Universidad Politécnica de Valencia

Departamento de Sistemas Informáticos y Computación

Reconocimiento de Formas e Inteligencia Artificial

On Clustering and Evaluation of Narrow Domain Short-Text Corpora

The accepted dissertation

of

David Eduardo Pinto Avendaño

in partial fulfillment of the requirements

to obtain the academic degree of

Doctor en Informática

under the supervision of

Dr. Paolo Rosso

Universidad Politécnica de Valencia (España)

Dr. Héctor Jiménez Salazar

Universidad Autónoma Metropolitana (México)

Valencia

July 2008

Acknowledgments

This Ph.D. thesis has mainly been supported by the following grant:

BUAP-701 PROMEP/103.5/-05/1536: The funds are given by both the "Secretaría de Educación Pública" (SEP) and the B. Autonomous University of Puebla (BUAP) in the framework of the "PROgrama para el MEjoramiento del Profesorado" (PROMEP) of Mexico.

This research work has also been supported by the following projects:

- MCyT TIN2006-15265-C06-04.
- PCI-AECI A/7067/06.
- "Programa de Apoyo a la Investigación y Desarrollo" (PAID-06-06) of the Universidad Politécnica de Valencia.

Dedicated to my family:

Sofia Paniagua Rivera,
You always has been there with me, supporting my dreams.
You know how much I love you.

David Pinto Paniagua & Angel Pinto Paniagua,
I hope that when you are old enough to read this book,
you are as proud of me as I am of both of you.

David Angel Pinto Paniagua,
You are the light at the end of the tunnel.
We love you, We miss you ...

Abstract

In this Ph.D. thesis we investigate the problem of clustering a particular set of documents namely *narrow domain short texts*.

To achieve this goal, we have analysed datasets and clustering methods. Moreover, we have introduced some corpus evaluation measures, term selection techniques and clustering validity measures in order to study the following problems:

1. To determine the relative hardness of a corpus to be clustered and to study some of its features such as *shortness*, *domain broadness*, *stylometry*, *class imbalance* and *structure*.

2. To improve the state of the art of clustering narrow domain short-text corpora.

The research work we have carried out is partially focused on "short-text clustering". We consider this issue to be 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).

Moreover, we study the domain broadness of corpora. A corpus may be considered to be *narrow* or *wide* domain if the level of the document vocabulary overlapping is high or low, respectively. In the categorization task, it is very difficult to deal with narrow domain corpora such as scientific papers, technical reports, patents, etc.

The aim of this research work is to study possible strategies to tackle the following two problems:

a) the low frequencies of vocabulary terms in short texts, and

b) the high vocabulary overlapping associated to narrow domains.

Each problem alone is challenging enough, however, dealing with narrow domain short texts increases the complexity of the problem significantly.

The clustering of scientific abstracts is even more difficult than the clustering of narrow domain short-text corpora. The reason is that texts belonging to scientific

papers often make use of 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. However, 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.

Due to the dynamic aspect of research, new interests could arise in a field and new sub-topics 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 this thesis, we deal with the treatment of narrow domain short-text collections in three areas: *evaluation*, *clustering* and *validation* of corpora.

The major contributions of the investigations carried out are:

1. The study and introduction of evaluation measures to analyse the following features of a corpus: *shortness*, *domain broadness*, *class imbalance*, *stylometry* and *structure*.

2. The development of the Watermarking Corpora On-line System, named WaCOS, for the assessment of corpus features.

3. A new unsupervised methodology (which does not use any external knowledge resource) for dealing with narrow domain short-text corpora. This methodology suggests first applying self-term expansion and then term selection.

We analysed different corpus features as evidence of the relative hardness of a given corpus with respect to clustering algorithms. In particular, the degree of *shortness*, *domain broadness*, *class imbalance*, *stylometry* and *structure* were studied.

We introduced some (un)supervised measures in order to assess these features. The supervised measures were used both to evaluate the corpus features and, even more importantly, to assess the gold standard provided by experts for the corpus to be clustered. The unsupervised measures evaluate the document collections directly

(i.e., without any gold standard) and, therefore, they may also be used for other purposes, for instance, to adjust clustering methods while being executed in order to improve the results.

The most successful measures were compiled in a freely functional web-based system that allows linguistics and computational linguistics researchers to easily assess the quality of corpora with respect to the aforementioned features.

The experiments conducted confirmed that the clustering of narrow domain short-text corpora is a very challenging task. However, the contributions of this research work are proof that it is possible to deal with this difficult problem as well as improve the results obtained with classical techniques and methods.

Resumen

En este trabajo de tesis doctoral se investiga el problema del agrupamiento de conjuntos especiales de documentos llamados *textos cortos de dominios restringidos*.

Para llevar a cabo esta tarea, se han analizados diversos corpora y métodos de agrupamiento. Mas aún, se han introducido algunas medidas de evaluación de corpus, técnicas de selección de términos y medidas para la validez de agrupamiento con la finalidad de estudiar los siguientes problemas:

1. Determinar la relativa difficultad de un corpus para ser agrupado y estudiar algunas de sus características como *longitud de los textos*, *amplitud del dominio*, *estilometría*, *desequilibrio de clases* y *estructura*.

2. Contribuir en el estado del arte sobre el agrupamiento de corpora compuesto de textos cortos de dominios restringidos

El trabajo de investigación que se ha llevado a cabo se encuentra parcialmente enfocado en el "agrupamiento de textos cortos". Este tema se considera relevante dado el modo actual y futuro en que las personas tienden a usar un "lenguaje reducido" constituidos por textos cortos (por ejemplo, blogs, snippets, noticias y generación de mensajes de textos como el correo electrónico y el chat).

Adicionalmente, se estudia la amplitud del dominio de corpora. En este sentido, un corpus puede ser considerado como *restringido* o *amplio* si el grado de traslape de vocabulario es alto o bajo, respectivamente. En la tarea de categorización, es bastante complejo lidiar con corpora de dominio restringido tales como artículos científicos, reportes técnicos, patentes, etc.

El objetivo principal de este trabajo consiste en estudiar las posibles estrategias para tratar con los siguientes dos problemas:

a) las bajas frecuencias de los términos del vocabulario en textos cortos, y

b) el alto traslape de vocabulario asociado a dominios restringidos.

Si bien, cada uno de los problemas anteriores es un reto suficientemente alto, cuando se trata con textos cortos de dominios restringidos, la complejidad del problema se incrementa significativamente.

El agrupamiento de resúmenes de artículos científicos es aún mas difícil que el agrupamiento de textos cortos de dominios restringidos. La razón es que los textos que pertenecen a artículos científicos a menudo usan secuencias de palabras tales como "en este artículo se presenta", "el objetivo principal es", "los resultados obtenidos", etc., lo cual obviamente incrementa el grado de similitud entre diferentes conjuntos de textos cortos. De esta manera, la correcta selección de términos cuando se agrupan textos es muy importante debido a que los resultados pueden variar significativamente.

El propósito de estudiar resúmenes de artículos científicos no está motivado de manera exclusiva por la alta complejidad de esta tarea, sino también porque en la mayoría de las bibliotecas digitales y otros repositorios (basados en el web) de información científica y técnica proporcionan acceso gratuito únicamente a los resúmenes y no a los textos completos.

Debido a la naturaleza dinámica de la investigación, nuevos intereses pueden surgir en una cierta área y nuevos sub-temas necesitan ser descubiertos a través de técnicas de agrupamiento con la finalidad de introducirlos posteriormente como nuevas categorías. Por lo tanto, el agrupamiento de resúmenes científicos viene a ser una necesidad real.

En esta tesis, se investiga el tratamiento de colecciones de textos cortos de dominios restringidos siguiendo tres ejes: *evaluación*, *agrupamiento* y *validación*.

Las contribuciones mayores de este trabajo doctoral son las siguientes:

1. El estudio y la introducción de medidas de evaluación para el análisis de las siguientes características de un corpus: *longitud de los textos*, *amplitud de dominio*, *desequilibrio de clases*, *estilometría* y *estructura*.

2. El desarrollo del sistema WaCOS (Watermarking Corpora On-line System) para la evaluación de características de corpus.

3. Una nueva metodología no supervisada (que no hace uso de recursos de conocimiento

externos) para tratar con corpora constituido de textos cortos de dominios restringidos. Esta metodología sugiere aplicar primero auto-expansión de términos y posteriormente una reducción de vocabulario mediante selección de terminos.

Se analizan diferentes características de corpus como una evidencia de la relativa dificultad de un corpus dado con respecto a ciertos algoritmos de agrupamiento. En particular, se estudia la *longitud de los textos*, *amplitud de dominio*, *desequilibrio de clases*, *estilometría* y *estructura*.

Se introducen algunas medidas supervisadas y no supervisadas para evaluar las características mencionadas anteriormente. Las medidas supervisadas se usan tanto para evaluar dichas características como para evaluar el gold estándar proporcionado por los expertos. Esto último se considera de gran relevancia. Por otro lado, las medidas no supervisadas evaluan las colecciones de documentos de manera directa (es decir, sin ningún gold estándar) y por lo tanto, pueden ser usadas con otros propósitos, por ejemplo, para ajustar parámetros de algoritmos de agrupamiento (durante su ejecución) con la finalidad de mejorar los resultados.

Las medidas de evaluación fueron integradas en un sistema gratuito y totalmente funcional basado en el web que permite a lingüistas puros y lingüistas computacionales evaluar fácilmente la calidad de corpora con respecto a las características antes mencionadas.

Los experimentos llevados a cabo confirman que el agrupamiento de textos cortos de dominios restringidos es una tarea difícil. Sin embargo, las contribuciones de este trabajo de investigación son evidencia de que es posible lidiar con este problema y además obtener mejoras en los resultados con respectos a aquellos obtenidos con técnicas y métodos clásicos.

Contents

List of tables

List of figures

Notation

Symbol	Meaning
D	Document collection
d,d_i	A document, the i-th document
$\overrightarrow{d_i}$	The vectorial representation of the i-th document
$\|D\|$	Cardinality of D, number of documents in D
$V(D)$	Vocabulary of the document collection D
$V(d)$	Vocabulary of the document d
t, t_i	A term, the i-th term
$tf(t_i, d_j)$	Term frequency of t_i within document d_j
$tf(t_i, D)$	Term frequency of t_i in collection D
tf_{ij}	Normalized term frequency of t_i in document d_j
$idf(t_i)$	Inverse document frequency of the term t_i in corpus D
$icf(t_i)$	Inverse class frequency of the term t_i in corpus D
$cf(t_i)$	The number of classes where the term t_i appears
$\mathcal{C}$	Clustering/categorization
$\mathcal{C}^{(i)}$	Clustering done by using only i vocabulary terms
C, C_i	Cluster set, i-th cluster obtained/set of documents
$\bar{C}_i$	The complement set of C_i
$\mathcal{C}^*$	Desired categorization/gold standard
C^*, C_i^*	Class set, i-th class/set of documents

Symbol	Meaning
$\bigcup$	Union
$\bigcap$	Intersection
$\mathbb{R}^+$	The positive real numbers
$\mathbb{N}$	The natural numbers
$\varphi : D \times D \rightarrow \mathbb{R}^+$	Similarity measure
$D_{KL}(P\|\|Q)$	The Kullback-Leibler divergence between P and Q
$D_{KLD}(P\|\|Q)$	The symmetrical Kullback-Leibler distance
$P(x)$	Probability of x
$P(A\|B)$	Conditional probability of A given B
$F(C_i, C_j^*)$	F-Measure of C_i with respect to C_j^*
F	Global F-Measure
$\mathcal{M}$	Similarity matrix
$Precision(C_i, C_j^*)$	Precision of C_i with respect to C_j^*
$Recall(C_i, C_j^*)$	Recall of C_i with respect to C_j^*

Chapter 1

Introduction

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

The challenges that researchers must deal with when working, for instance with web pages, are related to the structure of the web document content. Major web pages are written in natural language, and very often without any specific helpful structure. In other words, it is a problem of processing almost pure raw data, which is not an easy task.

We are particularly interested in the analysis of clustering and evaluation methods for text corpora. Document clustering consists of the assignment of documents to unknown categories. This task may be considered to be more difficult than supervised text categorization [127, 87] because the information about the category name and the correct structure of categorized documents is not provided in advance. The clustering of documents has been approached in different areas of text processing, such as text mining, summarization and information retrieval. In [117] and [67], for instance, the way 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 has also been studied in the literature. However, the difficulty of finding clustering methods that perform well on different data collections is a problem that

have existed for many years [23].

There exist sufficient examples that justify the study of document clustering for the analysis of Internet documents. 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 eleven 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 as an answer will be irremediably affected by each sense frequency of the word "Cancer" on 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, Vivísimo, Mooter y KartOO[2]); however, as we mentioned above, the accuracy may be affected by the frequency of the terms query on Internet and also by the possible ontologies used in the term clustering process.

Applications in different areas of natural language processing may include reranking of snippets in information retrieval, and automatic clustering of scientific texts available on the Web [105].

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 Internet. Therefore, there exists sufficient interest from the computational linguistic community to analyse the behaviour of classifiers when using short-text corpora [153, 54, 154, 111, 21, 11, 97]. If the short texts belong to the same domain (e.g. sports or physics) we say that they are narrow domain texts. If it is already difficult to cluster short texts, then if those documents are also narrow domain it greatly increases the complexity of the task.

The aim of this Ph.D. thesis is to investigate the problem of clustering a particular set of documents namely *narrow domain short texts*. To achieve this goal, we have

[1] `http://en.wikipedia.org/wiki/Cancer_(disambiguation)`

[2] `http://clusty.com/`;`http://vivisimo.com/`;`http://mooter.com/`;`http://kartoo.com/`

analysed datasets and clustering methods. Moreover, we have introduced some corpus evaluation measures, term selection techniques and clustering validity measures in order to study the following problems:

1. To determine the relative hardness of a corpus to be clustered and to study some of its features such as *shortness*, *domain broadness*, *stylometry*, *class imbalance* and *structure*.

2. To improve the state of the art of clustering narrow domain short-text corpora.

The rest of this chapter is structured as follows. In Section 1.1, we briefly describe the difficulties that need to be resolved when classifying short texts. Section 1.2 explains the problem of dealing with narrow domain corpora. Section 1.3 describes the great challenge of dealing with documents that are both narrow domain and short texts. Section 1.4 shows a particularisation of narrow domain short-text corpora, that is, scientific abstracts. Section 1.5 summarizes the challenges faced in this study. Section 1.6 highlights the contributions of this research work and finally, Section 1.7 presents the structure of this Ph.D. thesis.

1.1 Short texts

The research work we have carried out is partially focused on "short-text clustering". We consider this issue to be 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).

Short text corpora are text collections made up of documents containing few words as content. The principal characteristic of short texts is that the frequency of the terms is relatively low in comparison with their frecuency 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.

Formally, given a document d with vocabulary size $|V(d)|$ and the corresponding Short Representation of d ($SR(d)$), we may compute the Shortness Degree of d as

$SD(d) = \frac{log|V(d)|}{log|d|}$. For instance, if we have both, a full document d_F containing 1,700 words with a vocabulary size of 530, and a short representation of the same document $SR(d_F)$ (say an abstract) with cardinality 70 and vocabulary size equal to 48, the shortness of d_F and $SR(d_F)$ 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 consider that the equality $|V(d)| = |d|^{SD(d)}$ expresses the shortness degree of d, and, therefore, the vocabulary size is assumed to be a simple power function of $|d|$. The closer $SD(d)$ is to one, the shorter the document is. A *short text* (let us say 200-500 words) will have $SD(d) \approx 1$, whereas *very short texts*, such as a query input in a search engine (let us say 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 the fourth chapter of the thesis.

The differences between short texts and long documents for document representation and management is mainly twofold: high vocabulary dimensions and sparse data spaces. The average document similarity of short text collections is very low. 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 [99].

1.2 Narrow domain corpora

A corpus may be considered to be *narrow* or *wide* domain if the level of the document vocabulary overlapping is high or low, respectively. In the categorization task, it is very difficult to deal with narrow domain corpora such as scientific papers, technical reports, patents, etc [100].

In [6], vocabulary overlapping is calculated for the documents from the most different groups of a corpus that is composed of scientific documents from the computational linguistics field (e.g. “ambiguity” and “text processing”). They obtained about 70% of vocabulary overlapping between the two categories, which implies that the selected domain is rather *narrow*. Although it is desirable to assign each docu-

ment to one of these two categories, under normal conditions the documents could be merged into only one set by the application of almost any classifier. The complexity of clustering narrow domain corpora is highlighted in [129].

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* domain or *wide* domain. In Chapter 4 we introduce different approaches and formulae to calculate the degree of domain broadness of a corpus from a supervised and unsupervised viewpoint.

1.3 Narrow domain short-text corpora

The aim of this research work is to study possible strategies to tackle the following two problems:

a) the low frequencies of vocabulary terms in short texts, and

b) the high vocabulary overlapping associated to narrow domains.

Each problem alone is challenging enough, however, dealing with narrow domain short texts increases the complexity of the problem significantly.

In the literature, there exist some works that have studied the classification of narrow domain short-text corpora [78, 77, 6, 100, 26, 98, 57]. All of them agree about the high level of difficulty that is faced when classifying documents of this kind. The reason for this can be explained by the following analyses.

On the one hand, even if a document set is made up of short texts, if the vocabulary overlapping is low, the classification may be carried out easily. The reason is that it is easy to distinguish among the categories of the given corpus.

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 of term overlapping is still possible.

Therefore, 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.

Collections of scientific documents are an example of narrow domain short texts and, therefore, abstracts of scientific papers are a particularisation of narrow domain short-text corpora. We are implicitly interested in studying documents of this kind. In the following section, we present the challenges that arise with respect to the clustering of scientific abstract collections.

1.4 Scientific abstracts

The clustering of scientific abstracts is even more difficult than the clustering of narrow domain short-text corpora. The reason is that texts belonging to scientific papers often make use of 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. However, the correct selection of terms when clustering texts is very important because the results may vary significantly.

In fact, in [6], 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 one or two, sometimes three or four. In this situation, changing a keyword frequency by one can significantly change the clustering results.

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[3], the ACM[4], and others receive hundreds of publications that must be

[3]Conseil Européen pour la Recherche Nucléaire
[4]Association for Computing Machinery

categorized in some specific domain, often with an unknown number of categories a priori.

Let us take for instance, the PubMed[5], which is 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 an automatic computational-based system. Novel methods for classifying narrow domain short texts must be constructed to deal with this real problem.

Some approaches have tackled this particular problem with sucessful results. However, the applications are domain-dependent since they made use of supervised classifiers that were trained with data that were tagged with keywords extracted from domain-dependent thesauri [88]. However, in scientific domains, rarely there are linguistic resources to help in supervised categorization tasks due to the specific or narrow vocabulary 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 [104].

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

1.5 Challenges: corpora evaluation, clustering & validity

Once the high level of complexity that is involved when working with narrow domain short-text corpora has been clarified, we would like to highlight the *challenges* that must be faced when dealing with this particular kind of collection.

