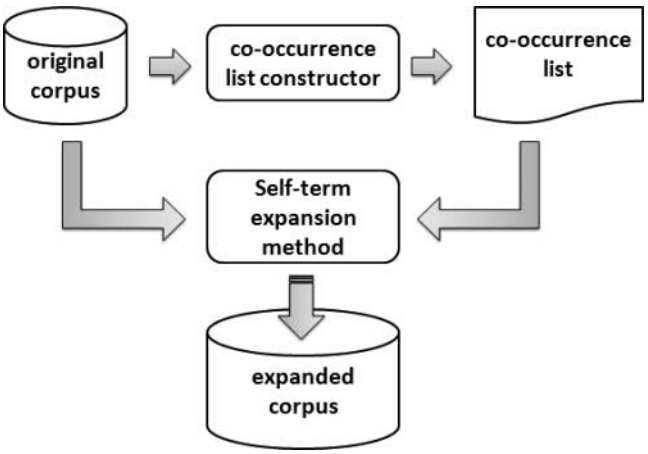

FIGURE 1. The Self-term expansion methodology.

itself before using external knowledge. The motivation relies on the fact that the applications that use external resources such as WordNet [30] are domain dependent since they make use of classifiers trained with data, which were tagged with keywords extracted from domain-dependent thesauri. Moreover, in narrow domains there is a lack of linguistic resources to help in the categorization task due to the specific or narrow vocabulary of the documents.

We propose a domain-independent term expansion technique that works without the help of any external resource. We called this approach as *self-term expansion* due to the fact that the term expansion is done by using only the same corpus to be clustered. The self-term expansion technique uses a co-occurrence term list calculated from the same target data set through the pointwise mutual information (PMI) measure [9, 49]. This list is then used to expand every term of the original corpus. Since the co-occurrence formula captures associations between term pairs, we consider that the self-term expansion magnifies both meaningful information and noise, but hopefully the former with a greater intensity than the latter. Therefore, we consider that employing a clustering algorithm on the self-expanded corpus should allow us to obtain better results. This hypothesis is confirmed from previous experiments [10, 23, 50] and those presented here, where the performance of the self-term expansion approach always outperforms baseline results obtained by clustering raw data(i.e. without expansion).

PMI is an information theory-based co-occurrence measure discussed in [9] for finding collocations. Given two terms t_i and t_j, the PMI formula (Equation (13)) calculates as follows 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 occurred alone:

$$\text{PMI}(t_i, t_j) = \log_2 \frac{P(t_i t_j)}{P(t_i) P(t_j)}. \tag{13}$$

The proposed self-term enrichment technique is formally described as follows.

Let $D = \{d_1, d_2, \ldots, d_n\}$ be a document collection with vocabulary $V(D)$. Let us consider a subset of $V(D) \times V(D)$ of *correlated terms* as $\mathcal{RT} = \{(t_i, t_j) | t_i, t_j \in V(D)\}$. The $\mathcal{RT}$ expansion of D is $D' = \{d'_1, d'_2, \ldots, d'_n\}$ such that, for all $d_i \in D$, the following two properties are satisfied: (i) if $t_j \in d_i$, then $t_j \in d'_i$, and (ii) if $t_j \in d_i$, then $t'_j \in d'_i$, with $(t_j, t'_j) \in \mathcal{RT}$. If $\mathcal{RT}$ is calculated by using the same target data set, then we say that D' is the *self-term expansion* version of D.

The degree of co-occurrence between a pair of terms (we use terms as synonym of 'content words', i.e. after stop listing) may be calculated through any co-occurrence method, since this model is based on the intuition that two terms are semantically similar if they appear in a similar set of contexts. This assumption comes from the Harris hypothesis (words with similar syntactic usage have similar meaning), which was proposed in [51].

Once the co-occurrence list has been obtained, the self-term expansion may be carried out. It simply concatenates each original term with its corresponding set of co-related terms. The next section explains the three TSTs used in the experiments carried out in this paper.

4.2. Term selection techniques

Several TSTs aimed at the clustering of the text task exist [52, 53]. However, we have considered three of them that better represent different term selection approaches. The following three unsupervised TSTs were used:

(i) *Document Frequency (DF)*: This technique assigns the value $\mathrm{df}(t_i)$ to each term t_i, where $\mathrm{df}(t_i)$ means the number of texts, in a collection, where t_i occurred. This technique assumes that low frequency terms will rarely appear in other documents, therefore, they will not have significance on the prediction of the category for a text.

(ii) *Term Strength (TS)*: The weight given to each term t_i is defined by the following equation:

$$\mathrm{ts}(t_i) = P(t_i \in d_A | t \in d_B), \text{ with } A \neq B, \qquad (14)$$

where d_A and d_B are two documents that must be as similar as a given threshold, i.e. $\mathrm{sim}(d_A, d_B) \geq \beta$. β must be tuned according to the values inside the similarity matrix. A more detailed description can be found in [52, 53].

(iii) *Transition Point (TP)*: A higher value of weight is given to each term t_i, as its frequency is closer to a frequency named the transition point (TP_V), which can be found by an automatic inspection of the vocabulary frequencies of each text, identifying the lowest frequency (from the highest frequencies) that it is not repeated; this characteristic comes from the formulation of Booth's law for low frequency words [54] (see [23] for a complete explanation of this procedure). The following equation shows how to calculate the final value:

$$\mathrm{idtp}(t_i, d) = \frac{1}{|\mathrm{TP}_V - \mathrm{freq}(t_i, d)| + 1}, \qquad (15)$$

where $\mathrm{freq}(t_i, d)$ is the frequency of the term t_i in the document d.

5. ASSESSING NARROW DOMAIN SHORT TEXTS

In order to be consistent with the presented approach, before applying the self-term expansion methodology, we first evaluate the features of the two corpora used in the experiments. The aim is to verify that these corpora are 'narrow domain' and composed of short texts.

A brief description of the data sets used in the experiments is given first. The evaluation of the three corpus features is presented later by using the previously introduced evaluation measures.

5.1. Data sets

We have selected two text collection sets for the experiments. The first one was used for testing the assessment measures, whereas the second one was employed for determining the performance of the self-term expansion methodology. The former data set was compiled from the Internet repositories and represents a sample of Internet texts which has been used at clustering tests in the computational linguistics field. The second data set was extracted from a Physics laboratory and it is composed of scientific abstracts on the physics topic.

We have pre-processed all the data collections by eliminating stop words and by applying the Porter stemmer [55]. The corpus features given in each table were obtained after applying this pre-processing phase.

5.1.1. *The CICLing-2002 corpus*

This corpus is made up of 48 documents from the *Computational Linguistics* domain, which corresponds to the *CICLing 2002* conference.[8] The categories and the distribution of them are shown in Table 1, whereas other features of this corpus are given in Table 2.