As stated above, there is no standard formula to measure the degree of domain broadness of a given corpus. The first challenge of this Ph.D. thesis is to propose a framework for the assessment of a set of corpus features that would be useful

[5] `http://www.ncbi.nlm.nih.gov`

to understand the nature of the documents from viewpoint of the *shortness* and *broadness*. The proposed measures will allow us to evaluate the relative hardness of corpora to be clustered and to study additional corpus features such as the particular writing style of scientific researchers. In general, we expect to be able to distinguish corpora that are composed of narrow domain short texts from those that are not.

By determining the degree of broadness and shortness of corpora we can test clustering methods in order to determine the complexity of classifying text collections of this type. This will enable us to analyse the possible components that could improve the obtained accuracy in the clustering task. This implies improving the state of the art of clustering narrow domain short text corpora.

Finally, a last challenge is to validate clustering results in the two following ways: First, we are interested in applying internal clustering validity measures in order to "validate" the quality of the obtained clusters by a given clustering method. Second, we also want to employ similar measures in order to assess the quality of gold standards.

1.6 Thesis contributions

In the above sections, we have presented the relevance of the problem to be discussed in this Ph.D. thesis. The purpose of this brief description is to present a general overview of the feasibility and the level of challenge of this work. In this thesis, we will deal with the treatment of narrow domain short-text collections in three areas: *evaluation*, *clustering* and *validation* of corpora.

The major contributions of the investigations carried out are:

1. The study and introduction of evaluation measures to analyse the following features of a corpus: *shortness*, *domain broadness*, *class imbalance*, *stylometry* and *structure* (See Chapter 4).

2. The development of the Watermarking Corpora On-line System, named WaCOS, for the assessment of corpus features (See Chapter 4).

3. A new unsupervised methodology (which does not use any external knowledge resource) for dealing with narrow domain short-text corpora (See Chapter 5). This methodology suggests first applying self-term expansion and then term selection.

We analysed different corpus features as evidence of the relative hardness of a given corpus with respect to clustering algorithms. In particular, the degree of *shortness*, *domain broadness*, *class imbalance*, *stylometry* and *structure* were studied.

We introduced some (un)supervised measures in order to assess these features. The supervised measures were used both to evaluate the corpus features and, even more importantly, to assess the gold standard provided by experts for the corpus to be clustered. The unsupervised measures evaluate the document collections directly (i.e., without any gold standard) and, therefore, they may also be used for other purposes, for instance, to adjust clustering methods while being executed in order to improve the results.

The most successful measures were compiled in a freely functional web-based system that allows linguistics and computational linguistics researchers to easily assess the quality of corpora with respect to the aforementioned features.

The new method of document representation based on self-term expansion highly improves the results of clustering narrow domain short-texts when using a classical document representation, such as the one based only on "bag of words". Moreover, the document representation technique proposed in this work makes it possible to obtain results similar to those that make use of external knowledge resources.

This fact is remarkable since, as mentioned above, there rarely exist linguistic resources in narrow domains to help in supervised categorization tasks due to the specific vocabulary of the documents. Self-term expansion allows a thesaurus to be obtained from the *same* dataset and then used to expand its own terms. Our study also investigates the performance of using this self-term expansion when different term selection techniques are employed. We have found that the best combination is to first expand the corpus and then to apply a term selection technique. Specifically, when we carried out experiments on the corpus of high energy particles domain (physics),

we observed that it was possible to improve the baseline by approximately 40% by using only the term expansion method. Furthermore, by using term selection after expanding the corpus, we obtained a similar performance with a 90% reduction in the full vocabulary.

The methodology proposed here can also be used in other practical applications such as automatic summary generation, clustering of snippets, homonymy discrimination, etc. In fact, we applied the self-term expansion methodology in one practical task known as *word sense induction*. We obtained the best results (with respect to completely unsupervised systems) in the international competition organised by the Association for Computational Linguistics.

The experiments conducted confirmed that the clustering of narrow domain short-text corpora is a very challenging task. However, the contributions of this research work are proof that it is possible to deal with this difficult problem as well as improve the results obtained with classical techniques and methods.

1.7 Thesis overview

The structure of this Ph.D. thesis is the following. Chapter 2 gives an overview of the clustering methods, clustering measures, term selection techniques and datasets used in this study. We decided to include this information at the beginning of the thesis in order to provide a fast reference for the items in this document that will be referred to frequently.

In Chapter 3, we analyse the implications of clustering narrow domain short-text corpora, studying the role of the term selection process as well as the instability of a term selection technique based on the selection of mid-frequency terms. We also make a comparison of different clustering methods in the narrow domain short-text framework. A similarity measure based on the distribution of term frequencies is proposed. Finally, we evaluate the performance of the term selection techniques on a standard narrow domain short-text corpus.

Chapter 4 proposes the use of several measures (most of which are introduced in this work) to assess different corpus features. These measures are tested on several

corpora and implemented in an on-line web-based system named WaCOS.

Chapter 5 presents a new methodology (based on term co-occurrence) for improving document representation for clustering narrow domain short texts. The self-term expansion methodology, which is independent of any external knowledge resource, greatly improves the results obtained by using classical document representation. This fact was confirmed in the practical task of word sense induction whose obtained results are shown in Chapter 6.

Finally, in Chapter 7, we study the impact of internal clustering validity measures by using narrow domain short-text corpora.

Chapter 2

Methods, techniques and datasets

In this chapter we define the clustering methods, term selection techniques and datasets which we used in our research experiments and that we will refer to throughout this Ph.D. thesis.

In the first section we describe the clustering methods and some important aspects such as the similarity measures and the external validity ones that may be employed. In Section 2.2, the term selection techniques used to reduce the vocabulary dimensionality are described. Finally, in Section 2.3 we illustrate the characteristics of the different data sets we used in our research work.

2.1 Clustering methods

Clustering analysis refers to the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait, often proximity, according to some defined distance measure [75, 85, 76]. Clustering methods are usually classified with respect to their underlying algorithmic approaches: hierarchical, iterative (or partitional) and density-based are some possible categories belonging to this taxonomy. In Figure 2.1 we can see the taxonomy presented in [82]. Hierarchical algorithms find successive clusters using previously established ones, whereas partitional algorithms determine all clusters at once. Hierarchical algorithms can be agglomerative ("bottom-up") or divisive ("top-down"); agglomerative algorithms

begin with each element as a separate cluster and merge the obtained clusters into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters. Iterative algorithms start with some initial clusters (their number either being unknown in advance or given a priori) and intend to successively improve the existing cluster set by changing their "representatives" ("centers of gravity" or "centroids"), like in K-Means [76] or by iterative node-exchanging (like in [66]). An interesting density-based algorithm is MajorClust [134], which automatically reveals the number of clusters, unknown in advance, and successively increases the total "strength" or "connectivity" of the cluster set by cumulative attraction of nodes between the clusters.

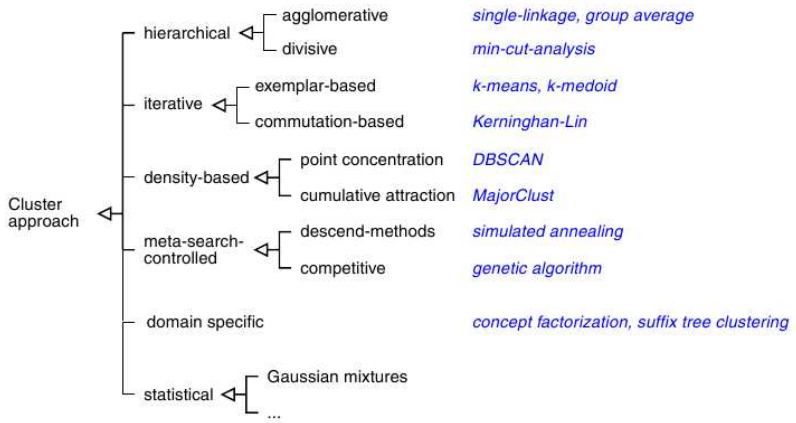


Figure 2.1: A taxonomy of clustering methods as presented in [82] (Reproduced with permission of the author).

In this thesis, we assume that the complete document clustering task may be carried out by executing at least the following three steps: (1) document representation; (2) calculus of a similarity matrix which represents the similarity degree among all the documents of the collection; (3) clustering of the documents. However, it is also feasible to apply an intermediate step called "dimensionality reduction", which may be performed by using some term selection technique.

Since we have tested different configurations for the experiments we carried out, we will briefly explain in this section how each one of the similarity measures and clustering methods works. The term selection techniques used in the experiments of

this Ph.D. thesis and their vocabulary dimensionality properties will be discussed in the next section of this chapter due to their optional use in the document clustering task. At the end of this section, a couple of external clustering validity measures are also explained.

2.1.1 Similarity measures

The clustering methods usually employ a similarity matrix which has already been calculated. 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 sub-sections we explain a set of similarity measures used in the experiments we carried out during our research work.

The Jaccard index

The Jaccard coefficient is a statistical measure used in natural language processing for comparing the similarity of a couple of documents [79]. It is defined as the cardinality of the intersection set divided by the cardinality of the union set of the sample texts. Given two documents, d_i and d_j, the Jaccard coefficient is a useful measure of the overlap that d_i and d_j share with their words. Formally,

$$Jaccard(d_i, d_j) = \frac{|d_i \bigcap d_j|}{|d_i \bigcup d_j|} \tag{2.1}$$

This measure does not care about term frequencies, which may be a considerable drawback in the most of the document clustering tasks. We will see in the following chapters that when using corpora with relative low frequencies, it will be better to use the Jaccard index instead of those measures that take into account the term frequencies, since the former is very easy and fast to be calculated compared with the others, and the similarity values obtained are very similar.

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 im-

portant a word is to a document in a corpus, using a vectorial representation. 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 [124] 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 D and a document d_j ($d_j \in D$), the *tf-idf* value for a term t_i in d_j is obtained by the product between the normalized frequency of the term t_i in the document d_j (tf_{ij}) and the inverse document frequency of the term in the corpus ($idf(t_i)$) as follows:

$$tf_{ij} = \frac{tf(t_i, d_j)}{\sum_{k=1}^{|d_j|} tf(t_k, d_j)} \tag{2.2}$$

$$idf(t_i) = log\left(\frac{|D|}{|d : t_i \in d, d \in D|}\right) \tag{2.3}$$

$$tf\text{-}idf = tf_{ij} * idf(t_i) \tag{2.4}$$

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_i and d_j, it is possible to calculate the cosine of the angle between these two vectors as follows:

$$\text{Cos}_\theta(\overrightarrow{d_i}, \overrightarrow{d_j}) = \frac{\overrightarrow{d_i} \cdot \overrightarrow{d_j}}{\left\|\overrightarrow{d_i}\right\| \left\|\overrightarrow{d_j}\right\|}$$

The Kullback-Leibler Distance

In 1951 Kullback and Leiber introduced a statistical measure of information with the purpose of calculating the asymmetric distance between two probability distributions associated with the same experiment [68]. 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 2.5.

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) log \frac{P(x)}{Q(x)} \tag{2.5}$$

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. In our research work we have used different symmetric Kullback-Leibler divergences (KLD). Each KLD corresponds to the definition of Kullback and Leibler [68], Bigi [16], Jensen [43], and Bennet [13] [157], respectively.

$$D_{KLD1}(P||Q) = D_{KL}(P||Q) + D_{KL}(Q||P) \tag{2.6}$$

$$D_{KLD2}(P||Q) = \sum_{x \in X} (P(x) - Q(x)) log \frac{P(x)}{Q(x)} \tag{2.7}$$

$$D_{KLD3}(P||Q) = \frac{1}{2} \left[D_{KL} \left(P|| \frac{P+Q}{2} \right) + D_{KL} \left(Q|| \frac{P+Q}{2} \right) \right] \tag{2.8}$$

$$D_{KLD4}(P||Q) = max \left(D_{KL}(P||Q) + D_{KL}(Q||P) \right) \tag{2.9}$$

KL and KLD have been used in many natural language applications like query expansion [30], language models [18], and categorization [16]. They have also been used, for instance, in speech processing applications based on statistical language modeling [33], and in information retrieval, for topic identification [17].

In order to use the aforementioned KLDs as similarity measures, we have considered to calculate the similarity between two documents d_i and d_j in an inverse function with respect to the distance defined in Equations (2.6), (2.7), (2.8), or (2.9).

In the text clustering model presented in this research work, the document d_j was represented by a term vector of probabilities $\overrightarrow{d_j}$ and the distance measure was, therefore, the KLD (the symmetric Kullbach-Leibler Divergence) between $\overrightarrow{d_i}$ and $\overrightarrow{d_j}$. However, in practice, often not all the terms in the vocabulary of d_j ($V(d_j)$) appear in the document d_i, therefore, for those terms it is useful to introduce a back-off probability 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 [32]).

In the smoothing model based on back-off, the 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 resulting definition of the smoothed document probability $\hat{P}(t_k|d_j)$ is shown in Eq. (2.10).

$$\hat{P}(t_k, d_j) = \begin{cases} \beta * P(t_k|d_j), & \text{if } t_k \text{ occurs in the document } d_j \\ \varepsilon, & \text{otherwise} \end{cases} \tag{2.10}$$

with

$$P(t_k|d_j) = \frac{tf(t_k, d_j)}{\sum_{x \in d_j} tf(t_k, d_j)} \tag{2.11}$$

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

Equation (2.10) must respect the following property:

$$\sum_{t_k \in d_j} \beta * P(t_k|d_j) + \sum_{t_k \notin d_j} \varepsilon = 1 \tag{2.12}$$

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

$$\beta = 1 - \sum_{t_k \notin d_j} \varepsilon \tag{2.13}$$

2.1.2 Hierarchical clustering methods

The Single Link Clustering method

Given a set D of documents to be clustered ($|D| = N$), and a $N \times N$ distance (or similarity) matrix, the following is the hierarchical clustering process presented in [62]:

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 [62].

The Lance & Williams recurrence

There exists 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 [69]. By means of the Lance and Williams recurrence an infinite number of hierarchical clustering methods may be implemented by using only one generic and simple computer program with quadratic spatial and cubic temporal costs.

Formally, let $\mathcal{M}$ 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 Lance and Williams recurrence shown in Eq. (2.14).

$$\mathcal{M}_{ij,k} = \alpha_i \mathcal{M}_{i,k} + \alpha_j \mathcal{M}_{j,k} + \beta \mathcal{M}_{i,j} + \gamma |\mathcal{M}_{i,k} - \mathcal{M}_{j,k}| \tag{2.14}$$

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. The value of α_c refers to both, α_i and α_j, whereas n_i, n_j, and n_k represent the number of documents in cluster i, j, and k, respectively.

Table 2.1: Six hierarchical clustering methods

Method	α_c	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
Weighted-average link	0.5	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 in order to find the maximum likelihood estimate of parameters in a probabilistic model, where the model depends on unobserved latent variables [94]. 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 implementation used in the experiments is the one of the Weka package [148].

2.1.3 Iterative clustering methods

The K-Star clustering method

The K-Star clustering method [130] 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 value in the similarity matrix $\varphi(d_i, d_j)$, and constructs a cluster (C_i) made up of 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 $\varphi(d_k, d_i) > \tau$, 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 τ defined as the average of the values in the similarity matrix.

The NN1 clustering method

The NN1 clustering algorithm [61] 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 know as K-NN [41], 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 ([76]):

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.

The DK-Means clustering method

We have proposed a deterministic version of the K-Means algorithm, which we have named DK-Means. The K-Means clustering method is executed after determining both, the number of expected clusters and a preliminar assignment of items to this K initial clusters. The number K and the preliminar assignment are obtained through the execution of the K-Star clustering method. Thus, we ensure the results do not vary over the execution of the same dataset. The global maximum, however, is not guaranteed.

2.1.4 Density-based clustering methods

The MajorClust clustering method

MajorClust executes iterative propagation of nodes into clusters according to the "maximum attraction wins" principle [134]. The algorithm starts by assigning each object to 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. Its characteristic of automatically discovering the target number of clusters makes it very attractive [133, 6, 105, 91], however a possible inconvenience may be considered when using it in the clustering task. It is assumed (but not proved) that the method is NP-Complete and, therefore, there not exist polynomial algorithms to

solve it efficiently and, therefore, the MajorClust would need considerable amount of time when calculating the clustering structure in large corpora.

The MajorClust algorithm

Input: object set D, similarity measure $\varphi : D \times D \rightarrow [0; 1]$, similarity threshold τ.
Output: function $\delta : D \rightarrow \mathbb{N}$, which assigns a cluster label to each item.

1. $i := 0$, ready := false
2. for all $p \in D$ do $i := i + 1$, $\delta(p) := i$ enddo
3. while ready = false do
   (a) ready := true
   (b) for all $q \in D$ do
      i. $\delta^* := \arg\max_j \{\Sigma_{\forall p} \varphi(p, q) | \varphi(p, q) \geq \tau, \delta(p) = j\}$[1].
      ii. if $\delta(q) \neq \delta^*$ then $\delta(q) := \delta^*$, ready := false
   (c) enddo
4. enddo

MajorClust automatically reveals the number of clusters and assigns each target document to exactly one cluster. However, in many real situations, there not exists an exact boundary between different clusters. A first attempt of fuzzifying this clustering method was introduced in [73]. The proposed clustering method assigns documents to more than one category by taking into account a membership function for both, edges and nodes of the corresponding underlying graph. Thus, the clustering problem is formulated in terms of weighted fuzzy graphs. The fuzzy approach permits to decrease some negative effects which appear in clustering of multi-categorized corpora.

[1]The similarity thershold τ is not a problem-specific parameter but a constant that serves for noise filtering purposes. Its typical value is 0.3.

2.1.5 External clustering validity measures

The quality of clustering results is often referred as "validity of document clustering" [82]. 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 taken into account: *internal* and *external* (often also called *objective* and *subjective*). The former validity indices allow to decide whether or not the obtained clusters are well developed with respect to the structural properties of the target clustering corpora. Whereas the latter indices compare the obtained clusters with respect to the gold standard, i.e., a classification given by an expert. In the following sub-sections, we present two external clustering measures we used in the research work carried out in this Ph.D. thesis.

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 $\mathcal{C} = \{C_1, \ldots, C_{|\mathcal{C}|}\}$ and a set of classes $\mathcal{C}^* = \{C_1^*, \ldots, C_{|\mathcal{C}^*|}^*\}$, the F-Measure between a cluster C_i and a class C_j^* is given by the following formula.

$$F(C_i, C_j^*) = \frac{2 \cdot Precision(C_i, C_j^*) \cdot Recall(C_i, C_j^*)}{Precision(C_i, C_j^*) + Recall(C_i, C_j^*)}, \tag{2.15}$$

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 class C_j^* are defined as follows:

$$Precision(C_i, C_j^*) = \frac{\text{Number of texts from cluster } i \text{ in class } j}{\text{Number of texts from cluster } i} = \frac{|C_i \bigcap C_j^*|}{|C_i|}, \tag{2.16}$$

and

$$Recall(C_i, C_j^*) = \frac{\text{Number of texts from cluster } i \text{ in class } j}{\text{Number of texts in class } j} = \frac{|C_i \bigcap C_j^*|}{|C_j^*|} \tag{2.17}$$

The global performance of a clustering method is calculated by using the values of $F(C_i, C_j^*)$, the cardinality of the set of clusters obtained, and normalizing it by

the total number of documents $|D|$ in the collection. The obtained measure is named F-Measure and it is shown in Equation (2.18).

$$F = \sum_{1 \leq i \leq |\mathcal{C}|} \frac{|C_i|}{|D|} \max_{1 \leq j \leq |\mathcal{C}^*|} F(C_i, C_j^*). \tag{2.18}$$

The supervised evaluation measure

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

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

Thereafter, the $\mathcal{M}$ mapping matrix is also used 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.

Other external clustering evaluation measures

The F-Measure is widely used to calculate the degree of similarity between the gold standard and the partition obtained by a given clustering algorithm, such as was expressed in Section 2.1.5. However, other external clustering validity measures also exist in literature [139, 8, 45, 46, 155]. For reference, we added some of them in Appendix A.

2.2 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 [42] [84]. 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 may led to use unsupervised term selection techniques instead of supervised ones. In the TST framework, a very insteresting research work has been carried out from several decades by using testors [70]. 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 [125] it was presented the fastest algorithm, which is not polinomial in complexity. Some works such as the one presented by Pons-Porrata *et al.* [108] 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 [74], whereas the third TST has been especially used in text categorization [149] [96]. The TP technique is a simple calculation procedure which has been used in many areas of computational linguistic: categorization or clustering of texts, keyphrases extraction, summarization, and weighting models for information retrieval systems (see [101, 100, 119, 118]). 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.

In order not to depend on hand-tagged corpora which could be so expensive in time to be created, in this research work we preferred to avoid the use of external

resources and only unsupervised term selection techniques were employed.

2.2.1 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 [156] and also on the refined studies of Booth [19] and Urbizagástegui [138]. 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 (2.19).

$$TP(d) = \frac{\sqrt{8 * I_1 + 1} - 1}{2} \tag{2.19}$$

where I_1 represents the number of words with frequency equal to 1 in a given text d [90] [138]. Alternatively, $TP(d)$ may be localised 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 [19], and it is useful when dealing with short texts which usually have terms with very low frequency.

Let t_i be the i-th term of the document d and $tf(t_i, d)$ the term frequency of that term. Let us consider the frequency-sorted vocabulary of d, i.e., $V'(d) = [(t_1, tf(t_1, d)), ..., (t_n, tf(t_n, d))]$ such that $tf(t_i, d) \geq tf(t_{i+1}, d)$. Therefore, $TP(d) = tf(t_i, d)$, with i equal to the minimum index such that $tf(t_{i+1}, d) = tf(t_{i+2}, d)$. The most important words are those which obtain the closest frequency values to $TP(d)$, i.e.,

$$V_{TP} = \{t_i | (t_i, tf(t_i, d)) \in V'(d), U_1 \leq tf(t_i, d) \leq U_2\} \tag{2.20}$$

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

For the representation schema, we considered that the important terms are those whose frequencies are closer to $TP(d)$. Therefore, a term with frequency very "close" to $TP(d)$ will get a high weight, and those "far" to it will get a weight close to zero. Therefore, for our experiments, the weight of each term t is calculated as in [118], i.e., inversely proportional with respect to the distance between its frequency and the TP frequency of d ($TP(d)$). The following equation shows how to obtain this value:

$$IDTP(t,d) = \frac{1}{|TP(d) - tf(t,d)| + 1} \tag{2.21}$$

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

2.2.2 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 [150]. 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 eliminating those rare 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.

2.2.3 The Term Strength technique

This technique was first introduced in [145] in order to improve the performance of document retrieval. Term Strength (TS) 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 a document d_i given that it has also appeared in the document d_j. The similarity between these

documents must be equal of higher than a given threshold τ. The weight given to each term t is then defined by the following equation:

$$TS(t) = P(t \in d_i | t \in d_j), \text{with } i \neq j,$$

τ 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 d_i and d_j to be more similar than τ. A more detailed description of the term strength technique may be found in [149] and [96].

2.3 Datasets

A corpus could be considered to be *narrow* or *wide* if the grade of the vocabulary overlapping is high or low, respectively. In a classification task, it is a high challenge to deal with narrow domain corpora, such as abstracts of scientific papers, technical reports, patents, etc. The obtained results are often very unstable or imprecise and the reason is that the document term frequencies are very low. Generally only 10% or 20% of the terms from the corpus vocabulary list occur in every document and their absolute frequency usually is one or two, and only sometimes three or four [6]. In this situation, changing a term frequency by one in the document representation may significantly deviate the classification results.

In the experiments we have carried out in our research work and that we will describe throughout the thesis, we investigated different features such as the corpus size and imbalance degree of different narrow and wide domain corpora.

The possibility of determining the *broadness* of a corpus (*narrow* vs. *wide*), its size and its imbalance degree allowed us to investigate how to tackle the challenging problem of clustering and validating narrow domain short-text corpora. The specific study of the aforementioned features was done and it is presented in Chapter 4, where we evaluate corpora in order to determine the level of the following four different characteristics: *broadness*, *shortness*, *imbalance*, and *stylometry*.

In the following sub-sections we describe each corpus into detail. We have pre-processed all these collections by eliminating stop words and by applying the Porter

stemmer [110]. The corpus features given in each table were obtained after applying this pre-processing phase.

2.3.1 Narrow domain short-text corpora

The *CICLing-2002* corpus

This corpus is made of 48 documents from the *Computational Linguistics* domain, which corresponds to the *CICLing 2002* conference[2]. The collection was first used by Makagonov et al. [77] in their experiments on clustering narrow domain abstracts. Even if very small, we consider it as a needed reference corpus. Moreover, its size made possible to manually validate the obtained results of a domain (computational linguistics) we are familiar with.

The distribution and other features of this corpus is shown in Tables 2.2 and 2.3, respectively. The topics of this corpus are also shown in Table 2.2.

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

Category	**Topics**	**# 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

As may be seen, in Table 2.3 we show features for both, the *full* and *abstracts* version of this collection. The number of documents and categories are exactly the

[2] `http://www.cicling.org`

Table 2.3: 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 abstracts	48	48
Total number of terms	80,109	3,382
Vocabulary size (terms)	7,590	953
Term average per abstract	1,668.94	70.45

same, however, it will be useful to study the effect of the term frecuencies in a short-text vs full collection.

With respect to the domain broadness, in [6] it is calculated that the vocabulary overlapping for the documents from the most different second and forth groups was about 70%. This implies that the selected domain is rather narrow.

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* [89]. It is composed of 2,922 abstracts from the *Physics* domain originally stored in the data server of the CERN[3]. The *hep-ex* corpus was released to be used in the automatic text categorization of documents task presented in [88]. 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 2.4 and 2.5. As can be seen, this corpus is totally unbalanced, narrow domain and made of short texts. Therefore, the use of this corpus may be very challenging.

[3]Centre Européen pour la Recherche Nucléaire; http://www.cern.ch/

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

Table 2.5: General 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

The WSI-SemEval collection

This data collection was provided by the organisers of the "Evaluating Word Sense Induction and Discrimination Systems" task of the SemEval 2007 workshop of the Association for Computational Linguistics[4].

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 manually annotated with OntoNotes senses [4].

[4] `http://nlp.cs.swarthmore.edu/semeval/tasks/task02/description.shtml`

In Table 2.6 are shown general features of the WSI-SemEval data collection. In this particular case, we present the average and not the complete features for each corpus (ambiguous word), due to the high number of corpora.

Table 2.6: Feature averages of the *WSI-SemEval* data collection

Feature	**Value**
Size of the corpus (bytes)	10,644,648
Number of ambiguous words	100
Number of sentences	27,132
Minimum number of categories (senses)	1
Maximum number of categories (senses)	11
Average number of categories (senses)	2.87
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 name of the ambiguous words (verbs and nouns) along with the number of their corresponding instances are presented in Table 2.7. Nouns are shown in Table 2.7(a), whereas the verbs are displayed in Tables 2.7(b) and 2.7(c). We may see that some instances, such as "share" and "say" have each one similar number of instances as other corpora used in our research work. We consider that this data collection is very important in the experiments we have carried out because it is not easy to have available a high number of corpora with similar characteristics like for instance the ones we are studying in this Ph.D. thesis, namely *broadness*, *shortness*, *imbalance*, and *stylometry*.

Table 2.7: The ambiguous words of the *WSI-SemEval* corpus

Word	**Instances**
president	1,056
chance	106
authority	111
base	112
rate	1,154
carrier	132
defense	141
condition	166
source	187
network	207
effect	208
development	209
job	227
hour	235
drug	251
power	298
share	3,061
position	313
move	317
management	329
capital	335
area	363
policy	370
value	394
order	403
plant	411
exchange	424
future	496
bill	506
system	520
part	552
point	619
state	689
space	81
people	869

(a) Nouns

Word	**instances**
explain	103
announce	108
cause	120
kill	127
remember	134
hope	136
allow	143
hold	153
end	156
produce	159
begin	162
report	163
build	165
raise	181
receive	184
find	202
lead	204
buy	210
see	212
set	216
come	229
grant	24
need	251
start	252
believe	257
do	268
say	2,702
work	273
examine	29
go	305
fix	34
negotiate	34
keep	340
rush	35
feel	398

(b) verbs

Word	**instances**
turn	402
ask	406
complain	46
improve	47
propose	48
attempt	50
purchase	50
contribute	53
regard	54
express	57
complete	58
promise	58
replace	61
affect	64
recall	64
remove	64
approve	65
claim	69
disclose	69
occur	69
enjoy	70
avoid	71
maintain	71
prove	71
prepare	72
exist	74
care	76
describe	76
join	86
estimate	90

(c) verbs

2.3.2 A new narrow-domain short text corpus

The absence of a specific forum for the evaluation of clustering narrow domain short texts task may be one of the reasons of the lackness of a gold standard for corpora of this kind. 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 MEDLINE[5]. 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* [104]. It consists of 900 abstracts related with the "Cancer" domain. This corpus has been used in some experiments, such as the ones presented at the CICLing 2007 conference [99]. More recently, in [58], the *KnCr* corpus was used (together with the *CICLing-2002* and *hep-ex* corpora) to show the possible correlation between subjective and objective (i.e., external and internal) clustering validity measures.

Below we explain the guidelines followed in order to construct 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 manually constructed, we attempted 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 the gold standard. Therefore, in the experiment we discarded the document itself and we considered only the document keywords. We selected three clustering methods for this experiment: two already implemented in the Weka machine learning software [148] (Expectation Maximization and K-Means) and one implemented by ourself (K-Star). A description of the three clustering methods can be seen in Section 2.1.

We used the F-Measure (see Section 2.18) for comparing each pair of clustering methods. The value of 0.51 in the third column of Table 2.8 is the F-Measure obtained by using the clusters obtained by the EM clustering method as the "gold standard",

[5] `ftp://ftp.nlm.nih.gov/nlmdata/sample/medline/`

whereas the evaluated clusters are the ones suggested by the K-Means algorithm. In general, none combination of pairs of clustering methods obtained a reliable F-Measure to confirm a possible gold standard. The results of this experiment reinforce the hypothesis that clustering narrow domain corpora seems to be really a difficult task, even if we know the keywords of each document or abstract.

Table 2.8: Obtained results 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, in order to construct the gold standard, we manually classified every document in its correct class. For the construction of the gold standard categories we used the ontology made available by the National Cancer Institute (NCI)[6]. The current OWL version of this ontology describes a hierarchy of cancer terms based on the anatomy kind and specifies the fine grain categories of this domain [7]. Tables 2.9 and 2.10 show the complete characteristics of this new cancer corpus. As can be seen, only 900 from the original 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.

[6] `http://ncimeta.nci.nih.gov/`
[7] `http://www.mindswap.org/2003/CancerOntology/`

Table 2.9: Categories of the *KnCr* corpus

Category	**# of abstracts**
blood	64
bone	8
brain	14
breast	119
colon	51
genetic studies	66
genitals	160
liver	29
lung	99
lymphoma	30
renal	6
skin	31
stomach	12
therapy	169
thyroid	20
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 short texts of this narrow-domain corpus. We implemented two hierarchical clustering methods (Single and Complete Link Clustering) and three iterative 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 abstracts instead of keywords, and by using two well-known vocabulary reduction techniques (Document Frequency and Term Strength), are presented in Table 2.11. We may observe a low F-Measure value for each clustering method, which highlights again the hardness of the task of the gold standard generation.

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

Clustering method	**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 or not the clustering of keywords (provided by the abstract authors) behaves better than when a vocabulary reduction technique is used, we carried out a third experiment. In this case we compared the results obtained by clustering keywords with EM, K-Means and K-Star methods with the gold standard we manually built. The results are presented in Table 2.12. We may see that the use of keywords instead of abstracts can lead to more confusion in the clustering narrow-domain short texts task. This may be due to the different keywords that could be added by the authors of scientific texts of the same topic That is, a little variation in the keyword set leads to classify similar documents in different classes.

We have made free available this new corpus by email request to the authors. However, a supervision by an expert of the cancer domain is needed to validate the quality of the obtained gold standard. We consider that this corpus, together with

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

Clustering method	***F*-Measure**
EM	0.20
K-Means	0.22
K-Star	0.22

its gold standard, will allow to test clustering algorithms on short texts of the cancer narrow domain.

2.3.3 Other kinds of corpora

In the research work we carried out, we have also used short texts corpora that were not narrow-domain. The goal was to study the characteristics of both, narrow and wide domain short-text corpora. We describe each corpus into detail in the following sub-sections.

Reuters

Reuters-21578[8] 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 and R52 sub-collections of Reuters-21578 since they are a single-label categorized dataset.

The characteristics of the R8 corpus are given in Tables 2.13 and 2.14, whereas the properties of the *training* and *test* version of the R52 are given in Tables 2.15 and 2.16.

Since both, the R8 and R52 corpora are used for the categorization task, it is usual to work with a training and a test version of the data. Therefore, each table shows the *training* and *test* subset features of these collections.

The construction of the aforementioned corpora was done in the following manner. We considered from the original Reuter-21578 collection only those documents with a

[8] `http://www.daviddlewis.com/resources/testcollections/reuters21578/`

Table 2.13: Number of documents in each category of the *R8-Reuters* corpus

Category	**Training**	**Test**	**Total**
trade	319	102	421
grain	78	34	112
monex-fx	366	130	496
crude	314	140	454
interest	202	87	289
acq	1,608	707	2,315
ship	121	43	164
earn	2,831	1,076	3,907

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

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

single topic and the classes which still have at least one training and one test example. In this manner, we obtained 8 of the ten most frequent classes and 52 of the original ninety. Following Sebastiani's convention [127], we called these sets R8 and R52.

From the viewpoint of the narrow-domain corpora, it is remarkable that from the ten most frequent classes of Reuters-21578 (R10), the classes corn and wheat (which are intimately related to the class grain) disapeared. Moreover, the wheat class lost many of its documents when we built the R8 dataset. This is important since in the experiments of this research work we are considering the R8 collection as a non-narrow domain one and, therefore, the obtained dataset is more skewed than the R10 one.

Table 2.15: Number of documents in each category of the *R52-Reuters* corpus

Category	**Train**	**Test**	**Total**
acq	1,596	696	2,292
alum	31	19	50
bop	22	9	31
carcass	6	5	11
cocoa	46	15	61
coffee	90	22	112
copper	31	13	44
cotton	15	9	24
cpi	54	17	71
cpu	3	1	4
crude	253	121	374
dlr	3	3	6
earn	2,840	1,083	3,923
fuel	4	7	11
gas	10	8	18
gnp	58	15	73
gold	70	20	90
grain	41	10	51
heat	6	4	10
housing	15	2	17
income	7	4	11
instal-debt	5	1	6
interest	190	81	271
ipi	33	11	44
iron-steel	26	12	38
jet	2	1	3

Category	**Train**	**Test**	**Total**
jobs	37	12	49
lead	4	4	8
lei	11	3	14
livestock	13	5	18
lumber	7	4	11
meal-feed	6	1	7
money-fx	206	87	293
money-supply	123	28	151
nat-gas	24	12	36
nickel	3	1	4
orange	13	9	22
pet-chem	13	6	19
platinum	1	2	3
potato	2	3	5
reserves	37	12	49
retail	19	1	20
rubber	31	9	40
ship	108	36	144
strategic-metal	9	6	15
sugar	97	25	122
tea	2	3	5
tin	17	10	27
trade	251	75	326
veg-oil	19	11	30
wpi	14	9	23
zinc	8	5	13

Table 2.16: Other features of the *R52-Reuters* corpus

Feature	**Train**	**Test**
Size of the corpus (Bytes)	2,837,999	1,010,066
Number of categories	52	52
Number of documents	6,532	2,568
Total number of terms	459,344	165,112
Vocabulary size (terms)	16,145	9,730
Term average per document	70.32	64.3

20 Newsgroups

The 20 Newsgroups dataset[9] is a well-known collection which has been used for benchmarking clustering algorithms. The corpus is made of 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 some preliminary experiments, we have used a small version of 20 Newsgroups called *Mini20Newsgroups*. The characteristics of this corpus are given in Tables 2.17 and 2.18. We specially made use of this reduced version of the *20Newsgroups* collection with experiments, such as clustering validity, where the time was crucial to tune particular parameters of the employed techniques and methods.

The fact, due to the major clustering evaluation measures introduced later in Chapter 4 may be executed moderately fast, it made us to consider to employ also the full version of the *20Newsgroups* collection in our research work.

Moreover, the *20Newsgroups* corpus may be seen as a wide domain corpus, since the categories of the collection range from religion, computer, to sports and, therefore, it may be used to compare the evaluation measures over narrow vs wide domain corpora.

In Tables 2.19 and 2.20 we show the properties of the entire collection for its training and test version.

[9] `http://people.csail.mit.edu/jrennie/20Newsgroups/`

Table 2.17: Number of documents in each category 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 2.18: General 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

The 4 Universities Dataset (*WebKb*)

The *WebKb* dataset is made of webpages collected by the World Wide Knowledge Base (Web→Kb) project of the Carnegie Mellon University text learning group, and downloaded from "The 4 Universities Data Set Homepage"[10]. These pages were originally collected from computer science departments of various universities in 1997 (Cornell, Texas, Washington, Wisconsin, and others), and manually classified into seven different classes: student, faculty, staff, department, course, project, and other.

[10] `http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/`

Table 2.19: Number of documents in each category of the *20Newsgroups* corpus

Category	**Train**	**Test**	**Total**
alt_atheism	480	319	799
comp_graphics	584	389	973
comp_os_ms-windows_misc	572	394	966
comp_sys_ibm_pc_hardware	590	392	982
comp_sys_mac_hardware	578	385	963
comp_windows_x	593	392	985
misc_forsale	585	390	975
rec_autos	594	395	989
rec_motorcycles	598	398	996
rec_sport_baseball	597	397	994
rec_sport_hockey	600	399	999
sci_crypt	595	396	991
sci_electronics	591	393	984
sci_med	594	396	990
sci_space	593	394	987
soc_religion_christian	598	398	996
talk_politics_guns	545	364	909
talk_politics_mideast	564	376	940
talk_politics_misc	465	310	775
talk_religion_misc	377	251	628

However, the version over the one we carried out the experiments was pre-processed and pruned to be used in experiments of single-label text categorization [29]. The difference with respect to the original corpus is that the classes *Department* and *Staff* were discarded because there were only a few pages from each university. They also discarded the class *Other* due to its high heterogeneity. They constructed a training and a test split for this dataset, in order to be consistent with their previous text categorization datasets, by randomly choosing two thirds of the documents for training and the remaining third for testing. The number of documents for each split and category is shown in Table 2.21. Other characteristics are shown in Table 2.22.