The collection was first used by Makagonov *et al.* [22] in their experiments on clustering narrow domain abstracts. Even though very small, we consider it as a necessary reference corpus because it has been used in other experiments [24, 25]. Moreover, its size made it possible to manually validate the obtained results.

5.1.2. *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* [56]. It is composed

[8] http://www.cicling.org.

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

Category	Topics	Number of abstracts
Linguistics	Semantics, syntax, morphology and parsing	11
Ambiguity	Word sense disambiguation, part of speech tagging, anaphora and spelling	15
Lexicon	Lexics, corpus and text generation	11
Text processing	Information retrieval, summarization and classification of texts	11

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

Feature	Full documents	Abstracts
Size of the corpus (bytes)	542 370	23 971
Number of categories	4	4
Number of texts	48	48
Total number of terms	80 109	3382
Vocabulary size (terms)	7590	953
Term average per text	1668.94	70.45

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

Category	Number of abstracts
Particle physics (experimental results)	2623
Detectors and experimental techniques	271
Accelerators and storage rings	18
Particle physics (phenomenology)	3
Astrophysics and astronomy	3
Information transfer and management	1
Non-linear systems	1
Other fields of physics	1
Unknown (originally named as XX)	1

of 2922 abstracts from the *Physics* domain originally stored in the data server of the CERN. The *hep-ex* corpus was released to be used in the automatic text classification of documents task presented in [31]. They used multiple categories for their experiments, however, for the purposes of our research work, we used the coarse grain categories of this corpus which implied to work with a single categorized data collection.

The distribution of the categories and other characteristics, such as the vocabulary size and the average size of the documents, are shown in Tables 3 and 4. As can be seen, this corpus is totally unbalanced, and consists of narrow domain short texts.

TABLE 4. General features of the *hep-ex* corpus.

Feature	Value
Size of the corpus (bytes)	962 802
Number of categories	9
Number of abstracts	2922
Total number of terms	135 969
Vocabulary size (terms)	6150
Term average per abstract	46.53

5.2. Analysis of the corpus features

We have evaluated three corpus features (domain broadness, shortness and stylometry) with the aforementioned narrow domain short text corpora. However, in order to compare the obtained results with not narrow domain short text corpora, we have also evaluated other nine corpora which have been usually used in the text clustering and/or classification task (*WSI-SemEval*, *WebKb-Training*, *WebKb-Test*, *R8-Test*, *R52-Test*, *R8-Training*, *R52-Training*, *20Newsgroups-Training*, *20Newsgroups-Test*). A complete description of these additional corpora may be found in [19].

A manual ranking among all the evaluated corpora was done in order to analyse the quality of the proposed corpus assessment measures. Five different human evaluators (two linguists and three computational linguists) were asked to rank the corpora according to each corpus feature. Thereafter, we use the Borda single-winner voting scheme[9] for determining the final ranking which was considered to be the gold standard.

Thus, in order to assess how well each broadness evaluation measure ranking correlates with their corresponding manually evaluated ranking, we have calculated one correlation coefficient among them. For the purpose of this analysis, besides considering that there are no tied ranks, we also have not made any assumptions about the frequency distribution of the evaluation measures. In order to make the correlation less sensitive to non-normality in distributions (rankings), we avoided using the well-known Pearson correlation coefficient [57]. Instead we would used the non-parametric measure of correlation named *Spearman's rank correlation coefficient* [58], however, the equidistance among the different corpora evaluation values could not be justified. Thus, the correlation between each pair of corpora evaluation measure ranks was calculated by means of the *Kendall tau rank correlation coefficient* [59], which is described as follows.

[9] http://en.wikipedia.org/wiki/Borda_count.

5.2.1. *Kendall tau rank correlation coefficient*

The Kendall tau coefficient (τ) is calculated as shown in the following equation.

$$\tau = \frac{2 \cdot C_P}{(k \cdot (k-1))/2} - 1, \quad (16)$$

where k is the number of items, and C_P is the number of concordant pairs obtained as the sum, over all the items, of those items ranked after the given item by both rankings.

The Kendall tau coefficient value lies between -1 and 1, and high values imply a high agreement between the two rankings. Therefore, if the agreement (disagreement) between the two rankings is perfect, then the coefficient will have the value of 1 (-1). In the case of obtaining the value 0, it is said that the rankings are completely independent.

5.2.2. *Measure by measure analysis*

We first analysed the domain broadness measure which is based on SLM. Tables 5 and 6 show the obtained corpora domain broadness evaluation with both the supervised SLMB and the unsupervised ULMB measures, respectively. The broadness ranking associated to the obtained value is shown in the third column, whereas the manual ranking is given in the fourth column. Both supervised and unsupervised measures agree on the fact that scientific documents are narrow domain, whereas news collections belong to a wide domain. In fact, the Kendall tau coefficient value for both the supervised and unsupervised measures are 0.82 and 0.56, respectively. As will be seen in Section 5.3, these values indicate a strong agreement between the automatic and the manual ranking. Now we may conclude that both measures perform well on evaluating the domain broadness degree of a given corpus in both supervised and unsupervised ways.

The vocabulary dimensionality-based evaluations SVB and UVB are shown in Tables 7 and 8, respectively. The evaluation of

TABLE 5. Ranking of domain broadness with SLMB (rank correlation value $\tau = 0.82$).

Corpus	SLMB	Automatic ranking	Manual ranking
CICLing-2002	38.92	1	1
WSI-SemEval	195.02	2	3
WebKb-Training	262.26	3	5
hep-ex	298.15	4	2
WebKb-Test	337.39	5	4
R8-Test	545.69	6	6
R52-Test	565.81	7	8
R8-Training	603.95	8	7
R52-Training	627.60	9	9
20Newsgroups-Training	694.38	10	11
20Newsgroups-Test	786.02	11	10

TABLE 6. Ranking of domain broadness with ULMB (rank correlation value $\tau = 0.56$).

Corpus	ULMB	Automatic ranking	Manual ranking
CICLing-2002	63.62	1	1
hep-ex	93.82	2	2
WSI-SemEval	130.62	3	3
R8-Test	134.60	4	6
R8-Training	135.87	5	7
R52-Training	143.10	6	9
R52-Test	177.54	7	8
WebKb-Test	218.85	8	4
20Newsgroups-Training	400.20	9	11
20Newsgroups-Test	455.38	10	10
WebKb-Training	628.60	11	5

TABLE 7. Ranking of domain broadness with SVB (rank correlation value $\tau = 0.67$).

Corpus	SVB	Automatic ranking	Manual ranking
WebKb-Test	0.44	1	4
WebKb-Training	0.50	2	5
CICLing-2002	1.73	3	1
WSI-SemEval	1.80	4	3
hep-ex	2.75	5	2
R8-Training	3.67	6	7
R8-Test	3.84	7	6
R52-Training	4.38	8	9
R52-Test	4.58	9	8
20Newsgroups-Test	5.21	10	10
20Newsgroups-Training	5.23	11	11

TABLE 8. Ranking of domain broadness with UVB (rank correlation value $\tau = 0.56$).

Corpus	UVB	Automatic ranking	Manual ranking
WebKb-Test	1.60	1	4
WebKb-Training	1.77	2	5
CICLing-2002	2.70	3	1
WSI-SemEval	3.06	4	3
hep-ex	3.07	5	2
R52-Training	4.62	6	9
R8-Training	4.76	7	7
R52-Test	4.82	8	8
R8-Test	4.89	9	6
20Newsgroups-Test	6.05	10	10
20Newsgroups-Training	6.08	11	11

TABLE 9. Ranking the corpus language stylometry with SEM (rank correlation value $\tau = 0.86$).

Corpus	SEM	Automatic ranking	Manual ranking
R8-Test	0.0980	1	1
R52-Test	0.1196	2	2
R8-Training	0.1420	3	4
20Newsgroups-Test	0.1437	4	3
20Newsgroups-Training	0.1543	5	6
R52-Training	0.1593	6	5
WebKb-Test	0.2273	7	7
WebKb-Training	0.2306	8	8
hep-ex	0.2711	9	10
CICLing-2002	0.3013	10	11
WSI-SemEval	0.4477	11	9

both rankings, automatically and manually obtained, were done with the Kendall tau correlation coefficient. The corresponding τ values were 0.67 and 0.56 for both SVB and UVB, respectively. The high degree of correspondence (Section 5.3) indicates a strong agreement between the automatic and the manual ranking. Therefore, we consider that also these proposed measures may be used to calculate the domain broadness degree of the corpora to be clustered.

The stylometry-based corpora evaluation measure determines the language style of writing. Thus, we expect to obtain a high value when the style is very specific, whereas a low value would indicate a general language writing style. In Table 9, we may see the obtained values by the SEM evaluation measure for *hep-ex* and *CICLing-2002* with respect to other different corpora. The obtained Kendall tau correlation coefficient is 0.86, and based on a test of significance we may conclude that this value implies a strong degree of agreement between the automatic and the manual ranking. Thus, the Kullback–Leibler distance between the Zipfian distribution and the term frequency distribution provides a very good indicator of the language writing style of a given corpus.

Tables 10 and 11 show the values obtained by using, respectively, DL and VL corpus evaluation measures. Table 10 shows the arithmetic mean of document sizes and Table 11 presents the mean ratio between the vocabulary and document size for each corpus.

As expected, the computed Kendall tau correlation coefficient value shows a high agreement between the manual and automatical rankings with the DL and VL evaluation measures, obtaining 0.96 and 0.78, respectively.

However, the values for the VDR measure shown in Table 12 are completely different. The correlation coefficient obtained a value of 0.05, which implies total independence between the two rankings. We consider that two issues affected the last result. On the one hand, the VDR measure is biased by the size of the

TABLE 10. Ranking of average document size obtained with DL (rank correlation value $\tau = 0.96$).

Corpus	DL	Automatic ranking	Manual ranking
hep-ex	46.53	1	1
WSI-SemEval	59.58	2	2
R8-Test	60.05	3	3
R52-Test	64.30	4	4
R8-Training	66.32	5	5
R52-Training	70.32	6	6
CICLing-2002	70.46	7	7
WebKb-Training	133.67	8	9
WebKb-Test	136.23	9	8
20Newsgroups-Test	138.73	10	10
20Newsgroups-Training	142.65	11	11

TABLE 11. Ranking of average document vocabulary size obtained with VL (rank correlation value $\tau = 0.78$).

Corpus	VL	Automatic ranking	Manual ranking
hep-ex	36.87	1	1
R8-Test	37.28	2	3
R52-Test	39.71	3	4
R8-Training	41.20	4	5
R52-Training	43.11	5	6
CICLing-2002	48.40	6	7
WSI-SemEval	50.30	7	2
WebKb-Training	77.13	8	9
WebKb-Test	79.42	9	8
20Newsgroups-Test	83.15	10	10
20Newsgroups-Training	84.32	11	11

TABLE 12. Mean ratio of vocabulary and document size computed with VDR (rank correlation value $\tau = 0.05$)

Corpus	VDR	Automatic ranking	Manual ranking
WSI-SemEval	0.9586	1	2
hep-ex	0.9394	2	1
CICLing-2002	0.9117	3	7
20Newsgroups-Test	0.8962	4	10
20Newsgroups-Training	0.8938	5	11
WebKb-Test	0.8902	6	8
WebKb-Training	0.8877	7	9
R8-Training	0.8865	8	5
R52-Training	0.8849	9	6
R52-Test	0.8843	10	4
R8-Test	0.8836	11	3

TABLE 13. *P*-values obtained for each assessment measure according to null hypothesis H_0.

Assessment measure	τ correlation value	*P*-value
SLMB	0.82	0.0002
ULMB	0.56	0.0084
SVB	0.67	0.0021
UVB	0.56	0.0084
SEM	0.86	0.0002
DL	0.96	0.0001
VL	0.78	0.0006
VDR	0.05	0.2

corpus, since the more the number of documents, the higher is the variation of the average document vocabulary. On the other hand, we consider that the manual ranking was based on the assumption that VDR will obtain performance similar to DL and VL. However, it seems that VDR assesses the complexity of the corpus (vocabulary vs. size) and not exactly the shortness of the documents.

5.3. Test of significance

We have carried out a test of significance in order to verify whether or not the τ correlation values are statistically significant. We defined the null hypothesis as follows:

H_0: The ranking values manually obtained by human evaluators and those automatically obtained by the assessment measures are independent of each other.

In Table 13, we may see both the τ correlation value and the *P*-value for each assessment measure. From the *P*-values obtained we may reject the null hypothesis for the assessment measures SLMB, ULMB, SVB, UVB, SEM, DL and VL with a significance level $\alpha = 0.01$ (this value corresponds to the probability of observing such an extreme value by chance). However, in the case of the VDR measure, we cannot reject H_0. This behaviour is derived from the wrong assumption of being a measure for evaluating the shortness of a document, a fact that was considered as the subjective judgment of the human evaluators.

5.4. Summary

We have developed the assessment measures as 'indicators' on the evaluation of corpora. We are quite confident about the performance of these measures for positive cases (narrow domain short texts), but we are not so confident with negative cases (which cannot be categorized as narrow domain short texts). We consider that these measures may be improved by taking into account, for instance, the fact that in the case of the *hep-ex* corpus there is a huge imbalance of classes, a fact that contributes to bias the supervised measures, in contrast with the behaviour of the unsupervised measures, where the data subsets are distributed uniformly.

The comparison of the obtained values with narrow domain short texts with respect to those corpora that do not share the same characteristics, i.e. not narrow domain short texts, permits us to experimentally determine a set of range values which may be used to discriminate between these two categories.