Table 2.20: General features of the *20Newsgroups* corpus

Feature	**Train**	**Test**
Size of the corpus (Bytes)	9,785,329	6,337,651
Number of categories	20	20
Number of documents	11,293	7,528
Total number of terms	1,610,938	1,043,828
Vocabulary size (terms)	54,580	43,720
Term average per document	142.65	138.73

Table 2.21: Number of documents in each category of the *WebKb* corpus

Category	**Train**	**Test**	**Total**
project	335	166	501
course	620	306	926
faculty	745	371	1,116
student	1,083	537	1,620

Table 2.22: General features of the *WebKb* corpus

Feature	**Train**	**Test**
Size of the corpus (Bytes)	2,541,674	1,296,139
Number of categories	4	4
Number of documents	2,783	1,380
Total number of terms	371,989	187,990
Vocabulary size (terms)	7,287	4,798
Term average per document	133.67	136.23

Chapter 3

Clustering narrow domain short-text corpora

Nowadays, most digital libraries and other web-based repositories of scientific and technical information provide free access only to *abstracts* as opposed to full texts access of the documents. Some repositories, like the renowned MEDLINE[1], and the Conseil Européen pour la Recherche Nucléaire (CERN)[2], receive hundreds of publications every day that must be categorized into some specific domain, often with an unknown number of categories a priori. For instance, the PubMed[3] is an on-line search engine for MEDLINE articles which has an index of more than 16 millions of abstracts. This huge volume of information would be impossible to managed without the help of an automatic computational-based system, and, therefore, in order to deal with this real problem it is necessary to construct novel methods for classifying *narrow domain short texts*.

A number of approaches exist in order to tackle this particular problem. In [88] for instance, the author proposes a tool, called TECAT, with the aim of automatically categorizing texts from restricted domains. These experiments were tested in the high energy physics domain. The author made use of supervised classifiers trained

[1] `http://www.nlm.nih.gov`
[2] `http://library.cern.ch`
[3] `http://www.ncbi.nlm.nih.gov`

with data which were tagged with keywords extracted from the DESY thesaurus[4]. However, in scientific domains it is rare to find linguistic resources to help in supervised classification tasks due to the specific or narrow vocabulary of the documents. In this case, the clustering approach would be used instead. However, we must take into account that some difficulties that would arise. We know, for example, that sometimes the use of scientific document keywords (rarely provided by authors) may be insufficient to perform a good clustering. This could lead to obtaining a lower performance than when using the abstracts on the clustering process [104].

Clustering of scientific abstracts is a particularization of the categorization of narrow domain short-text corpora. The possible implications of the latter problem are therefore inherited by the former one. In this case, we are referring to the complexity of dealing with both, high vocabulary overlapping and low term frequencies. However, when these kind of texts also belong to scientific papers the difficulty increases due to the continued use of words such as, for instance: "in this paper we present...", "the aim is", "the results", etc.

In fact, in [6] 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 one or two, sometimes three or four. In this situation, changing a keywords frequency by one can significantly change the clustering results.

The previous assumptions highlight the high challenge implied when dealing with scientific abstract corpora. This work is motivated by the difficulty of the task as well as the potential applications of it, as mentioned in the following paragraphs.

Emerging areas of text writing such as the so called blogs (or weblogs) is another example of narrow domain short-text corpora different than scientific abstracts. Blogs

[4]The Deutsches Electron Synchrotron (DESY); `http://library.desy.de`

are reverse chronological sequences of highly opinionated and personal on-line commentaries. There exists a high interest on exploiting computational approaches in order to analyse blogs. The main industrial and scientific aim is to outreach opinion formation and monitoring the reaction of public with respect to specific events. This new interesting area of research was investigated, for instance, in [21] in order to analyse the effectiveness of tags for classifying blog entries by selecting the top 350 tags from a blog-based information retrieval (named Technorati [40]) and measuring the similarity of all articles sharing the same tag. They claimed that tags are useful for clustering blogs into *broad* categories, but less effective in indicating the *particular content* of an article. Moreover, the comparison in that paper between the obtained results and a randomly constructed set of document clusters showed that aside from the low obtained cosine similarity, the improvement is rather low (less than 0.5).

These facts are interesting because they reinforce our hypothesis that the clustering hardness of documents may be highly related to two important characteristics: the corpus *shortness* and the corpus domain *broadness*.

Many web pages are made of, or considered to be, short-texts. News, document titles, abstracts, FAQs, etc., are some examples of the high volume of short texts available in Internet. There exists sufficient interest from the computational linguistic community in analysing the behaviour of classifiers when using short-text corpora [153, 54, 154, 111, 21, 11, 97].

Moreover, we are witnesses to a new era of text communication where people are using, and most likely will continue to use, "small-language". Blogs, snippets, emails, chats, etc., are some examples of this particular mode of communication. This type of communication tends to be personal and often uses a restricted vocabulary (i.e., narrow domain).

The above mentioned two corpus features represent the corpus average document size (*shortness*) and whether or not the vocabulary is very domain-dependent (*domain broadness*). After investigating the studies in literature on categorization and clustering of short texts (see for instance [21]), we might assume that there exist different levels of difficulty when clustering documents. We believe that corpus hardness degree depends on, at least, the aforementioned characteristics. Moreover, we also

consider that these features are independent of the classifier used. However, further study should analyse this issue in more detail.

Given the following kinds of corpora: Short-Text (ST), Narrow Domain (ND), Narrow Domain Short-Text (NDST) and Scientific Abstract (SA), we might hypothesize that the relation shown in Eq. (3.1) holds. We represent a hierarchical relationship that these kinds of corpora have in terms of the difficulty of clustering.

$$SA \subseteq NDST \subseteq ST \subseteq ND \textbf{ or } SA \subseteq NDST \subseteq ND \subseteq ST \tag{3.1}$$

In Figure 3.1 we present a simple taxonomy which reflects the above considered hypothesis of corpus/document hardness levels.

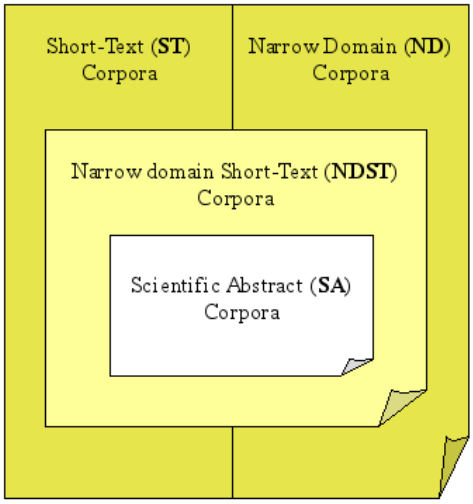


Figure 3.1: Hypothesis of document hardness levels

The main purpose of this chapter is to study the behaviour of clustering methods, and also of different term selection techniques, when dealing with narrow domain short-text corpora. We consider that the analysis which is the objective of this chapter besides being quite a challenging one, should be highly beneficial for the aforementioned emerging areas of application.

In the following section we discuss the differences between clustering and categorization, their advantages and application areas. In Section 3.2 we present a formal definition of clustering. Section 3.3 presents previous work of other authors in the

clustering narrow domain short-text corpora task. The experimental results we have obtained are shown and discussed in Section 3.4. Finally, concluding remarks are given in Section 3.5.

3.1 Clustering vs. categorization

Document classification subsumes two types of text analyses: clustering and categorization. The difference between the two is that the latter uses a predefined number of classes or categories with their corresponding tags, whereas in the former approach, the number and the tag for each category is to be discovered. Since in categorization the classes are known a priori, categorization algorithms usually take advantage of them by using supervised algorithms with some kind of training step. Clustering algorithms, on the other hand, have been linked with unsupervised classification; however, they may also use supervised techniques in order to discover the expected clusters.

In the literature, when comparing the same task, it can be seen that categorization algorithms consistently outperform clustering ones. This is to be expected, since the categorization uses training data in order to feed a supervised classifier and, thereafter, the obtained model is used in order to compute the classification task using test data. This advantage cannot be sustained when dealing with certain domains that do not have training data because either, constructing such data is not permissive in terms of time, or the domain is narrow and the taxonomy is not clearly defined. Categorization of scientific texts is an example of the last case.

Another advantage of categorization algorithms is that, once we have trained the model (which usually takes a lot of time), the time needed for evaluating a test set is quite fast. However, we are restricted to categorize the input data only with some of the categories used in the training phase. This drawback can be successfully overcome with clustering algorithms by discovering not only the expected categories but also new ones, which can be highly beneficial.

In summary, when training data is available and it is required to restrict the assignment of categories to a fixed classification taxonomy, it is preferable to use

categorization instead of clustering algorithms. However, when dealing with specific domains with dynamic categories, clustering algorithms are the best choice.

The potential application areas of document classification include, but are not restricted to, analysis of document databases such as internet blogs, patent documents, scientific papers, automatic forwarding messages at help desks, e-mail filtering, enhancement of internet search engines by means of cluster-based information retrieval, etc. It is remarkable that many of these tasks require either, to discover new categories in the classification process or to work with narrow domain corpora and, therefore, the use of clustering is justified.

3.2 The clustering hypothesis

The document clustering task may be informally expressed as the partitioning of a document collection into subsets (clusters), so that the documents in each subset (ideally) share some common trait, often proximity, according to some defined distance measure.

In the information retrieval framework, there also exists a clustering hypothesis which was formulated as follows: "closely associated documents tend to be relevant to the same requests" [117]. There we may see the request as the expected categories or classes of each desired cluster.

The clustering hypothesis may only be verified through experimental work on a large number of collections. However, we can depict a mathematical formulation for it as follows.

Given a document collection $D = \{d_1, d_2, \cdots, d_{|D|}\}$, a clustering of D is a partition into k subsets $\mathcal{C} = \{C_1, C_2, \cdots, C_k | C_i \subseteq D\}$, such that $\bigcup_{i=1}^{k} C_i = D$. If $C_i \bigcap C_j = \emptyset$, for $i \neq j$, then we are dealing with "hard" clustering, otherwise it is named "soft" clustering or "fuzzy" clustering.[5]

The final aim of document clustering is to discover the optimal partition of D, and, therefore, given a similarity function of the document set, $\varphi : D \times D \rightarrow \mathbb{R}^+$, the

[5]In this research work, if not specifically expressed, we will always deal with "hard" clustering.

best cluster set $\mathcal{C}$ is considered of good quality if it is found by maximizing the intra-clusters similarity and, at the same time, minimizing the inter-clusters similarity. This may be defined by two optimization formulae, as shown in Eq. (3.2) and (3.3).

$$\text{Maximize} \quad \left[\sum_{i=1}^{k} \sum_{\substack{\forall d_r, d_s \in C_i \\ 1 \leq r < s \leq |C_i|}} \varphi(d_r, d_s) \right] \tag{3.2}$$

$$\text{Minimize} \quad \left[\sum_{i=1}^{k} \sum_{\substack{\forall d_r \in C_i, \forall d_s \in C_l \\ 1 \leq i \neq l \leq k}} \varphi(d_r, d_s) \right] \tag{3.3}$$

In practice, the aim is to obtain the optimal partition with respect to a given gold standard $\mathcal{C}^*$ constructed by using human criteria. This may be a drawback if we consider that human beings may have different criteria when classiying documents. Whether it exists a huge problem or not when dealing with lexical versus semantic clustering is matter of a deeper study which is beyond of this investigation.

In the experiments carried out in this research work, the F-Measure is used to calculate the degree of similarity between the gold standard and the partition obtained by a given clustering algorithm, such as was expressed in Section 2.1.5. However, other external clustering validity measures, such as Huberts Γ statistic, purity, inverse purity, B^3, etc also exist in literature [139, 8, 45, 46, 155]. For reference, we have added some of these validity measures in Appendix A

3.3 Related work

Classification is a very old area that has been studied from different perspectives (see [86, 37] for an introduction to machine learning). On the one hand, the particular case of short text classification has been investigated in the past in various research projects from both, a categorization and a clustering perspective. In [153], for instance, it is described a method for improving the classification of short text strings by using a combination of labeled training data, plus a secondary corpus of unlabeled but related longer documents. It is shown that such unlabeled background

knowledge can greatly decrease error rates, particularly if the number of examples or the size of the strings in the training set is small. They performed their experiments with three different corpora: technical papers, sport and banking news headers and, webpage titles. This approach was later enhanced in other works such as the one that exploits transductive Latent Semantic Indexing (LSI) in short text classification problems [154]. LSI is a technique that allows to find a low-rank approximation to a original term-document matrix (describing the occurrences of terms in documents). The goal is to analyse the relationships between the documents of the collection and the terms they contain. In this way, an approximated matrix may be used instead of the original term-document matrix which may be considered noisy and/or too large for computing resources. The approximated matrix is calculated by applying Singular Value Decomposition (SVD). SVD consists of decomposing the original matrix into three different matrices: two orthonormal ones and a diagonal matrix. The values of the diagonal matrix are called the singular values.

In [111] it is presented another approach that has used LSI. However, in this case LSI was applied as a data preprocessing method before applying a text classifier based on Independent Component Analysis (ICA) in order to ameliorate the problems related with particular characteristics of short texts which usually have little overlap in their feature terms. Unfortunately, in cases like this one, techniques such as ICA were shown not to work well. When using ICA in text categorization, the goal is to calculate the independent components of these documents and, thereafter, to use them to represent the documents in the categorization process. In many cases good categorization effects are obtained. From the experiments on Chinese short texts, the authors concluded that the sequence LSI→ICA, provide better categorization effects. However, we consider that the two known drawbacks of LSI (semantic interpretation of the obtained dimensions and the assumed ergodic hypothesis of words in documents) may be carefully taken into account when combining both, LSI and ICA.

In [54], the categorization of short documents is considered; the authors developed a method for automating document categorization in a digital library. Their method is based on "itemsets", extending the traditional application of the *apriori algorithm* (introduced in [5]) which they claimed to be suitable for automatic categorization of

short documents, such as abstracts and summaries. The authors follow the notation of Agrawal [5] but in this case they consider the items to be terms and the itemsets to be basket of terms. In the training phase, the itemsets are calculated automatically over a manually constructed training corpus of classes and documents assigned to their corresponding class. In the categorization phase, the itemsets for the test corpus are calculated again automatically and, thereafter, each document is assigned to a class based on the sum of products of its itemset weights and the ones obtained in the training phase. One of the drawbacks of the apriori algorithm is that it needs a threshold in order to discriminate the good and not so good itemsets. However, according to the authors the method computes quickly and, therefore, it could be used in commercial applications.

The main concern when dealing with short texts is the huge volume of information that classification algorithms must handle. Let us consider short documents such as paper abstracts and emails, quite commonly available in Internet. The major clustering algorithms become very inefficient when have to deal with very large amount of data with very high-dimensional representation. One recent approach presented in [142] proposes a frequent term-based parallel clustering algorithm to be used in very large short-text database. This "itemset"-based procedure is also similar to the one presented in [54]. However, the same authors claimed in [141] that the performance is better than the *apriori algorithm* presented in [5]. The aim in [142] was to improve accuracy of clustering and obtain good scalability when processing huge amounts of data. In fact, in [140] a similar procedure is executed by calculating the top-k frequent term sets to produce k initial means that are used as the initial clusters which are further refined by the k-means clustering method.

One way of improving the categorization of collections made up of short documents is by enriching the document representation by means of external knowledge resources. For instance, in [97] it is proposed a clustering algorithm based on concept similarity. In this case, Chinese terms are splitted into concepts by using a lexicon known as HowNet [114] which is an on-line common-sense knowledge database unveiling inter-conceptual relations and inter-attribute relations of concepts as connoting in lexicons of the Chinese and their English equivalents. In [11], a method for improving the

accuracy of clustering short texts by enriching their representation with additional features from Wikipedia[6] is proposed. The fact that small length texts have low frequency terms is the main detonating that has led researchers to investigate different techniques of improving short text classification.

The main problem, from our particular viewpoint, is that all these proposals are supervised, since they require external resources constructed in advance by human beings (hand-labeled) and, therefore, training data is required. Moreover, these knowledge databases are usually domain dependent which is a considerable drawback when dealing with narrow domain collections whose terms are not usually covered by generic lexicons. In fact, research work dealing with narrow domain clustering are even more difficult to find in the literature. In [26], for instance, the tasks of categorization and clustering of narrow domain documents are investigated. The experiments for the categorization task were carried out by using Bernoulli mixtures for binary data, and in the case of the clustering task, by means of Steins MajorClust method (see Section 2.1.4 of Chapter 2). The proposed method for clustering narrow domain short texts extracts sense clusters from abstracts, exploiting the WordNet [39] 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. Experiments with clustering narrow domain texts are also presented in [129], where the complexity of this particular task is highlighted.

As already stated, the aim of this chapter is to study the behaviour of clustering methods and different term selection techniques when dealing with corpora made of documents that are both: narrow domain and short. As far as we know, research work in this field has rarely been carried out, since the categorization of narrow domain short-text corpora is a relatively new challenge, requiring further investigation due to the current "fashionable" use of small-language as particular mode of communication. This fact could also be derived from the great challenge that this problem implies,

[6] `http://www.wikipedia.org`

since the results obtained are very unstable and even imprecise when clustering abstracts of scientific papers, technical reports, patents, etc. Some related work was presented in [77], where simple procedures to improve results by an adequate selection of keywords and a better evaluation of document similarity was proposed. The authors used as corpora two collections retrieved from the Web. The first collection was composed by a set of 48 abstracts (40 Kb) from the CICLing 2002 conference (describe in the previous chapter); the second collection was composed by 200 abstracts (215 Kb) from the IFCS-2000[7] 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. 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 [6] 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 [78] in order to perform the word selection process. They cluster the stemmed terms by using the same document vector space representation. Finally, they smooth the frequency of the each final obtained term index t_k in the document collection D with the following formula: $\log(1 + tf(t_k, D))$. The purpose of the latter formula was to ameliorate the effects of dealing with the low frequencies of the terms. The authors based their experiments on the first CICLing collection used by Makagonov et al. [77], and they succeeded in improving those results. In the final discussion, Alexandrov et al. stated that

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

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 [77], suggesting that, for an open access via Internet, digital libraries should provide document images of full texts for the papers and not only abstracts.

In the following section we describe the several experiments we have carried out on narrow domain short-text corpora.

3.4 Experimental results

In this section we study the performance of different clustering methods and term selection techniques in the clustering of narrow domain short texts task. The aim is to investigate possible strategies that would be used in order to tackle the problem of both, a) the low frequencies of vocabulary terms in short texts, and b) the high vocabulary overlapping associated to narrow domains.

For this purpose, we have structured this section into four parts. First, we analyse the impact of using term selection techniques when clustering scientific abstracts (experiment 1). Next, we study the behaviour of different clustering methods in order to determine those more suitable for their use in further experiments (experiment 2). Thereafter, we investigate the application of a new document similarity measure based on the Kullback-Leibler divergence in the clustering of documents (experiment 3). Finally, we show the plotting of the F-Measure as a function of different ranges of vocabulary reduced data of a standard narrow domain short-text corpus (experiment 4). The vocabulary reduction is calculated with three different term selection techniques and the F-Measure is obtained by comparing the gold standard of this standard corpus with the obtained clusters by the execution of a selected clustering method.

3.4.1 Experiment 1: The role of the term selection process

In our first works presented in [61] and in [60], we used a novel technique for term selection based on mid-frequency terms, named Transition Point. The experiments carried out were tested with the *CICLing-2002* corpus. The obtained findings motived us to use also this term selection technique in the evaluation of a bigger size corpus [100] and to compare the results with other two term selection techniques used in literature: Document Frequency and Term Strength.

This first experiment was carried out using the reference narrow-domain abstracts corpus of *hep-ex* [89]. The K-Star clustering method was employed because it may automatically discover the number of clusters in a totally unsupervised way reducing, therefore, the number of variables to analyse. The similarity among the documents was calculated by means of the Jaccard similarity function due to the fact that Jaccard is faster to be computed than the cosine measure and also obtains comparable results when dealing with short text corpora. Our main concern was to evaluate the behaviour of the three unsupervised term selection techniques described earlier in Section 2.2 in the task of clustering abstracts of a narrow domain.

Exp. 1.1: A test over a subset of the *hep-ex* corpus

In order to have a first idea of the behaviour of each one of the term selection techniques we used in these experiments, we carried out a test over a subset of the *hep-ex* corpus. This subset was 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 for the K-Star clustering method was tuned on this collection, concluding that it should be the similarity average among all the documents.

We calculated the F-Measure values for every term selection technique executed over different percentages of the collection vocabulary (from 600 to 2,000 terms).

Given a percentage of the collection vocabulary, the highest scored terms, by using the TP, DF and TS techniques were selected. Therefore, the comparison among the TSTs was done through the complete range of vocabulary size. The DF and

TS techniques selected a range from 2% to 70% of vocabulary terms. This range corresponds to between 21 to 1,700 of the total terms of the collection. Given a similar range of total terms, the TP selection technique took from 5 to 30 terms from each text. In Fig. 3.2, the results of these three techniques are shown; the horizontal axis represents the number of terms and the vertical axis the F-Measure. In order to apply the TS technique, a similarity matrix was calculated as 3-tuples (t_i,t_j,$\varphi(t_i, t_j)$) and it was sorted according to $\varphi(t_i, t_j)$, and then $TS(t)$ was computed for all terms. This first calculation produced only 1,349 terms and, therefore, in order to keep all of them, the threshold β was fixed to 0.

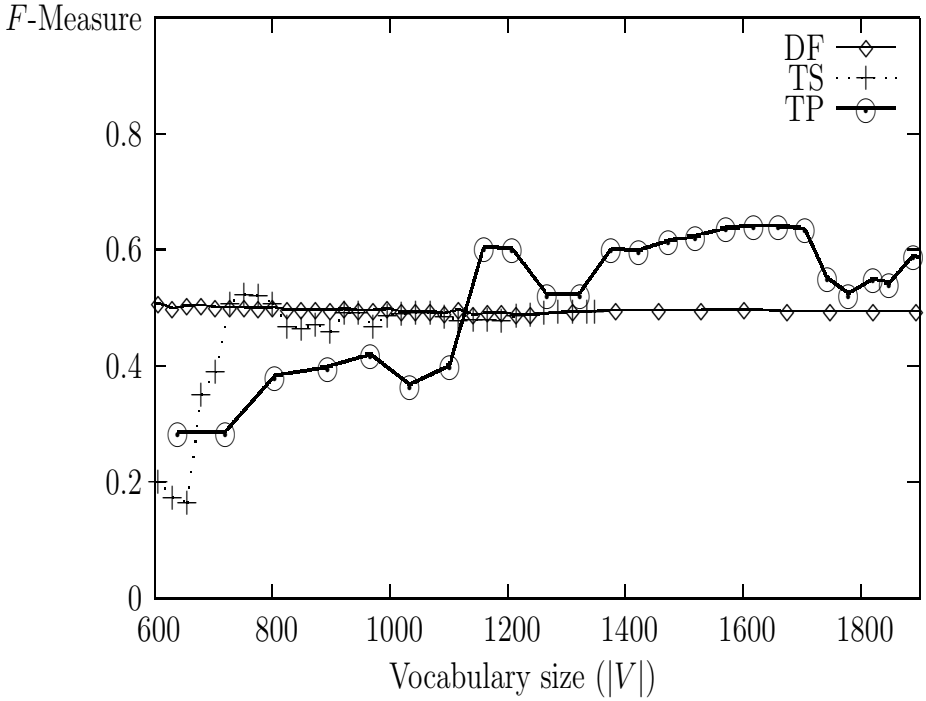


Figure 3.2: Behaviour of DF, TS and TP techniques in a subset of the *hep-ex* corpus.

On the one hand, the DF technique was very stable over the different thresholds of selection, since from the smallest selection thresholds, it included the most frequent terms in the text collection, contributing to maintain a minimum level of similarity during the clustering task. The baseline, i.e., the clustering done without term selection (F-Measure=0.5004), indicates that the DF selected terms keep coreference between both, the original dataset and the reduced versions of it. On the other hand,

the TS technique reached the maximum F-Measure after 700 terms. Moreover, after 900 terms it obtained stability as well as the DF technique did.

The TP technique outperformed the other two techniques. 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 in the abstracts. Moreover, we consider that some methods or measures are more stable than others when they use global data features instead of local ones. In this way, it is expected that low changes in the data lead to small changes in the final results. For instance, let us consider the DF term selection technique which is quite stable among different thresholds of vocabulary reduction. This behaviour may be derived from the fact that DF considers the term frequency among all the documents of the collection which is a global data feature. The TP technique uses instead a local criterion and, therefore, small changes in term frequencies may lead to different final results. One may purposefully design stable methods/measures by taking into account at least the following two considerations: 1) the use of global features representing the input dataset and, 2) the application of stabilisation methods during the execution. The following experiment presents an analysis of the TP selection process in order to control its instability.

Exp. 1.2: Analysis of the instability of TP

Although the TP technique obtained the highest F-Measure, 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 estimate 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 should be taken into account in order to find a possible formula which indicates the optimal number of terms to be selected by the TP technique. This hypothesis is explained into detail in the following paragraph.

Let $\mathcal{C}^{(i)}$ be a clustering of the text collection made of vocabulary-reduced texts, i.e.,

by whose terms which have been obtained by applying the TP technique and including only the i terms from each original text with frequency value closer to $TP(d)$ (see Section 2.2.1). Let $V(\mathcal{C}^{(i)})$ be the vocabulary of $\mathcal{C}^{(i)}$ and $\bar{df}_{V(\mathcal{C}^{(i)})}$ the $DF(t)$[8] average of terms t that belong to $V(\mathcal{C}^{(i)})$ but do *not* belong to $V(\mathcal{C}^{(i-1)})$. The $\bar{df}_{V(\mathcal{C}^{(i)})}$ value is close-related with the similarity among the texts. Clearly, the lowest value of $\bar{df}_{V(\mathcal{C}^{(i)})}$ is 1, and it means that the new terms added to $V(\mathcal{C}^{(i-1)})$ are not shared by the texts of $\mathcal{C}^{(i)}$. In our experiments it was observed that a decreasing in the $\bar{df}_{V(\mathcal{C}^{(i)})}$ value ($\bar{df}_{V(\mathcal{C}^{(i)})} < \bar{df}_{V(\mathcal{C}^{(i-1)})}$) contributed to change instances from an incorrect cluster to a correct one. Terms with low $\bar{df}_{V(\mathcal{C}^{(i)})}$ seem to help to distribute texts into the clusters and, therefore, we may use $\bar{df}_{V(\mathcal{C}^{(i)})}$ as an indicator of the goodness of a selection $\mathcal{C}^{(i)}$. In other words, this hypothesis may be used as an internal vocabulary-reduction validity measure.

Whenever the number of clusters (N_i) decreases after applying the clustering method to $\mathcal{C}^{(i)}$, a lower $\bar{df}_{V(\mathcal{C}^{(i)})}$ value with respect to $\bar{df}_{V(\mathcal{C}^{(i-1)})}$ means that new terms added to the vocabulary $V(\mathcal{C}^{(i)})$ increase the similarity between texts in $\mathcal{C}^{(i)}$. In such conditions $\bar{df}_{V(\mathcal{C}^{(i)})}$ indicates a good selection. One manner to express the above description is as follows: "a good clustering supposes that $\bar{df}_{V(\mathcal{C}^{(i)})}$ should be smaller than $\bar{df}_{V(\mathcal{C}^{(i-1)})}$ and N_i should be greater than N_{i-1}".

We can now define the goodness of the terms selected for $\mathcal{C}^{(i)}$ as shown in Eq. (3.4).

$$dfN_i(\mathcal{C}^{(i)}, \mathcal{C}^{(i-1)}) = \frac{(N_i - N_{i-1}) \times (\bar{df}_{V(\mathcal{C}^{(i)})} - \bar{df}_{V(\mathcal{C}^{(i-1)})})}{N_i} \tag{3.4}$$

In Table 3.1 a neighbour of the maximum value of dfN_i is shown. Row 1 shows the i number of terms selected by the TP technique; row 2, the size of the vocabulary of $\mathcal{C}^{(i)}$; row 3, the normalised 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-Measure does. Thus, dfN_i may be used for determining the optimal clustering set $\mathcal{C}^{(i)}$ that must be used in the clustering task.

[8] $DF(t)$ is the document frequency of the term t; see Section 2.2.2

Table 3.1: Some normalised values of dfN_i

i	**20**	**21**	**22**	**23**	**24**
$\|V(\mathcal{C}^{(i)})\|$	1,572	1,619	1,661	1,706	1,744
dfN_i	0.573	0.621	1.027	0.584	0.990
F**-Measure**	0.637	0.6411	0.6415	0.636	0.551

Exp. 1.3: Test over the full *hep-ex* corpus

A last experiment was carried out by using the entire collection and applying the three term selection techniques. Due to the small document length of the *hep-ex* corpus, the noisy words had a notably effect, mainly in the TP technique. TP selects one term at the time for each text and, therefore, a wrong selection may be crucial in the clustering task. In some cases, this selection process include/eliminate words that would change dramatically the composition of texts and, therefore, the canonic threshold used as parameter in the K-Star clustering method. We attempted to tackle this problem with an enrichment/expansion of terms selected by TP. It is not possible to solve this task by using related terms dictionaries like WordNet [39], since the terminology of the texts in question is very specialised (see [61]). The problem was finally overcomed by using co-ocurrence terms as an approximation to related words. We use the Pointwise Mutual Information (PMI) in order to compute the correlation degree among two given terms. PMI is an information theory based co-ocurrence measure discussed in [80] for finding collocations. Given two terms t_i and t_j, the PMI formula (see Eq. (3.5)) calculates 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 ocurrs alone.

$$PMI(t_i, t_j) = \log_2 \frac{P(t_i t_j)}{P(t_i)\ P(t_j)} \tag{3.5}$$

By taking into account the enrichment procedure, we defined a refined term selection technique based on the transition point technique in order to improve the results obtained over the full *hep-ex* corpus. This novel technique was named *Transi-*

tion Point and pointwise Mutual Information (TPMI), and basically uses $IDTP(t, D)$ (see Section 2.2.1) for the vocabulary reduction process and pointwise mutual information for enriching the selected terms. TPMI is then a refinement of the selection technique provided by TP. This technique is formally expressed as follows.

Let TP_V be the term associated to the transition point of the text $d = \{t_1, \ldots, t_k\}$. We can calculate the PMI score of each term t_i as $PMI(TP_V, t_i)$. The TPMI will assign as final score:

$$TPMI(t_i, d) = IDTP(t_i, d) * PMI(TP_V, t_i) \tag{3.6}$$

The results obtained by using this refined technique are shown in Figure 3.3.

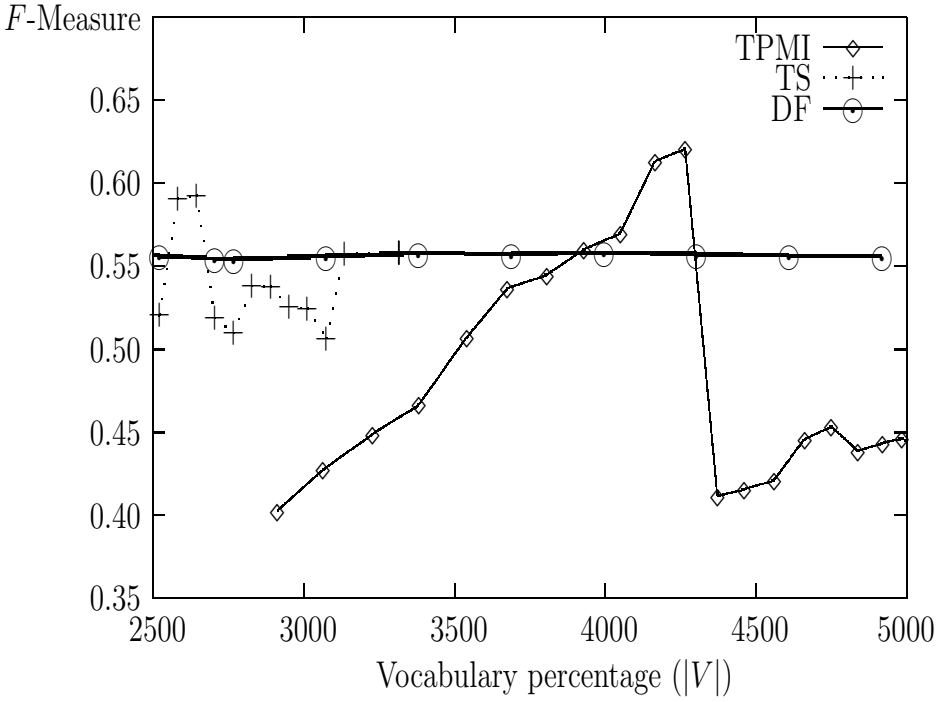


Figure 3.3: Behaviour of DF, TS and TPMI term selection techniques

We may see that this approach obtained the best value of F-Measure when we executed the experiments over the full *hep-ex* collection. Very similar clustering results were obtained for the DF and TS techniques, again over the full *hep-ex* corpus. The TS technique reached the maximum F-Measure (0.5925) with the 43% of the vocabulary

(the exact number of terms selected was 2,644), and only 3,318 terms were greater than the given threshold β. The DF technique showed to be very stable since it maintained its F-Measure values very close to the baseline (0.5919). The TPMI technique had a good high peak (F-Measure=0.6206) selecting only 20 terms from each document and obtaining a vocabulary size of 4,268 terms.

3.4.2 Experiment 2: A comparative study of different clustering methods using narrow domain short-text corpora

Clustering short texts of a narrow domain would imply using (1) a term selection technique and, (2) a clustering method. In order to investigate a possible dependence of the term selection techniques with respect to the employed clustering method, we carried out a preliminary comparative study of clustering algorithms with two datasets, the already used *hep-ex* corpus and the *CICLing-2002* one. The aim was to compare clustering results with different corpora (in size and balance).

As first step, *the term selection process*, we have used the three unsupervised techniques described in Section 2.2 in order to sort the vocabulary of each corpus in a non-increasing order according to the score of each TST. 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 a clustering method*, 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 the description of the mentioned clustering methods, see Section 2.1.

The experiments were carried out by using v-fold cross validation [25]. 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. Conse-

quently, the obtained results will not be casual through the use of a specific clustering method and a specific data collection. The v number of folds was set to ten and five respectively, for the *hep-ex* and *CICLing-2002* document collections.

In Tables 3.2(a) and 3.2(b) we show the maximum F-Measure values obtained for each term selection technique by using the five different clustering methods, considering two different corpora 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. However, once more TP shows to have an unstable behaviour.

Table 3.2: Maximum F-Measure obtained with five different clustering methods and three term selection techniques over both, (a) the *CICLing-2002* and, (b) the *hep-ex* corpora

	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

(a) The *CICLing-2002* corpus

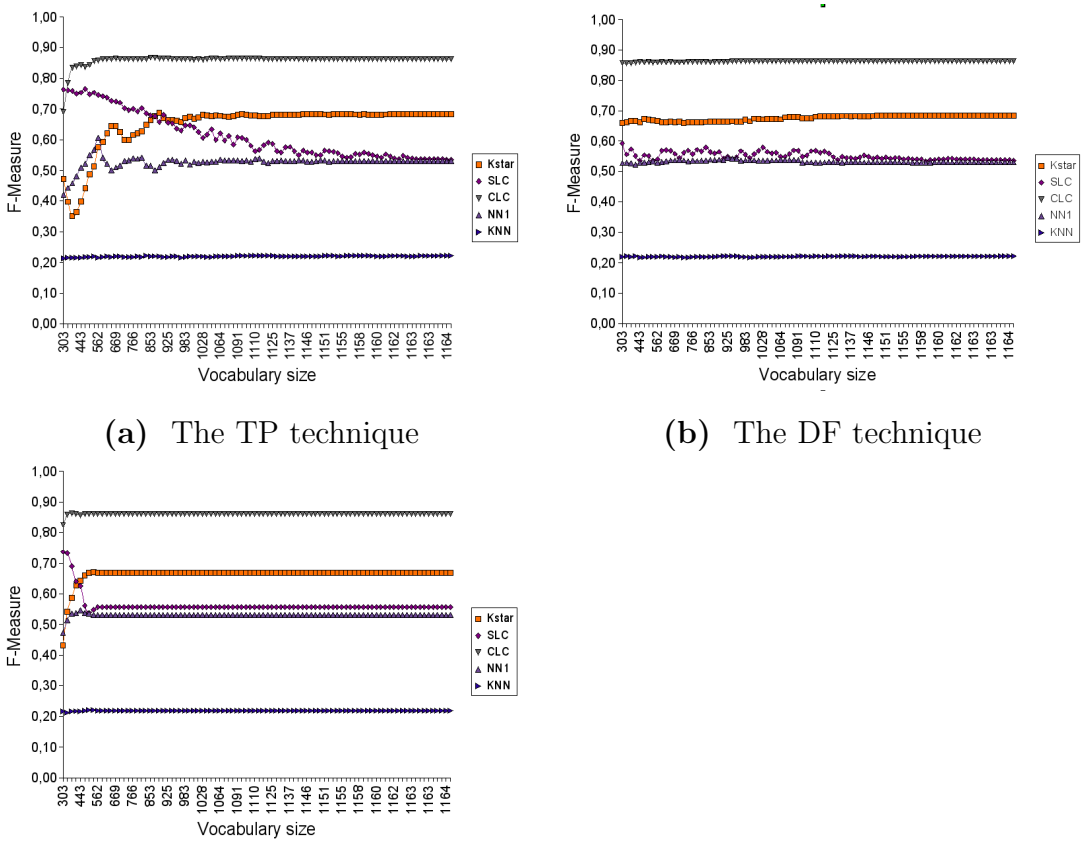
	TP	**DF**	**TS**
KStar	0.69	0.68	0.67
SLC	0.77	0.59	0.74
CLC	0.87	0.86	0.86
NN1	0.61	0.54	0.55
KNN	0.22	0.22	0.22

(b) The *hep-ex* corpus

The above results show that the term selection technique seems to be quite independent from the clustering method which is employed. In order to further investigate this hypothesis, we carried out an analysis of each selection technique and the five different clustering methods with bigger collection, i.e., the *hep-ex* corpus.