In summary, we have introduced a set of assessment measures which may be useful to easily evaluate different corpus features. The final aim of this section is to provide an automatic mechanism to determine whether a corpus is narrow domain short text or not.

Additionally, we have integrated the presented corpus assessment measures in a freely available on-line system.[10] which may be used, for instance by linguists and computational linguists, in order to fully evaluate three different features associated to a given corpus: *shortness*, *stylometry* and *domain broadness*.

6. CLUSTERING SELF-EXPANDED NARROW DOMAIN SHORT TEXTS

In this section, we present the results obtained by applying the proposed methodology to narrow domain short text corpora. The aim is to highlight the need for improving text representation for narrow domain short documents. In summary, we are interesting in analysing the use of the self-term expansion technique together with TSTs for clustering narrow domain short text corpora.

We first observed the behaviour of the application of each TST on the complete collection (*baseline* results) before the clustering process is performed. Thereafter, we carried out a set of tests for verifying how the self-term expansion technique may improve these baseline results. In our particular case, we have focused on using two unsupervised clustering methods: *K*-Star [60] and DK-Means [19], in order 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 methodology. The use of *K*-Star is justified by the fact that it is a completely unsupervised clustering method which automatically discover the number of clusters. This characteristic was very important for the experiments, because it allowed us to decrease the number of variables to analyse. On the other hand, we used a deterministic version of the well-known *K*-Means method as a comparison parameter. We would like to highlight the fact that this paper does not intend to be a comprehensive study of clustering

[10]The Watermarking Corpora On-line System is a web-based tool available at http://nlp.dsic.upv.es:8080/watermarker/ and http://nlp.cs.buap.mx/watermarker/.

algorithms, but to be an indicator of the performance of clustering 'self-term expanded' corpora.

The three unsupervised TSTs were used to sort the corpus vocabulary in a non-increasing order with respect to the score of each TST (idtp(t_i, D), df(t_i) and, ts(t_i)). 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 5 and 10 partitions, respectively, for the *CICLing-2002* and the *hep-ex* corpus that, as formally investigated in the previous section, were shown to be narrow domain short text (abstracts) corpora.

For the evaluation of the quality of the results, we compared the obtained clusters with respect to the gold standard by using the F_1-measure, which is a particular version of the general formulation of F_β-measure [1]. We used the F_1-measure because we were interested in comparing the obtained results with those obtained by previous works on clustering narrow domain short texts. In general, the F_1-measure is an external clustering measure that compares the clusters obtained by some clustering method with respect to the categories given by an expert. Formally, given a set of clusters $\mathcal{C} = \{C_1, \ldots, C_{|\mathcal{C}|}\}$ and a set of categories $\mathcal{C}^* = \{C_1^*, \ldots, C_{|\mathcal{C}^*|}^*\}$, the F_1-measure between a cluster C_i and a category C_j^* is given by the following formula.

$$F_1(C_i, C_j^*) = \frac{2 \cdot \text{Precision}(C_i, C_j^*) \cdot \text{Recall}(C_i, C_j^*)}{\text{Precision}(C_i, C_j^*) + \text{Recall}(C_i, C_j^*)}, \quad (17)$$

where $1 \leq i \leq |\mathcal{C}|$, $1 \leq j \leq |\mathcal{C}^*|$. The precision and the recall between a cluster C_i and a category C_j^* are defined as follows:

$$\text{Precision}(C_i, C_j^*) = \frac{|C_i \cap C_j^*|}{|C_i|}, \quad (18)$$

and

$$\text{Recall}(C_i, C_j^*) = \frac{|C_i \cap C_j^*|}{|C_j^*|}. \quad (19)$$

The global performance of a clustering method is calculated by using the values of $F_1(C_i, C_j^*)$, the cardinality of the set of clusters obtained, and normalizing it with respect to the total number of documents $|D|$ in the collection. The obtained measure is referred to as the F_1-measure and it is shown in the following equation:

$$F_1\text{-measure} = \sum_{1 \leq i \leq |\mathcal{C}|} \frac{|C_i|}{|D|} \max_{1 \leq j \leq |\mathcal{C}^*|} F_1(C_i, C_j^*). \quad (20)$$

6.1. Co-occurrence methods

In order to determine the correct method for calculating the list of co-occurrence used in the self-term expansion process, we have tested two different co-occurrence methods (n-grams and PMI) with different thresholds. We investigated the behaviour of each TST with respect to the use of the two self-term expansion techniques.

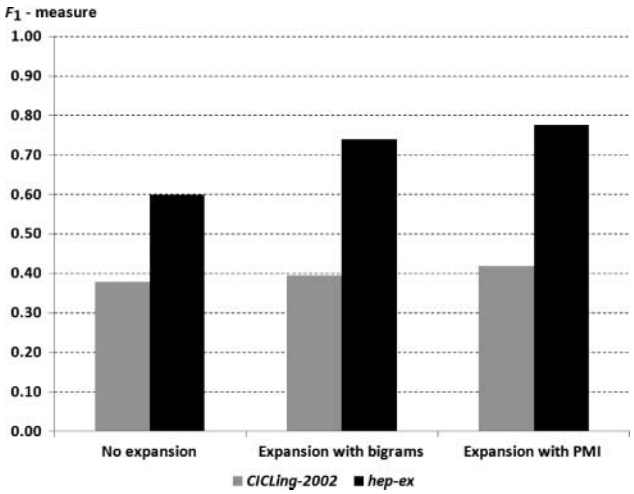

FIGURE 2. Effect of self-term expanding both the *CICLing-2002* and the *hep-ex* narrow domain short text corpus with two co-occurrence methods.

The experimental results showed that it is possible to obtain a considerable improvement when using bigrams of frequency greater than or equal to 4, and PMI with a threshold equal to 7. However, since the bigram counting is considered to be part of the PMI formula, it was expected that PMI would outperform the bigrams results. The obtained F_1-measure values confirmed the previous hypothesis. In fact, in Fig. 2 we may see how the baseline results calculated by clustering the full corpus (without expansion) are highly improved by just using the self-term expansion technique. We consider that this behaviour is derived from the following hypothesis. The addition of co-related terms to the original data set implies an increase of both noise and meaningful information in the expanded corpus. However, the valuable information added to the expanded corpus is considerably higher than the noise introduced which makes it possible to improve the original results.

Having observed that the best of the two co-occurrence methods analysed was the PMI, we carried out a further set of experiments with the two *CICLing-2002* and *hep-ex* corpora. The focus was mainly to investigate the performance of the self-term expansion technique over the two full corpora and their subsets. For this purpose, we expanded the full version of each corpus and, thereafter, we constructed the corresponding subsets of both the expanded and unexpanded version of the corpus with the following techniques of term selection: TP, TS and DF.

In Table 14, we show the maximum (MAX), minimum (MIN) and average (AVG) value obtained when applying the self-term expansion methodology to the *CICLing-2002* corpus. The K-Star clustering method was executed over nine different

TABLE 14. Analysis of the self-term expansion methodology by using the *K*-Star clustering method over the *CICLing-2002* corpus.