The performance of each term selection technique (TP, DF, and TS) over the *hep-ex* corpus by using the five clustering methods is shown in Figures 3.4(a), 3.4(b), and 3.4(c), respectively. It may be seen that the complete link clustering method obtains the best results for all the TSTs. The disadvantage of this clustering method family (CLC and SLC) is that the threshold used to cut-off the obtained hierarchical cluster structure was previously tuned over a subset of the *hep-ex* corpus (it strongly depends of the training corpus features). The KNN method obtained instead very poor results.

The iterative clustering method, K-Star, is the second best one that performed well over this corpus. The obtained results of this method and the advantage of being completely unsupervised makes it a good candidate for further experiments.



(a) The TP technique

(b) The DF technique

(c) The TS technique

Figure 3.4: F-Measure of the three term selection techniques as a function of the vocabulary size for the five clustering methods we considered (over the *hep-ex* corpus).

Figure 3.5 shows the standard deviation for different vocabulary percentages extracted from the complete vocabulary of the *hep-ex* corpus employing the TSTs. By obtaining the average of the three TSTs, we may observe that there exists some independence (with exception of the SLC method) on the behaviour of each clustering method, which suggests that the term selection process is independent from the clustering method.

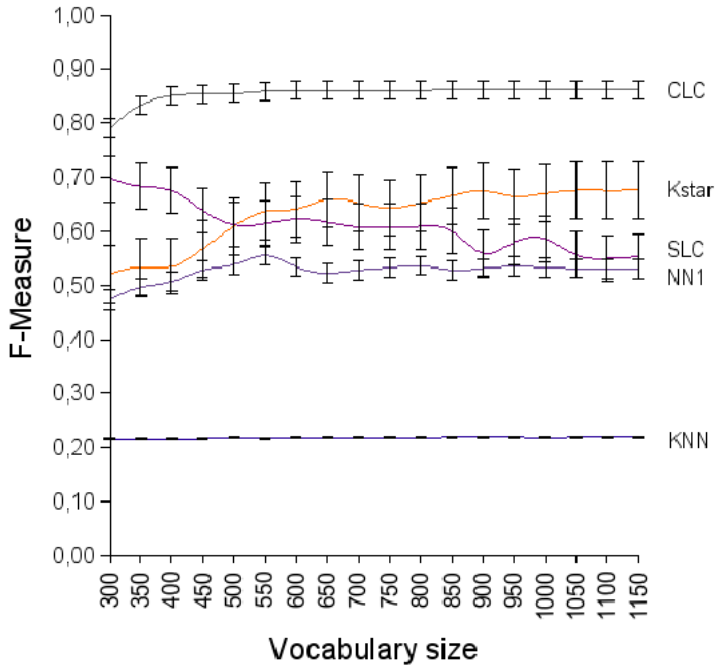


Figure 3.5: Average behaviour of all the TSTs analised with the five different clustering methods (using the *hep-ex* corpus)

3.4.3 Experiment 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 attempted to address this problem with the use of a new measure of distance between documents. The proposed similarity measure is based on the symmetric Kullback-Leibler distance which 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 carried out some experiments over two different narrow-domain corpora and our findings indicate that it is possible to use this measure for the addressed problem obtaining comparable results than when using the Jaccard similarity measure. The complete theoretical basis for the proposed clustering similarity measure is described in Section 2.1.1.

We used the three TP, DF and TS unsupervised techniques in order to sort the corpora vocabulary in a non-increasing order, with respect to the score of each TST.

Thereafter, we 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 involved the use of clustering methods; the three different clustering methods that obtained the best results in the previous experiment (Single Link Clustering, Complete Link Clustering, and K-Star) were used.

We carried out a v-fold cross validation evaluation for the experiments using five partitions for the *CICLing-2002* corpus and ten for the *hep-ex* collection. The quality of the obtained clusters was determined by means of the F-Measure.

In the experiments we have carried out, the TP technique slightly improved the DF and TS results, which reinforce the hypothesis made in [100]. Moreover, we have observed that there is not a significant difference between any of the symmetric KL distances. Therefore, we consider that the simplest of the used distances should be used. Tables 3.3 and 3.4 show, respectively, our evaluation results for all the Kullback-Leibler distances we implemented, over the *CICLing-2002* and *hep-ex* corpora. 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 term selection techniques, respectively. In the first column we have named as *KullbackOriginal*, *KullbackBigi*, *KullbackJensen* and, *KullbackMax*, the symmetric distance defined by Kullback and Leibler [68], Bigi [16], Jensen [43], and Bennet [13] [157], respectively.

Table 3.3: 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

The evaluation of the different Kullback-Leibler Distances (KLD) is presented in Tables 3.5 and 3.6. Our best approach (named *PintoetAl*) is compared with the results reported in [103]. We observed that the use of KLD obtained comparable results with respect to the Jaccard similarity measure. We believe that this behaviour is derived

Table 3.4: 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

from the size of each text. A smoothing process is needed; unfortunately, the number of document terms that does not appear in the corpus vocabulary may be extremely high. Further work should investigate this issue.

Table 3.5: 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 3.6: 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

3.4.4 Experiment 4: Evaluating with a standard narrow domain short-text corpus

The aim of this section is to verify the performance of the TP, DF and TS term selection techniques over the standard WSI-SemEval collection. This dataset has been used by other researchers in the WSI competition of SemEval and, therefore, we are able to compare our results directly with those of the ACL competition.

Figure 3.6 illustrates the performance over different percentages of vocabulary reduction of three different term selection techniques. Each point in the plotting

means the arithmetic mean F-Measure over the 100 corpora that made of the *WSI-SemEval* collection.

It may be seen that the best global value is obtained by the TP technique with an approximately 20% percentage of vocabulary reduction.

The DF technique obtained similar results than the TP did, but at difference of the latter technique, DF reduced in almost 80% the corpus vocabulary without a significantly loss of F-Measure. The TS term selection technique performed worst than the other TSTs.

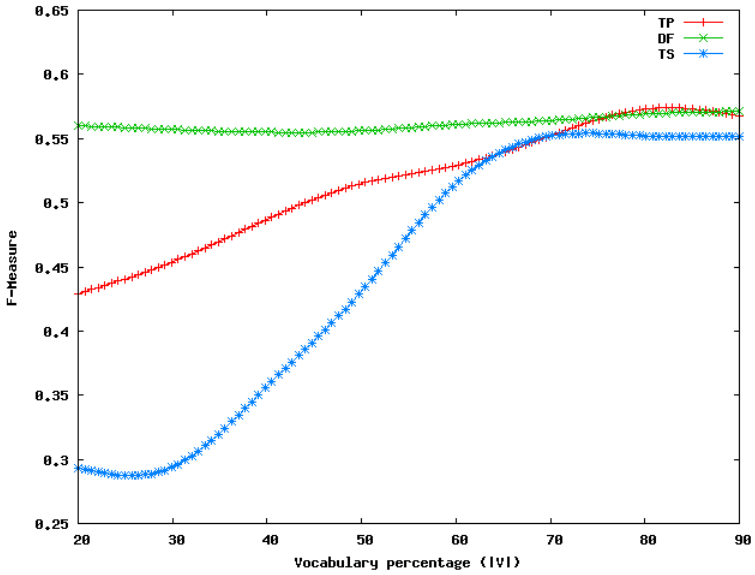


Figure 3.6: Clustering the *WSI-SemEval* collection with the K-Star clustering method

Table 3.7 shows the F-Measure obtained in the experiment carried out; whereas Table 3.8 presents the results each one of six clustering-based systems obtained participating in the WSI SemEval competition.

The TP and DF term selection techniques outperform two of the six systems with the consequent advantage of vocabulary reduction. In this test we may observe the impact of using term selection techniques when clustering short texts of narrow domains.

Table 3.7: F-Measure values obtained with the *WSI-SemEval* collection

Reduction	**TP**	**DF**	**TS**
80%	0.429	0.560	0.293
70%	0.454	0.557	0.294
60%	0.487	0.555	0.357
50%	0.515	0.556	0.431
40%	0.529	0.561	0.515
30%	0.552	0.564	0.552
20%	0.573	0.569	0.552
10%	0.568	0.571	0.552

Table 3.8: Standard F-Measure evaluation of the *WSI-SemEval* collection

System →	**1**	**2**	**3**	**4**	**5**	**6**
F-Measure →	0.787	0.663	0.661	0.639	0.561	0.379

3.5 Concluding remarks

3.5.1 Experiment 1

In the first experiment, we have proposed a new use of the transition point technique in the task of clustering of abstracts of a narrow-domain. We used as a corpus a set of documents (*hep-ex*) of the *High Energy Physics* domain, which led to experimenting with real collections composed of very short texts. The findings obtained from the execution of three unsupervised techniques (DF, TS and TP) was that TP outperformed the other two techniques over a subset of *hep-ex*. However, when the full document collection was used, the new TPMI term selection technique had to be developed in order to improve the previous unstable results obtained by the TP technique. TPMI takes advantage of a dictionary of related terms which is constructed over the *same* collection, by using pointwise mutual information since common or general-purpose dictionaries are not very useful (due to the very specialised narrow-domain vocabularies). After the calculation of a baseline in both the full corpus and

a subset of it, the experiments that were carried out allowed us to verify that the TPMI technique outperformed the other approaches.

Due to the instability of TP, we carried out an analysis to understand its behaviour and be able to determine the number of terms needed in the task. We observed that it does not seem possible to determine the number of terms that a term selection technique must choose in order to carry out the clustering task. It is vitally important to investigate further the study of stability control for the TP term selection technique. In fact, we consider this to be the key to obtaining a stable internal vocabulary-reduction validity measure.

3.5.2 Experiment 2

In this second experiment, we have carried out a comparative study of the behaviour of five clustering methods which were applied to two narrow domain short-text corpora (*hep-ex* and *CICLing-2002*) with very different characteristics. The documents of the two datasets are abstracts of scientific papers of very restricted scientific domains (high energy physics and computational linguistics). We have observed that the transition point technique obtained slightly better 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 would suggest that there is an independence between the term selection techniques and the clustering methods.

3.5.3 Experiment 3

In this experiment, we studied 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 were few differences in the use of any of the symmetric KL distances analysed. We evaluated the proposed approach with three different narrow domain short-text corpora, and our findings indicated that it is possible to use this measure to tackle this problem. We obtained results that were comparable to those that use the Jaccard similarity measure. Nevertheless, due to the fact that the KLD distance measure is computa-

tionally more expensive than the Jaccard one, this faster measure was used in the experiments described in the next chapters.

Even if we implemented the KLD to use it for clustering narrow domain short texts, we consider that this distance measure could also be employed for clustering more general domain and large size text corpora. The use of a smooth procedure should be useful since the vocabulary of each document is more similar to the corpus vocabulary. We consider that a performance improvement could be obtained by using a term expansion method before calculating the similarity matrix with the analysed KLD.

3.5.4 Experiment 4

In this last experiment, we studied the impact of the term selection techniques in a standard data collection. We compared the obtained results with those reported in [4]. The TP and DF term selection techniques outperformed two of the six systems with the additional advantage of vocabulary reduction.

In order to sum up the behaviour of the TP, DF and TS term selection techniques over three narrow domain short-text corpora, we show the plot of executing the same TSTs with both, the *CICLing-2002* and *hep-ex* corpora. Figures 3.7 and 3.8, respectively, show the F-Measure as a function of different reduced versions of the corresponding *CICLing-2002* and *hep-ex* full corpus.

The TP technique usually obtained good maxima; however, once more this technique becomes quite unpredictable. The TS technique obtained good terms for representing the documents with a very reduced vocabulary, but determining the vocabulary reduction threshold still remains a real problem, and, therefore, also the complexity of this term selection technique. Finally, the DF technique provided a very stable and fast procedure of term selection but moderate results. We consider this last term selection technique the most adequate for the future experiments.

Although the presented plots show in some cases low values of F-Measure (specially compared with standard evaluations such as SemEval), we will see in Chapter 5 and 6, how we succeeded to considerably improve the current obtained results.

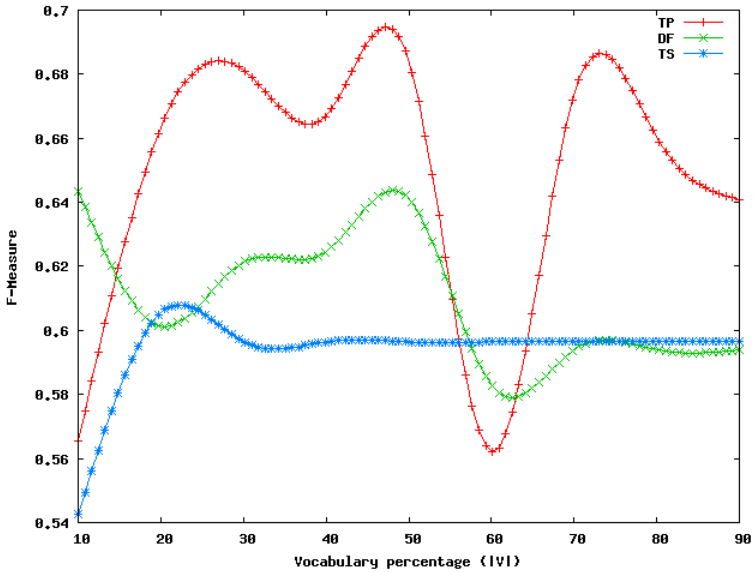


Figure 3.7: TSTs behaviour with the *CICLing-2002* narrow domain short-text corpus

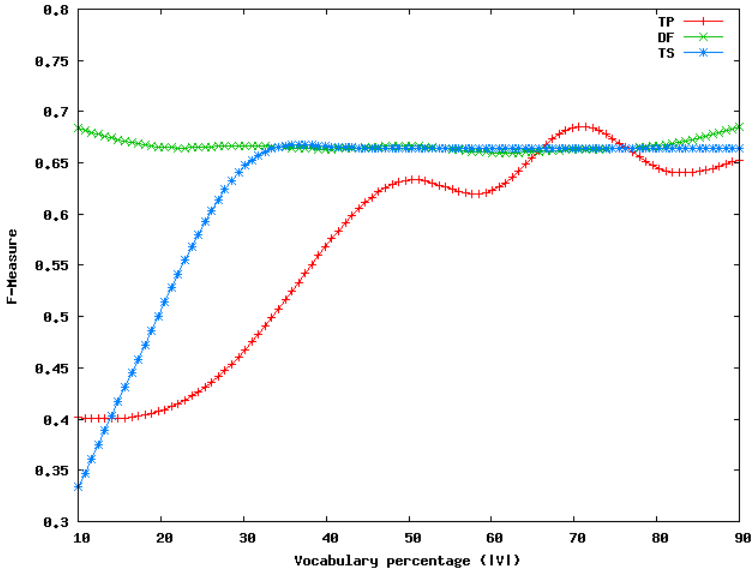


Figure 3.8: TSTs behaviour with the *hep-ex* narrow domain short-text corpus

Chapter 4

Evaluation of narrow domain short-text corpora

Evaluation of 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, it 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 may bias the expected results in a competition.

Moreover, when dealing with raw text corpora, if it is possible to find a set of features involved in 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 [35], the authors attempted to determine the relative hardness of different Reuters-21578[1] subsets by executing various supervised classifiers. However, in their research it is not defined any measure 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.

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578/

The aim of our proposal is to evaluate classifier-independing features which could be involved in the hardness of a given corpus. As far as we know, research in this field nearly have been carried out in literature.

For the purpose of our investigation, we took into account five different corpus features, namely *domain broadness*, *shortness*, *class imbalance*, *stylometry* and *structure*. 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 it exists a narrow gap between the gold standard of a corpus and the categories obtained through the execution of some classifier system.

The description of the features to be investigated together with related work is given as follows.

Domain broadness The goal is to evaluate the broadness of a given corpus. We assume (see for instance [144]) that it is easier to classify documents belonging to very different categories, for instance "sports" and "seeds", than those belonging to very similar ones, e.g. "barley" and "corn" (Reuters-21578). The attempt is to indicate the *domain broadness* degree of a given corpus. A binary classifier would assign, respectively, the tags *wide* to the former "sports-seeds" collection and *narrow* to the latter "barley-corn" one.

Shortness The term frequency is crucial for the majority of the similarity measures. When dealing with very short texts, the frequency of their vocabulary is very low and, therefore, the clustering algorithms have the problem of dealing with similarity matrices containing very low values. Therefore, we believe that independently of the clustering method used, the average text length of the corpus to be clustered is an important feature that must be considered when evaluating its relative difficulty. The formula introduced by Herdan [48] has extensively been used for measuring lexical richness of documents [137] such as, vocabulary richness for authorship attribution [50].

Class imbalance The document distribution across the classes is another feature that we consider important to take into account. There may be different lev-

els of difficulties depending on whether or not the corpus is more balanced. This feature is even more relevant when the corpus is used with the purpose of benchmarking different classifiers, for instance in the different tasks of an international competition such as SemEval[2]. The *imbalance* degree of a given corpus is also closely-related to the external corpus validation measure used (e.g. F-Measure) and, therefore, the obtaining of a single value for measuring it will clearly be beneficial. Two research projects that deal with the problem of class imbalance are the ones presented in [59] and [88]. Particularly, in the former paper it is claimed that class (category) imbalances hinder the performance of standard classifiers.

Stylometry It refers to the linguistic style of a writer. The goal is to determine the authorship of a set of documents. Even if in our case, the aim is not to attribute the autorship but to distinguish between scientific and other kind of texts. Due to the specific writing style of researchers, when the collection to be clustered is scientific then a new level of difficulty arises. There have been carried out several works on the statistical study of writing style (stylometry) field [36] which is stills an active research area [28, 51].

Structure The aim is to evaluate structural properties of the categories distribution given by the experts that manually classified a given corpus to be clustered. We validate the similarity and dissimilarity of the suggested groups or categories of the gold standard providing a single value which represents the structure of the document collection. When applied to the structure previously detected by some classifier, this area is named *clustering validity* and has been studied in the past by different authors. For instance, in [83] and Meyer07 different internal clustering quality measures are investigated. The Dunn's indices, for example, showed to perform well also in the experiments presented by Bezdek et al. in [15, 14], among others.

The corpora evaluation measures which are presented in this chapter may be

[2]http://nlp.cs.swarthmore.edu/semeval/

applied to all kinds of corpora, but we are expecting that they would be particularly useful when clustering documents of a narrow domain short-text corpus.

The supervised vs. unsupervised nature of each of the mentioned clustering evaluation measures is very important, since some of them may be obtained without any knowledge of the expected distribution of the documents, whereas other measures are focused to evaluate the gold standard of the target corpus. Those measures that do not need any information besides the document collection would be used to either evaluate general features of the collection or to improve clustering results in an unsupervised way, whereas the supervised measures will be devoted to evaluate the classification of the "experts".

In the following four sections we present measures for each of the previous briefly discussed corpora evaluation features. Each measure will be explicitly presented as supervised and/or unsupervised. In the case of the supervised ones, we will measure the *quality of the gold standard*, whereas when using unsupervised measures we will directly evaluate the corpus.

In Section 4.6 we discuss the obtained experimental results after evaluating several corpora with all the mentioned evaluation measures. In Section 4.7 we show a Web-based system for evaluating clustering corpora, which we have named "WaCOS: The Watermarking Corpora On-line System". Finally, the concluding remarks are given.

4.1 Domain broadness evaluation measures

The domain broadness of a given corpus is a very important classifier-independent feature that should be considered when evaluating a data collection to be clustered, in order to determine its possible relative hardness. 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 corpora domain broadness degree from a vocabulary-based perspective. We present the supervised and unsupervised version of the three approaches, one based on statistical language modeling, another based on vocabulary dimensionality and the last one based on the vocabulary overlapping.

4.1.1 Using statistical language modeling

Statistical Language Modeling (SLM) is commonly used in different natural language application areas such as machine translation, part-of-speech tagging, information retrieval, etc ([20, 81, 109]). However, it has been originally known by its use in speech recognition (see for instance [9]) which stills the most important application area.

Informally speaking, the goal of SLM consists in building a statistical language model 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.

This first approach makes use of statistical language modeling 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 variants, namely supervised and unsupervised.

In our particular case, we have considered that every hand-tagged category of a given corpus to be clustered has a language model. Therefore, if this model is very similar to the rest of models which were calculated for the other categories, then we could affirm that the corpus is narrow domain. The degree of broadness may be approximated by evaluating this proposed *supervised* approach over several corpora. Our proposal approaches in an unsupervised way the problem of determining the domain broadness of a given corpus. In fact, we calculate language models for v partitions of the corpus without any knowledge about the expert document categorization (gold standard). Following, we present the formal definition of n-grams based SLM and the two SLM-based evaluation measures.

The n-gram model

The n-gram model is up to now the most widely used SLM. We may express the probability of a sequence of n words (string) S: $P(S)$ through the chain rule as shown

in Eq. (4.1) and (4.2).

$$P(S) = P(w_1)P(w_2|w_1)P(w_3|w_1w_2)...P(w_n|w_1...w_{n-1}) \tag{4.1}$$

$$= \prod_{i=1}^{n}(P(w_i|w_1...w_{i-1})) \tag{4.2}$$

Since the complete n-gram model is difficult to calculate as n increases, an approximation of that model has been proposed by using bigram models. That is, we make the approximation that the probability of a word only depends on the presence of the immediately preceeding word. The bigram-based approximation of $P(S)$ may be expressed as shown in Eq. (4.3).

$$P(S) = \prod_{i=1}^{n} P(w_i|w_{i-1}) \tag{4.3}$$

Another sucessful approximation of the complete n-gram model is when using the trigram model (i.e., with $n = 3$). In this case, the immediately previous two words are used to condition the probability of the next word. The trigram-based model of $P(S)$ may be expressed as shown in Eq. (4.4).

$$P(S) = \prod_{i=1}^{n} P(w_i|w_{i-2}, w_{i-1}) \tag{4.4}$$

In order to estimate the parameters of the proposed models (bigrams or trigrams), one can use a large training corpus and calculate the n-gram frequencies. This procedure is shown in Equations (4.5) and (4.6) for bigrams and trigrams, respectively.

$$P(w_i|w_{i-1}) = \frac{freq(w_{i-1}\, w_i)}{freq(w_{i-1})} \tag{4.5}$$

$$P(w_i|w_{i-2}, w_{i-1}) = \frac{freq(w_{i-2}\, w_{i-1}\, w_i)}{freq(w_{i-2}\, w_{i-1})} \tag{4.6}$$

where $freq(w_i\, w_{i+1})$ is the number of times the sequence of words $w_i\, w_{i+1}$ is observed in the training corpus.

One of the problems that has to face the approximation of language models by using training corpora is that, even using a very large collection, there will exist a number of sequences of words that will not be seen when constructing the language model and, therefore, these n-grams will have zero probability. The way that SLM specialists solve this problem is by means of a smooth estimate of the probability of unseen events. There are various smoothing schemes which have been proposed such as backing-off, co-occurrence smoothing, count re-estimation, and deleted interpolation ([9]). In our particular case, we have used the Witten-Bell smoothing method [147] which calculates the discounted probability as shown in Eq. (4.7).

$$P(w_i|w_{i-n+1}\cdots w_{i-1}) = \frac{freq(w_{i-n+1}\cdots w_i)}{freq(w_{i-n+1}\cdots w_{i-1}) + W} \tag{4.7}$$

where W is the number of distinct words which follow $w_{i-n+1}\cdots w_{i-1}$ in the training data. If the model is based only on unigrams, then this value corresponds to the vocabulary size.

Perplexity and entropy

In order to compare similarities between two different language models we may use entropy. This information theory based technique allows to estimate how good a language model might be by averaging the log probability on per word basis for a piece of new text not used in building the language model [121].

For instance, if we are insterested in obtaining the similarity of language models between a new set of sequences from a test corpus and the training one, we may compute the entropy between them. Given a new sequence of n words (S), the entropy H on a per word basis of S is defined as follows [121]:

$$H = \lim_{n\to\infty} -\frac{1}{n}\sum_{w_i\in S} P(w_i)\log_2 P(w_i) \tag{4.8}$$

The probability of a set of word sequences belonging to a test corpus (LM_{Test}) of being used by the language model calculated in the training corpus ($LM_{Training}$) is computed by means of the average *logprob* (cross entropy) on a per word basis which

is shown in Eq. (4.9).

$$H(LM_{Test}|LM_{Training}) = -\frac{1}{|LM_{Test}|}\sum_{S\in LM_{Test}} P_{LM_{Test}}(S)\log_2 P_{LM_{Training}}(S) \quad (4.9)$$

Since the perplexity of a language is defined as two raised to the entropy power, the perplexity of a model (with respect to a test corpus) is defined as two raised to the logprob power (see Eq. (4.10)).

$$Perplexity(LM_{Test}|LM_{Training}) = 2^{H(LM_{Test}|LM_{Training})} \quad (4.10)$$

Perplexity like entropy may be computed per sentence or per word. Therefore, most of SLM toolkits provide both results. In the experiments we have carried out, we have used the SRILM toolkit which was primarily developed to be used in speech recognition, statistical tagging and segmentation. A detailed reference for this freely available[3] toolkit can be found in [135].

4.1.2 Two approaches for domain broadness assessment

Due to the fact that the perplexity is by definition dependent on the text itself, we should make sure that the text chosen is representative of the entire corpus [22]. In fact, in [121] 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.

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

The supervised domain broadness evaluation measure

Given a corpus D with a gold standard made up of k classes $\mathcal{C}^* = \{C_1^*, C_2^*, \cdots, C_k^*\}$. We obtain the language model of all the classes except C_i^* ($\bar{C_i^*}$) and, thereafter, we

[3] `http://www.speech.sri.com/projects/srilm/download.html`

compute the perplexity of the obtained language model with respect to the model of C_i^*. That is, we use the class C_i^* as a test corpus and the remaining ones as a training corpus in a leave one out process. Formally, the *Supervised* Language Modeling Based (SLMB) approach for determining the domain broadness degree of the corpus D may be obtained as shown in Eq. (4.11).

$$SLMB(D) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(Perplexity(C_i^*|\bar{C}_i^*) - \mu(Perplexity(\mathcal{C}^*))\right)^2} \tag{4.11}$$

where

$$\mu(Perplexity(\mathcal{C}^*)) = \frac{\sum_{i=1}^{k} Perplexity(C_i^*|\bar{C}_i^*)}{k} \tag{4.12}$$

The unsupervised domain broadness evaluation measure

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

The main problem consists in finding the correct way of splitting the corpus in order to allow that the evaluation by using statistical language models could make sense. An immediate solution would split the document collection in percentages (e.g. 10) and use for instance 10% or 20% for test and the remaining for training. This approach should work well for evaluating narrow domain corpora, but it would not be useful for wide domain corpora, since the expected model for training and test would be quite similar. 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 Modeling Based (ULMB) approach for assessing the domain broadness of a text corpus is calculated as follows.

Given a corpus D splitted into subsets C_i of l documents, we calculate the perplexity of the language model of C_i with respect to the model of a training corpus composed by all the documents not contained in C_i ($\bar{C}_i$).

Formally, given $\bar{C}_i \bigcup C_i = D$ such as $\bar{C}_i \bigcap C_i = \emptyset$ and k =Integer($\frac{|D|}{|C_i|}$) with $|C_i| \approx l$, the *unsupervised* broadness degree of a text corpus D may be obtained as shown in Eq. (4.13).

$$ULMB(D) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\Big(Perplexity(C_i|\bar{C}_i) - \mu(Perplexity(\mathcal{C}))\Big)^2} \tag{4.13}$$

where

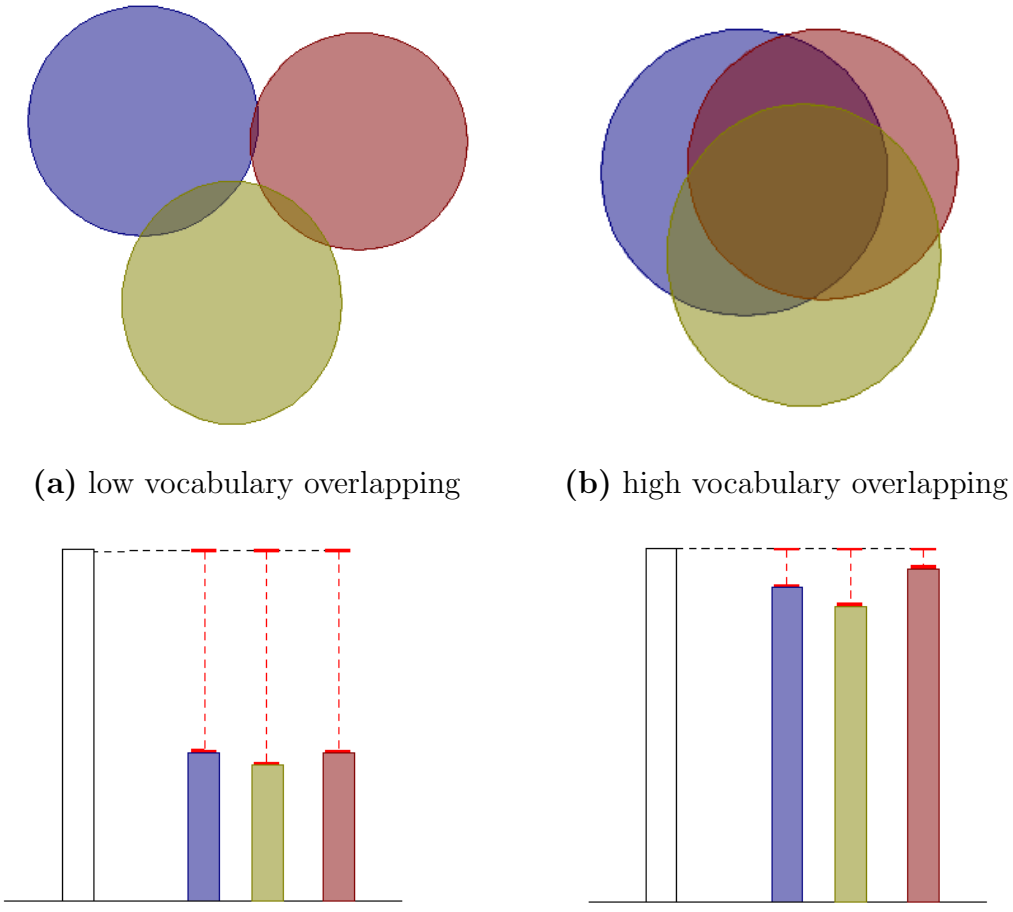
$$\mu(Perplexity(\mathcal{C})) = \frac{\sum_{i=1}^{k} Perplexity(C_i|\bar{C}_i)}{k} \tag{4.14}$$

4.1.3 Using vocabulary dimensionality

This measure of domain broadness assumes that corpora subsets which 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 vocabularies obtained from subsets of this corpus (with respect to the full corpus vocabulary) is greater than the one of a narrow domain corpus.

A graphical representation of this hypothesis is shown in Figure 4.1. In Figure 4.1(a) it is represented the low overlapping vocabulary expected when dealing with wide domain corpora, whereas in Figure 4.1(b) we show a high overlapping among all the classes that belong to a narrow domain corpus.

Figures 4.1(b) and 4.1(d) complement in graphical way the above hypothesis. They illustrate the same idea in terms of the vocabulary dimensionality. The white bar represents the corpus vocabulary, whereas other bars mean the size of each class vocabulary. In a wide domain corpus it is expected that the contribution of each class vocabulary to the corpus one will be lower than when using a narrow domain corpus. In the same figure, the sum of error lines will indicate the domain wideness degree. However, this value must be normalised with respect to the corpus size because the vocabulary of a text collection is highly depending of the number of documents.



(a) low vocabulary overlapping

(b) high vocabulary overlapping

(c) wide domain corpus

(d) narrow domain corpus

Figure 4.1: The use of vocabulary dimensionality on the assessment of domain broadness

The supervised approach

Given a corpus D with a gold standard made up of k classes $\mathcal{C}^* = \{C_1^*, C_2^*, \cdots, C_k^*\}$. If $|V(D)|$ is the cardinality of the complete document set vocabulary and $|V(C_i^*)|$ the vocabulary size of the class C_i^*, the *Supervised* Vocabulary Based (SVB) measure for the domain broadness of D may be written as shown in Eq. (4.15).

$$SVB(D) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(\frac{|V(C_i^*)| - |(V(D))|}{|D|}\right)^2} \tag{4.15}$$

The unsupervised approach

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 classes are unknown, we could then use each document instead of the corpus classes. Formally, given a corpus made up of n documents $D = \{d_1, d_2, \cdots, d_n\}$, 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 Eq. (4.16).

$$UVB(D) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{|V(d_i)| - |(V(D))|}{|D|}\right)^2} \tag{4.16}$$

4.1.4 Using vocabulary overlapping

As we mentioned at the beginning of this chapter, the automatic evaluation of the relative hardness of corpora to be categorized nearly has been attended in the literature. Sebastiani et al [127] studied the relative hardness of differents subsets of the Reuters-21578 data collection. They approached this problem by comparison of the performance of different classifiers.

We are more interested on relying the experiments without the use of any classifier and, therefore, we use the overlapping vocabulary inter categories/classes for this

purpose. We assume the category vocabulary overlapping to be closely-related with the corpus domain broadness and, therefore, with its relative hardness.

The measure we are presenting was introduced first in [105]. It calculates the vocabulary overlapping degree of a given document set. This measure may be used as both, external or internal clustering validity measure (i.e., taking into account the gold standard or not). For the particular purpose of this chapter, we calculate the vocabulary overlapping degree of each category suggested by a given gold standard.

Formally, given a corpus D with a gold standard made up of k classes, $\mathcal{C}^* = \{C_1^*, C_2^*, ..., C_k^*\}$, the Macro-average Relative Hardness[4] (MRH) of D could be expressed as shown in Eq. (4.17), whereas the micro-average Relative Hardness of D (mRH) may be calculated as shown in eq. (4.18).

$$MRH(D) = \frac{1}{k(k-1)/2} \sum_{\substack{i,j=1 \\ i<j}}^{k} \varphi(C_i^*, C_j^*) \tag{4.17}$$

$$mRH(D) = \sum_{i=1}^{k} \left(\frac{|C_i^*|}{|D| \cdot (k-1)} \sum_{\substack{j=1 \\ i \neq j}}^{k} \varphi(C_i^*, C_j^*) \right) \tag{4.18}$$

The similarity φ among two classes may be obtained by using either, the Jaccard coefficient or the cosine measure in order to determine their overlapping (see Equation (4.19) and (4.22), respectively). However, other less or more sophisticated measures could also be used, such as the one presented in [63] to calculate the plagiarism degree between two texts.

$$\varphi(C_i^*, C_j^*) = \frac{|C_i^* \bigcap C_j^*|}{|C_i^* \bigcup C_j^*|} \tag{4.19}$$

In the above formulae we have considered each class C_j^* as the "document" obtained from the concatenation of all the documents belonging to the class C_j^*. In Equation (4.20), $weight(t_i, C_j^*)$ is the weight of the term t_i in the j-th class C_j^*; $icf(t_i)$ (Eq. (4.21)) is the inverse *class* frequency of the term t_i and, finally, the similarity among two classes (Eq. (4.22)) is given by the cosine of the angle between the

[4]The concept of "relative hardness" of a corpus was introduced in [127]

vectorial representation of the classes of a given corpus. We have named the MRH (mRH) calculated with Eq. (4.19) as MRH-J (mRH-J), whereas the one that uses Eq. (4.22) as MRH-C (mRH-C).

$$weight(t_i, C_j^*) = tf(t_i, C_j^*) * icf(t_i) \tag{4.20}$$

$$icf(t_i) = log\left(\frac{|\mathcal{C}^*|}{cf(t_j)}\right) \tag{4.21}$$

where $cf(t_j)$ is the number of classes where the term t_i appears in.

$$\varphi(C_i^*, C_j^*) = \frac{\sum_l w_{il} * w_{jl}}{\sqrt{\sum_l w_{il}^2} * \sqrt{\sum_l w_{jl}^2}} \tag{4.22}$$

4.2 Stylometry-based evaluation measure

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

It has been observed that when the collection to be clustered is scientific then a new level of difficulty arises [100]. This observation may have its basis in 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 writing style (stylometry) field [36] which is stills an active research area [28, 51].

For the analysis of slylometry introduced in this chapter, 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 terms frequency distribution may be described by a particular distribution named “Zipfian”. This is a particularisation of a more general fact depicted in [156] which establishes the many types of data that could be described by the Zipfian distribution.

The approach presented in this section is similar to the one presented in [12] for the arabic language, however, we have restricted the analysis to determine whether or not a corpus is written by a group of persons with the same linguistic style.