	F_1-measure		
Approach name	MAX	MIN	AVG
TP	0.68	0.56	0.64
TP-Exp	0.70	0.61	0.67
DF	0.64	0.58	0.61
DF-Exp	0.66	0.59	0.63
TS	0.61	0.54	0.59
TS-Exp	0.63	0.60	0.61

TABLE 15. Analysis of the self-term expansion methodology by using the *K*-Star clustering method over the *hep-ex* corpus.

	F_1-measure		
Approach name	MAX	MIN	AVG
TP	0.57	0.31	0.47
TP-Exp	0.76	0.67	0.73
DF	0.59	0.55	0.56
DF-Exp	0.83	0.75	0.77
TS	0.58	0.24	0.52
TS-Exp	0.78	0.66	0.74

reduced versions of the complete corpus. We used the three TSTs in order to obtain these reduced corpora (selecting from 10 to 90% of the complete vocabulary). As may be seen, in all the cases, the expanded approach (TP-Exp, DF-Exp and TS-Exp) outperforms the other one (TP, DF and TS), independently of the TST applied.

Table 15 shows the behaviour of the self-term expansion methodology, in this case, over the *hep-ex* corpus. This table, which was again obtained by executing the *K*-Star clustering method, shows the behaviour of each TST separately. We may see the obtained improvement which is independent of the TST used.

It is noteworthy that when we applied the self-term expansion before applying the TST, the best results are obtained with a very small size of the vocabulary. The discrimination of noisy terms is well performed by each TST. For the *hep-ex* corpus, in particular, we have seen that the DF technique is the one which performs better in comparison with the other two TSTs. In addition to the fact that the DF technique obtains the best F_1-measure results, it also reduces the corpus vocabulary to ∼90%.

The methodology performed better over the *hep-ex* corpus than over the *CICLing-2002* one. We consider that the moderated results obtained with the *CICLing-2002* corpus are justified by the small number of documents in the text collection. We conclude that the self-term expansion technique did not have enough contexts to determine the correct co-occurrence between terms of the corpus vocabulary.

In order to investigate the behaviour of the self-term expansion methodology by using another clustering method over the same corpus subset, we carried out further experiments employing what we have named the DK-Means clustering algorithm [19]. The difference between this method and the classic *K*-Means consists of the initialization of the clustering method in order to make it deterministic. In particular, we have initialized *K*-Means with the final results obtained by *K*-Star (including the clusters and their number).

The performance of each TSL with the *CICLing-2002* and the *hep-ex* corpora by using the DK-Means clustering method is shown in Figs 3 and 4, respectively.

With respect to the TSL used after the self-term expansion in order to reduce the vocabulary size, the best results were obtained employing DF (which is also very simple and quite easy to calculate). In the experiments presented, we may observe that independent of the corpus and clustering method used, the DF TST always performed well.

We may conclude that the task of clustering narrow domain short text corpora obtains better results if the original corpus has been previously enriched employing the self-term expansion methodology. This is due to the high vocabulary overlapping associated to this kind of corpora, which allows one to determine co-occurrence relationships that may be useful when expanding the original terms of the documents. Up to now, we have confirmed experimentally that this hypothesis is correct for these two narrow domain short text corpora. However, its veracity to other kinds of corpora has not been fully investigated.

With respect to the time complexity of the proposed approach, it depends on the time complexity of the co-ocurrence measure used, plus the time complexity of the term selection method. There are some methods and measures that behave worse in case of large size data sets than others. For instance, in the case of the TSTs used in this paper, the TS term selection method is worse than TP and DF.

Finally, we would like to discuss the performance of the expanding techniques across different percentages of vocabulary reduction. In Figs 3 and 4, we may see the variability of F_1 scores throughout the successive percentages of vocabularies selected by the TSTs. We may say that the self-term expansion process increases instability of TSTs with an unexpected behaviour (we did not discover a well-defined pattern). This issue is related with some corpus features such as the number of target classes, the size of these classes and the number of documents in the collection. Even if the instability of TSTs becomes an important issue to be investigated (see [23] for an example), we consider that the advantages of using a self-term expanded corpus over the non-expanded one in the clustering task are sufficient to consider the self-term expansion methodology an important contribution.

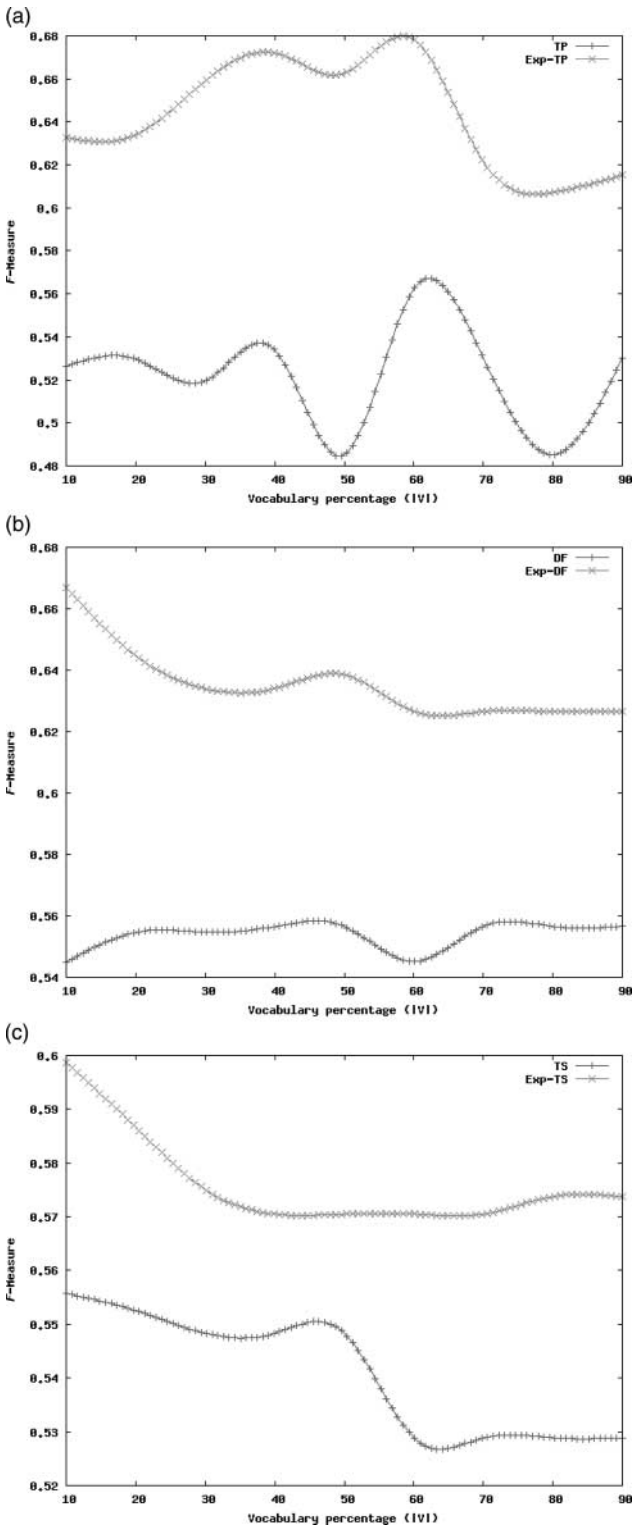

FIGURE 3. Analysis of the behaviour of each TST in the self-term expansion methodology by using the DK-Means clustering method on the *CICLing-2002* corpus. (a) *The TP technique*; (b) *The DF technique*; (c) *The TS technique*.

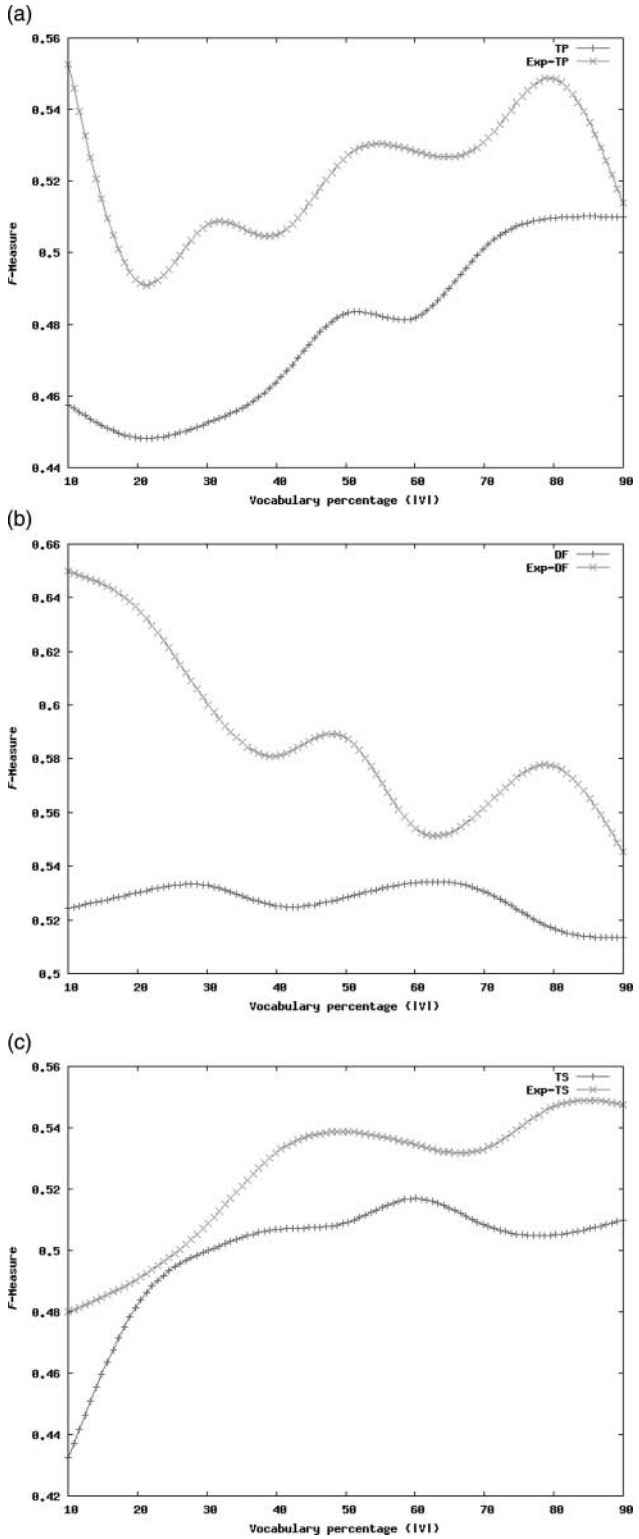

FIGURE 4. Analysis of the behaviour of each TST in the self-term expansion methodology by using the DK-Means clustering method on the *hep-ex* corpus. (a) *The TP technique*; (b) *The DF technique*; (c) *The TS technique*.

7. CONCLUSIONS

Clustering narrow domain short texts is a very challenging task, because of the *high overlapping* that exists among all the documents and the *low frequencies* that the corpora terms have. In order to tackle these two problems, we have presented a new methodology which increments the frequency of each term by expanding them with a set of co-related terms. Clustering narrow domain short texts by using the proposed methodology assumes that the two corpus features: *shortness* and *narrow domain* exist; therefore, we have introduced novel measures in order to automatically determine whether or not a corpus is made up of narrow domain short texts.

Two unsupervised formulae for domain broadness analysis were introduced, whereas three already known term frequency-based measures were used for the evaluation of shortness. We also introduced a novel unsupervised formula in order to determine whether or not a corpus is written with a similar writing style, even if the corpus is written by one or several authors. This last measure was used to distinguish scientific documents from those that are not.

Moreover, we introduced supervised variants of domain broadness detection, which could be easily used to evaluate classifications of clustering corpora provided by human experts (gold standard). We consider this issue of high relevance given the current use of gold standards in international competitions.

The self-term expansion methodology we have introduced allows the baseline corpus to be enriched by adding co-related terms from an automatically constructed lexical knowledge resource obtained from the *same* target data set (and not from an external resource). This was done by using two different co-occurrence techniques based on bigrams and PMI, respectively. The experiments showed that the PMI outperforms the bigrams co-occurrence technique given that the latter is statistically included in the former. Our empirical analysis has shown that it is possible to significantly improve clustering results by first performing the self-term expansion and optionally applying thereafter the term selection process.

The experiments were carried out on two real collections extracted from the CICLing-2002 conference and the CERN research centre. The corpora contain abstracts of scientific papers related to the computational linguistics domain and the high-energy particles' narrow domain, respectively. The self-term expansion method effectively improved the baseline F_1-measure by $\sim$ 30%. Furthermore, by using the term selection after expanding the corpus, we obtained a similar performance with a 90% reduction in the full vocabulary.

Until now, we have observed that the above behaviour is associated with the clustering of narrow-domain short text corpora since the enrichment process carried out by the methodology benefits from the high overlapping that usually exists in corpora of this kind. However, the enriching process is based on the frequency of term co-ocurrences and, therefore, the performance of the proposed methodology will be directly proportional to the number of documents in the collection to be clustered.

ACKNOWLEDGEMENTS

We thank the reviewers for their many helpful comments and suggestions.

FUNDING

This work was supported by MICINN project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I + D + i) and the CONACYT research project number 106625.

REFERENCES

[1] Sebastiani, F. (2002) Machine learning in automated text categorization. *ACM Comput. Surv.*, **34**, 1–47.

[2] Montejo-Ráez, A. and Ureña-López, L.A. (2006) Binary classifiers versus AdaBoost for labeling of digital documents. *Procesamiento del Lenguaje Natural*, **37**, 319–326.

[3] Rijsbergen, C.J.V. (1979) *Information Retrieval* (2nd edn) Department of Computer Science, University of Glasgow, Glasgow, Scotlan, UK.

[4] Kowalski, G. (1997) *Information Retrieval Systems Theory and Implementation*. Kluwer Academic Publishers, Norwell, MA, USA.

[5] Hearst, M.A. and Pedersen, J.O. (1996) Re-examining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. *Proc. 19th Annual Int. ACM SIGIR Conf. Research and Development in Information Retrieval—SIGIR'96*, Zurich, Switzerland, August 18–22, pp. 76–84. Association for Computing Machinery (ACM), New York, NY, USA.

[6] Jardine, N. and van Rijsbergen, C.J. (1971) The use of hierarchical clustering in information retrieval. *Inf. Storage Retr.*, **7**, 217–240.

[7] Tombros, A., Villa, R. and van Rijsbergen, C.J. (2002) The effectiveness of query-specific hierarchic clustering in information retrieval. *Inf. Process. Manag.*, **38**, 559–582.

[8] Buckley, C. and Lewit, A.F. (1985) Optimizations of Inverted Vector Searches. *Proc. 8th Annual Int. ACM SIGIR Conf. Research and Development in Information Retrieval—SIGIR'85*, Montreal, Quebec, Canada, June 5–7, pp. 97–110. Association for Computing Machinery (ACM), New York, NY, USA.

[9] Manning, D.C. and Schütze, H. (2003) *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

[10] Pinto, D., Benedí, J.M. and Rosso, P. (2007) Clustering Narrow-Domain Short Texts by Using the Kullback–Leibler Distance. *Proc. CICLing Conference—CICLing'07*, Mexico city, Mexico, February 18–24, Lecture Notes in Computer Science 4394, pp. 611–622. Springer, Berlin.

[11] Moussiades, L. and Vakali, A. (2005) Pdetect: a clustering approach for detecting plagiarism in source code datasets. *Comput. J.*, **48**, 656–670.

[12] Zelikovitz, S. and Hirsh, H. (2000) Transductive LSI for Short Text Classification Problems. *Proc. 17th Int. Conf. Machine Learning—ICML'00*, Palo Alto, CA, USA, June 29–July 2, pp. 1183–1190. Morgan Kaufmann, San Francisco, CA.

[13] Hynek, J., Jezek, K. and Rohlik, O. (2000) Short Document Categorization Itemsets Method. *Proc. 4th Practice of Knowledge Discovery in Databases—PKDD'00*, Lyon, France, September 13–16, Lecture Notes in Computer Science 1910, pp. 9–14. Springer, Berlin.

[14] Zelikovitz, S. and Marquez, F. (2005) Transductive learning for short-text classification problems using latent semantic indexing. *Int. J. Pattern Recognit. Artif. Intell.*, **19**, 143–163.

[15] Pu, Q. and Yang, G.-W. (2006) Short-Text Classification based on ICA and LSA. *Proc. 3rd Int. Symp. Neural Networks (Advances in Neural Networks)—ISNN'06*, Chengdu, China, May 28–June 1, Lecture Notes in Computer Science 3971, pp. 265–270. Springer, Berlin.

[16] Brooks, C.H. and Montanez, N. (2006) An Analysis of the Effectiveness of Tagging in Blogs. *Proc. AAAI Spring Symp. Computational Approaches to Analyzing Weblogs*, Stanford, CA, USA, March 21–23, pp. 9–14. AAAI Spring Symposium Series, CA, USA.

[17] Banerjee, S., Ramanathan, K. and Gupta, A. (2007) Clustering Short Texts Using Wikipedia. *Proc. 30th Annual Int. ACM SIGIR conference on Research and Development in Information Retrieval—SIGIR'07*, Amsterdam, The Netherlands, July 23–27, pp. 787–788. Association for Computing Machinery (ACM), New York, NY, USA.

[18] Peng, J., Yang, D.-Q., Wang, J.-W., Wu, M.-Q. and Wang, J.-G. (2007) A Clustering Algorithm for Short Documents Based on Concept Similarity. *Proc. IEEE Pacific Rim Conf. Communications, Computers and Signal Processing—PACRIM'07*, Victoria, B.C., Canada, August 22–24, pp. 42–45. IEEE, Piscataway, NJ, USA.

[19] Pinto, D. (July 2008) On clustering and evaluation of narrow domain short-text corpora. PhD Dissertation, Polytechnic University of Valencia, Spain.

[20] Steinbach, M., Karypis, G. and Kumar, V. (2000) A Comparison of Document Clustering Techniques. *Proc. 6th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (workshop on text mining)—KDD'00*, Boston, MA, USA, August 20–23, pp. 109–110. Association for Computing Machinery (ACM), New York, NY, USA.

[21] Herdan, G. (1964) *Quantatative Linguistics*. Butterworth, London.

[22] Makagonov, P., Alexandrov, M. and Gelbukh, A. (2004) Clustering Abstracts Instead of Full Texts. *Proc. Text, Speech and Dialogue Conference—TSD'04*, Brno, Czech Republic, September 8–11, Lecture Notes in Artificial Intelligence 3206, pp. 129–135. Springer, Berlin.

[23] Pinto, D., Jiménez-Salazar, H. and Rosso, P. (2006) Clustering Abstracts of Scientific Texts Using the Transition Point Technique. *Proc. CICLing Conference—CICLing'06*, Mexico city, Mexico, February 19–25, Lecture Notes in Computer Science 3878, pp. 536–546. Springer, Berlin.

[24] Alexandrov, M., Gelbukh, A. and Rosso, P. (2005) An Approach to Clustering Abstracts. *Proc. 10th Int. Conf. Application of Natural Language to Information Systems—NLDB'05*, Alicante, Spain, June 15–17, Lecture Notes in Computer Science 3513, pp. 8–13. Springer, Berlin.

[25] Makagonov, P., Alexandrov, M. and Sboychakov, K. (2000) Keyword-based Technology for Clustering Short Documents (Selected Papers). *Comput. Res.*, **2**, 105–114.

[26] Homayouni, R., Heinrich, K., Wei, L. and Berry, M.W. (2005) Gene clustering by latent semantic indexing of medline abstracts. *Bioinformatics*, **21**, 104–115.

[27] Buscaldi, D., Juan, A., Rosso, P. and Alexandrov, M. (2006) Sense Cluster-based Categorization and Clustering of Abstracts. *Proc. CICLing Conf.—CICLing'06*, Mexico city, Mexico, February 19–25, Lecture Notes in Computer Science 3878, pp. 547–550. Springer, Berlin.

[28] Ingaramo, D., Pinto, D., Rosso, P. and Errecalde, M. (2008) Evaluation of Internal Validity Measures in Short-Text Corpora. *Proc. CICLing Conference—CICLing'08*, Haifa, Israel, February 17–23, Lecture Notes in Computer Science 4919, pp. 555–567. Springer, Berlin.

[29] Stein, B. and Nigemman, O. (1999) On the Nature of Structure and its Identification. *Proc. 25th Int. Workshop on Graph-Theoretic Concepts in Computer Science—WG'99*, Ascona, Switzerland, June 17–19, Lecture Notes in Computer Science 1665, pp. 122–134. Springer, Berlin.

[30] Fellbaum, C. (1998) *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, USA.

[31] Montejo-Ráez, A. (2006) Automatic text categorization of documents in the high energy physics domain. PhD Dissertation, Granada University, Spain.

[32] Pinto, D. and Rosso, P. (2006) KnCr: A Short-Text Narrow-Domain Sub-corpus of Medline. *Proc. Human Language Technologies Conference—TLH'06*, San Luis Potosí, Mexico, September 18–20, Advances in Computer Science, pp. 266–269. ENC, Mexico.

[33] Debole, F. and Sebastiani, F. (2005) An analysis of the relative hardness of Reuters-21578 subsets. *J. Am. Soc. Inf. Sci. Technol.*, **56**, 584–596.

[34] Wibowo, W. and Williams, H.E. (1999) On Using Hierarchies for Document Classification. *Proc. 4th Australasian Document Computing Symposium*, Coffs Harbour, Australia, December, pp. 31–37. Australasian Document Computing Symposium, Australia.

[35] Brants, T., Popat, A.C., Xu, P., Och, F.J. and Dean, J. (2007) Large Language Models in Machine Translation. *Proc. Joint Conf. Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, Prague, Czech Republic, June, pp. 858–867. Association for Computational Linguistics.

[36] Màrquez, L. and Padró, L. (1997) A Flexible POS Tagger Using an Automatically Acquired Language Model. *Proc. 35th Annual Meeting on Association for Computational Linguistics—ACL'97*, Madrid, Spain, July 7–12, pp. 238–245. Association for Computational Linguistics.

[37] Ponte, J.M. and Croft, W.B. (1998) A Language Modeling Approach to Information Retrieval. *Proc. 21th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval—SIGIR'98*, Melbourne, Australia, August

24–28, pp. 275–281. Association for Computing Machinery (ACM), New York, NY, USA.

[38] Bahl, L.R., Jelinek, E. and Mercer, R.L. (1983) A maximum likelihood approach to continuous speech recognition. *IEEE Trans. Pattern Analy. Mach. Intell.*, **5**, 179–190.

[39] Brown, P.F., Pietra, V.J.D., deSouza, P.V., Lai, J.C., and Mercer, R.L. (1992) Class-based n-gram models of natural language. *Comput. Linguist.*, **18**, 467–479.

[40] Roukos, S. (1997) Language Representation. Survey of the State of the Art in Human Language Technology. Cambridge University Press, Cambridge, MA, USA.

[41] Herdan, G. (1960) *Type-Token Mathematics: A Textbook of Mathematical Linguistics*. Mouton & Co., The Hague, The Netherlands.

[42] Diederich, J., Kindermann, J., Leopold, E. and Paass, G. (2004) Authorship attribution with support vector machines. *Appl. Intell.*, **19**, 109–123.

[43] Can, F. and Patton, J. M. (2004) Change of writing style with time. *Comput. Humanit.*, **38**, 61–82.

[44] Hoover, D.L. (2007) Corpus stylistics, stylometry, and the styles of henry james. *Style*, **41**, 174–203.

[45] Zipf, G.K. (1949) *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA, USA.

[46] Hotho, A., Staab, S. and Stumme, G. (2003) Ontologies Improve Text Document Clustering. *Proc. 3rd IEEE Int. Conf. Data Mining—ICDM'03*, San Francisco, CA, USA, May 1–3, pp. 1–4. SIAM, Philadelphia, PA, USA.

[47] Sedding, J. and Kazakov, D. (2004) WordNet-based Text Document Clustering. *Proc. 20th Conf. Computational Linguistics* (*3rd Workshop on Robust Methods in Analysis of Natural Language Data*)*—COLING'04*, Geneva, Switzerland, August 23–27, pp. 104–113. COLING, Switzerland.

[48] Reforgiato-Recupero, D. (2007) A new unsupervised method for document clustering by using WordNet lexical and conceptual relations. *Inf. Retr.*, **10**, 563–579.

[49] Church, K., Gale, W., Hanks, P. and Hindle, D. (1991) Using Statistics in Lexical Analysis. In Zernik, U. (ed.), *Lexical Acquisition: Exploiting On-lines Resources for Build a Lexicon*, pp. 115–164. Erlbaum, Englewood Cliff, NJ.

[50] Pinto, D., Rosso, P., and Jiménez-Salazar, H. (2007) UPV-SI: Word Sense Induction Using Self Term Expansion. *Proc. 4th Int. Workshop on Semantic Evaluations—SemEval'07*, Prague, Czech republic, June 23–24, pp. 430–433. Association for Computational Linguistics (ACL), Stroudsburg, PA, USA.

[51] Harris, Z. (1954) Distributional structure. *Word*, **10**, 146–162.

[52] Yang, Y. (1995) Noise Reduction in a Statistical Approach to Text Categorization. *Proc. 18th Annual Int. ACM SIGIR Conf. Research and Development in Information retrieval—SIGIR'95*, Seattle, WA, United States, July 9–13, pp. 256–263. Association for Computing Machinery (ACM), New York, NY, USA.

[53] Pekar, V., Krkoska, M. and Staab, S. (2004) Feature Weighting for Co-occurrence-based Classification of Words. *Proc. 20th Conf. Computational Linguistics—COLING'04*, Geneva, Switzerland, August 23–27, pp. 799–805. COLING, Switzerland.

[54] Booth, A.D. (1967) A law of occurrences for words of low frequency. *Inf. control*, **10**, 386–393.

[55] Porter, M.F. (1980) An algorithm for suffix stripping. *Program*, **14**, 130–137.

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

[57] Rodgers, J.L. and Nicewander, W.A. (1988) Thirteen ways to look at the correlation coefficient. *Am. Stat.*, **42**, 59–66.

[58] Lehmann, E.L. and D'Abrera, H.J.M. (1998) *Nonparametrics: Statistical Methods Based on Ranks*. Prentice-Hall, Englewood Cliffs, NJ.

[59] Kendall, M. (1938) A new measure of rank correlation. *Biometrika*, **30**, 81–89.

[60] Shin, K. and Han, S.Y. (2003) Fast Clustering Algorithm for Information Organization. *Proc. CICLing Conf.—CICLing'03*, Mexico city, Mexico, February 16–22, Lecture Notes in Computer Science 2588, pp. 619–622. Springer, Berlin.