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 Eq. 4.23 and the expected Zipfian distribution of terms as shown in Eq. (4.24). We used the classic version of the Zipf's law and, therefore, s was set to 1.

$$P(t_i) = \frac{tf(t_i, D)}{\sum_{t_i \in V(D)} tf(t_i, D)} \tag{4.23}$$

$$Q(t_i) = \frac{1/i^s}{\sum_{r=1}^{|V(D)|} 1/r^s} \tag{4.24}$$

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 Eq. (4.25).

$$SEM(D) = \sum_{t_i \in V(D)} P(t_i) log \frac{P(t_i)}{Q(t_i)} \tag{4.25}$$

4.3 Shortness-based evaluation measures

These evaluation measures calculate features derived from the length of a text, such as the maximum term frequency per document and the ratio between the document vocabulary size and the document length. The term frequency, for instance, is crucial for the major of similarity measures. When dealing with very short texts, we expect that the frequency of their vocabulary terms will be very low. Therefore, the clustering algorithms will have problems for detecting the correct classification, 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 normalised 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 $D = \{d_1, d_2, \cdots, d_n\}$, we present three *unsupervised* text length-based evaluation measures which take into account the level of shortness [49]. In the first and second approaches, we directly calculated the arithmetic mean of Document Lengths (DL) and Vocabulary Lengths (VL) as shown in Eq. (4.26) and Eq. (4.27), respectively. In Eq. (4.28) it is shown the third measure, introduced in [48], that obtains the average of Vocabulary vs. Document cardinality Ratios (VDR).

$$DL(D) = \frac{1}{n}\sum_{i=1}^{n}|d_i| \tag{4.26}$$

$$VL(D) = \frac{1}{n}\sum_{i=1}^{n}|V(d_i)| \tag{4.27}$$

$$VDR(D) = \frac{\log\left(VL(D)\right)}{\log\left(DL(D)\right)} \tag{4.28}$$

4.4 Class imbalance degree assessment measure

The document assignation to categories leads to identify those corpora with almost the same number of documents in each class/category as *balanced* or *unbalanced*. The class *imbalance* degree is an important feature that must be considered when corpora is categorized, since according to the imbalance degree there could exist different levels of difficulty. In fact, in [59] it is affirmed that class (category) imbalances hinder the performance of standard classifiers.

This feature is even more relevant when the corpus is used for benchmarking different classifiers, for instance in an international competition such as SemEval. Let us suppose that the corpus is totally unbalanced and, that for some reason there exist some clue of that. This fact could lead some participants to force their system to obtain the least possible number of clusters in order to get the best performance. In these conditions it would be quite difficult to carry out a fair evaluation and, therefore, understand what is(are) the best system(s).

For the purpose, we introduce a new *supervised* class imbalance evaluation formula. First, we assume that given a corpus D to be categorized with a pre-defined gold standard made up of k classes ($\mathcal{C}^* = \{C_1^*, C_2^*, \cdots, C_k^*\}$), the Expected Number of Documents per Class (ENDC) will be:

$$ENDC(D) = \frac{|D|}{k} \tag{4.29}$$

The aim of the proposed measure is to determine the Class Imbalance (CI) degree of a supervised corpus which has a gold standard. Thus, the *supervised* measure is calculated as the standard deviation of D with respect to the expected number of documents per class in the gold standard as shown in Eq. (4.30).

$$CI(D) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\Big(|C_i^*| - ENDC(D)\Big)^2} \tag{4.30}$$

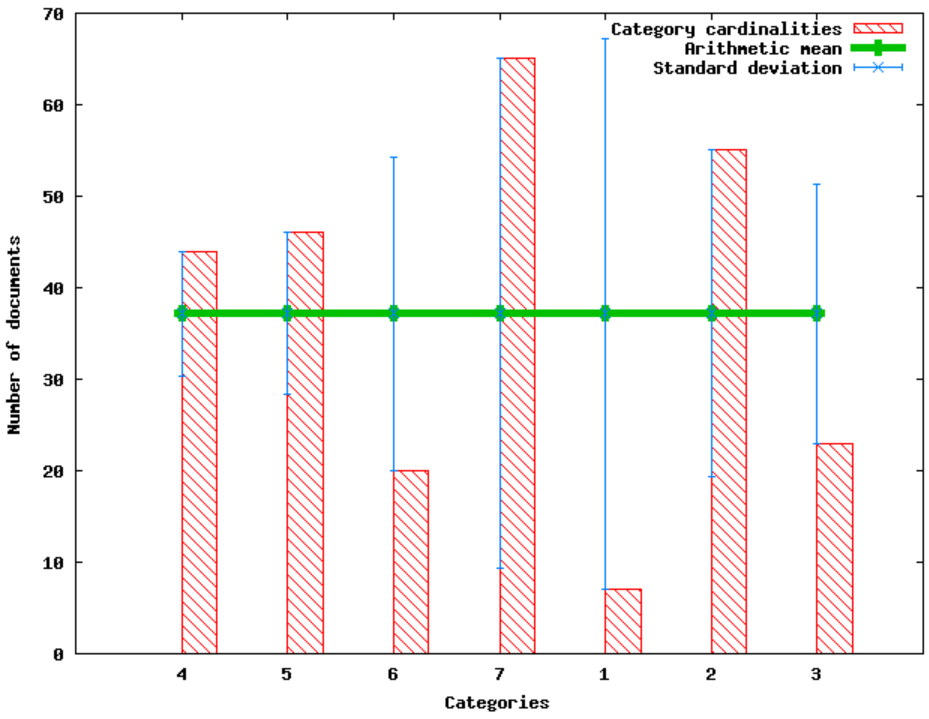


Figure 4.2: Example of class imbalance degree of a corpus

In Figure 4.2 we show the graphical representation of the class/category imbalance degree of a corpus and the error calculated by using the average number of documents assigned to each category (i.e., the difference with respect to the $ENDC(D)$ value).

4.5 Structure-based evaluation measures

Document clustering may be seen as the problem of finding a structure in a collection of unlabeled data [134, 152]. Therefore, we may use Internal Clustering Validity Measures (ICVM) to calculate the structural degree that would have a document distribution among a set of fixed categories. In particular, we have used two selected graph-based ICVM: one approach based on the Dunn index family and the expected density measure (see [15], [14] and [132]). However, a wide number of ICVMs also exist in literature [45, 46, 132, 57]

4.5.1 The Dunn index family

The Dunn index family identifies cluster sets that are compact and well separated. Let $C = \{C_1, ...C_k\}$ be a clustering of a set of objects D, $\delta : C \times C \rightarrow \mathbb{R}^+$ be a cluster to cluster distance and $\Delta : C \rightarrow \mathbb{R}^+$ be a cluster diameter measure. Then all the measures of the following form are called Dunn indices:

$$Dunn(C) = \frac{min_{i \neq j} \delta(C_i, C_j)}{max_i \Delta(C_i)} \tag{4.31}$$

For our analysis we have used Eq. (4.32) and Eq. (4.33).

$$\delta(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i, y \in C_j} \varphi(x, y) \tag{4.32}$$

$$\Delta(C_i) = \frac{\sum_{x,y \in C_i} \varphi(x, y)}{|C_i|(|C_i| - 1)/2} \tag{4.33}$$

where $\varphi : D \times D \rightarrow \mathbb{R}^+$ is a function which measures the distance between two objects. High values of $Dunn(C)$ correspond to a better structure and, therefore, we

will say that the higher is the value of $Dunn(C)$, the better is the cluster partition proposed by the expert.

4.5.2 The Expected Density Measure

A graph G = $\langle V, E, w\rangle$ may be called sparse if $|E| \in \mathcal{O}(|V|)$, whereas it is called dense if $|E| \in \mathcal{O}(|V|^2)$. Then we may compute the density θ of a graph from the equation $|E| = |V|^\theta$ where $w(G) = |V| + \sum_{e\in E} w(e)$, in the following way:

$$w(G) = |V|^\theta \Leftrightarrow \theta = \frac{\ln(w(G))}{\ln(|V|)} \tag{4.34}$$

θ can be used to compare the density of each induced subgraph $G' = \langle V', E', w'\rangle$ with respect to the density of the initial graph G. G' is sparse (dense) compared to G if $\frac{w(G')}{|V'|^\theta}$ is smaller (bigger) than 1. The expected density measure was formally introduced in [132]. Let $\mathcal{C} = \{C_1, .., C_k\}$ be the clustering of a weighted graph $G = \langle V, E, w\rangle$ and $G_i = \langle V_i, E_i, w_i\rangle$ be the induced subgraph of G with respect to the cluster C_i. Then the *Expected Density* $\overline{\rho}$ of a clustering $\mathcal{C}$ is obtained as shown in Eq. (4.35).

$$\overline{\rho}(\mathcal{C}) = \sum_{i=1}^{k} \frac{|V_i|}{|V|} \cdot \frac{w(G_i)}{|V_i|^\theta} \tag{4.35}$$

A high value of $\overline{\rho}$ should indicate a good clustering. Therefore, if we use the cluster distribution given by a gold standard, we will consider that the higher is the value of $\overline{\rho}$, the better is the gold standard.

4.6 Experimental results

In this section we present the obtained experimental results. The evaluation of the previously introduced corpus evaluation measures was done with standard corpora whose characteristics may be seen in Section 2.3. In order to easily understand the acronym given to each evaluation measure, we present in Table 4.1 the short

and complete names for each measure together with its corresponding category of evaluation and a tag identifying whether or not the measure is supervised.

Table 4.1: The corpus assessment measures

Short name	**Complete name**	**Category**	**Approach**
$SLMB$	Language model perplexity	Domain broadness	Supervised
$ULMB$	Language model perplexity	Domain broadness	Unsupervised
SVB	Category vocabulary dimensionality	Domain broadness	Supervised
UVB	Document vocabulary dimensionality	Domain broadness	Unsupervised
$mRH\text{-}J$	Vocabulary overlapping with Jaccard	Domain broadness	Supervised
$mRH\text{-}C$	Vocabulary overlapping with Cosine	Domain broadness	Supervised
SEM	Zipfian vs real term-frequency distribution	Stylometry	Unsupervised
DL	Text length	Shortness	Unsupervised
VL	Text vocabulary size	Shortness	Unsupervised
VDR	Vocabulary vs document length	Shortness	Unsupervised
CI	Document distribution	Class imbalance	Supervised
$Dunn$	A Dunn index family based measure	Structure	Supervised
$\overline{\rho}$	Expected density measure	Structure	Supervised

Table 4.2 illustrates all the evaluation measures which compute the domain broadness degree of a given corpus. The remaining evaluation measures are shown in Table 4.3. Each evaluation measure will be analised and discussed separately in the rest of this section.

4.6.1 Quality analysis of the corpus evaluation measures

In order to assess how well each broadness evaluation measure ranking correlates with their corresponding manually evaluated ranking, we have calculated one correlation coefficient between them. For the purpose of this measure by measure analysis, besides we have considered that there are no tied ranks, we have not made any assumptions about the frequency distribution of the evaluation measures. We would preferred to use the non-parametric measure of correlation named *Spearman's rank correlation coefficient* [71], however, since the equi-distance between the different corpora evaluation value could not be justified, then the correlation between each corpora

Table 4.2: The broadness-based corpus evaluation measures

Corpus	$SLMB$	$ULMB$	SVB	UVB	mRH-J	mRH-C
CICLing-2002	38.92	63.62	1.73	2.70	0.2036	0.0450
hep-ex	298.15	93.82	2.75	3.07	0.0542	0.2228
WSI-SemEval	195.02	130.62	1.80	3.06	0.0560	0.0183
WebKb-Training	262.26	628.60	0.50	1.77	0.6134	0.0835
WebKb-Test	337.39	218.85	0.44	1.60	0.6199	0.1200
R52-Training	627.60	143.10	4.38	4.62	0.0932	0.0563
R52-Test	565.81	177.54	4.58	4.82	0.0797	0.0485
R8-Training	603.95	135.87	3.67	4.76	0.2116	0.0353
R8-Test	545.69	134.60	3.84	4.89	0.1814	0.0454
20Newsgroups-Training	694.38	400.20	5.23	6.08	0.2472	0.0287
20Newsgroups-Test	786.02	455.38	5.21	6.05	0.2405	0.0300

Table 4.3: The remaining corpus evaluation measures

Corpus	SEM	DL	VDR	CI	$Dunn$	$\overline{\rho}$
CICLing-2002	0.3013	70.46	0.9117	0.0361	0.9991	0.8724
hep-ex	0.2711	46.53	0.9394	0.2795	0.9463	0.9433
WSI-SemEval	0.4477	59.58	0.9586	0.2263	0.9958	0.9875
WebKb-Training	0.2306	133.67	0.8877	0.0965	0.9985	0.7370
WebKb-Test	0.2273	136.23	0.8902	0.0966	0.9977	0.7098
R52-Training	0.1593	70.32	0.8849	0.0674	0.7601	1.1881
R52-Test	0.1196	64.30	0.8843	0.0677	0.8207	1.2196
R8-Training	0.1420	66.32	0.8865	0.1714	0.9605	1.1680
R8-Test	0.0980	60.05	0.8836	0.1689	0.9693	1.1895
20Newsgroups-Training	0.1543	142.65	0.8938	0.0040	0.9879	0.6192
20Newsgroups-Test	0.1437	138.73	0.8962	0.0050	0.9921	0.6046

evaluation measure rank was calculated by means of the *Kendall tau rank correlation coefficient* [65], which is described as follows.

Kendall tau rank correlation coefficient

The Kendall tau coefficient (τ) is calculated as shown in Eq. (4.36).

$$\tau = \frac{2 \cdot P}{(k \cdot (k-1))/2} - 1 \tag{4.36}$$

where k is the number of items, and 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 case of obtaining the value 0, then it is said that the rankings are completely independent.

Measure by measure analysis

We first analysed the broadness measure by means of statistical language modeling. Tables 4.4 and 4.5 show the obtained corpora 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 third column. Both supervised and unsupervised measures agree on 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 respectively 0.82 and 0.56. These values, indicate a strong agreement between the automatic and manual rankings. Now we may conclude that both measures perform well on evaluating the broadness degree of a given corpus in both, supervised and unsupervised ways.

The vocabulary dimensionality-based evaluations SVB and UVB are shown in Tables 4.6 and 4.7, respectively. The evaluation of 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 indicates a strong agreement between the automatic

Table 4.4: Ranking domain broadness with $SLMB$ (rank correlation value τ=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 4.5: Ranking domain broadness with $ULMB$ (rank correlation value τ=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

and manual rankings. Therefore, we consider that also these proposed measures may be used to calculate de broadness degree of clustering corpora.

The last way to measure the domain broadness of clustering corpora was done by means of the overlapping degree of vocabulary between the categories of the given corpus. In Tables 4.8 and 4.9 we may see the obtained results for the micro-averaged RH evaluation measures for the Jaccard-based (mRH-J) and Cosine-based (mRH-C) approach, respectively.

Table 4.6: Ranking domain broadness with SVB (rank correlation value τ=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 4.7: Ranking domain broadness with UVB (rank correlation value τ=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

The measure of quality of the automatic rankings by using the Kendall tau coefficient measure gave a value 0.09 for mRH-J and -0.05 for mRH-C.

The obtained values indicate that the measures are not a good indicator of corpus domain broadness. The reason of this behaviour is due to the fact that we have used the arithmetic mean of overlapping vocabulary as the final measure instead of the standard deviation.

The sylometry-based corpora evaluation measure determines the language style

Table 4.8: Ranking domain broadness with $mRH\text{-}J$ (rank correlation value τ=0.09)

Corpus	$mRH\text{-}J$	**Automatic ranking**	**Manual ranking**
hep-ex	0.0542	1	2
WSI-SemEval	0.0560	2	3
R52-Test	0.0797	3	8
R52-Training	0.0932	4	9
R8-Test	0.1814	5	6
CICLing-2002	0.2036	6	1
R8-Training	0.2116	7	7
20Newsgroups-Test	0.2405	8	10
20Newsgroups-Training	0.2472	9	11
WebKb-Training	0.6134	10	4
WebKb-Test	0.6199	11	5

Table 4.9: Ranking domain broadness with $mRH\text{-}C$ (rank correlation value τ=$-$0.05)

Corpus	$mRH\text{-}C$	**Automatic ranking**	**Manual ranking**
WSI-SemEval	0.0183	1	3
20Newsgroups-Training	0.0287	2	11
20Newsgroups-Test	0.0300	3	10
R8-Training	0.0353	4	7
CICLing-2002	0.0450	5	1
R8-Test	0.0454	6	6
R52-Test	0.0485	7	8
R52-Training	0.0563	8	9
WebKb-Training	0.0835	9	4
WebKb-Test	0.1200	10	5
hep-ex	0.2228	11	2

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 4.10 we may see the obtained values by the SEM evaluation measure for different corpora. The obtained Kendall tau correlation coefficient is 0.86 which 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 frequencies distribution results a very good indicator of the language writing style of a given corpus.

Table 4.10: Ranking the corpus language stylometry with SEM (rank correlation value τ=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

Tables 4.11 and 4.12 show the values obtained by using respectively, DL and VL corpus evaluation measures. Table 4.11 shows the arithmetic mean of document sizes and Table 4.12 presents the mean ratio between the vocabulary and document size for each corpus.

Table 4.11: Ranking of average document size obtained with DL (rank correlation value τ=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

As expected, the computed Kendall tau correlation coefficient value show a high agreement between the manually and automatically rankings with the DL and VL

Table 4.12: Ranking of average document vocabulary size obtained with VL (rank correlation value τ=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 4.13: Mean ratio of vocabulary and document size computed with VDR (rank correlation value τ=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

evaluation measures, obtaining respectively 0.96 and 0.78.

However, the values for the VDR measure shown in Table 4.13 are completely different. The correlation coefficient gave a value of 0.05, which means a 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 corpus, since the more are 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 similar performance than DL and VL did. However, it seems that VDR assess the complexity of the corpus (vocabulary vs. size) and not exactly the shortness.

In Table 4.14 we may see the values corresponding to the CI corpus evaluation measure. The higher is the value, the more unbalanced the corpus is, whereas the lower is the value the more balanced it is. It results that *hep-ex* is the most unbalanced collection, whereas, the *20-Newsgroups* and *CICLing-2002* are more balanced. In the case of corpora containing both, *training* and *test* split, it was considered in the manual ranking that both should be equally balanced and, therefore, when calculating the correlation coefficient, any of the two ranking values could be used. The obtained Kendall tau correlation tau is equal to one.

Table 4.14: Ranking of corpus balancing computed with CI (rank correlation value τ=1.00)

Corpus	CI	**Automatic ranking**	**Manual ranking**
20Newsgroups-Training	0.0040	1	1
20Newsgroups-Test	0.0050	2	2
CICLing-2002	0.0361	3	3
R52-Training	0.0674	4	4
R52-Test	0.0677	5	5
WebKb-Training	0.0965	6	6
WebKb-Test	0.0966	7	7
R8-Test	0.1689	8	8
R8-Training	0.1714	9	9
WSI-SemEval	0.2263	10	10
hep-ex	0.2795	11	11

Finally, Figures 4.15 and 4.16 show the ranking for both the measure based on the Dunn index family and the one based on the measure of expected density ($\overline{\rho}$), respectively. On the one hand, the Kendall tau correlation coefficient for the automatic ranking calculated with the *Dunn* formula ($\tau = -0.09$) indicates a clear independence of this ranking with respect to the manual ranking. The *Dunn* formula used in this work is only one of the multiple variants that may be used for

calculating the structure of a corpus gold standard. The results obtained with this particular implementation of the Dunn index family are not good. However, other ways of calculating this measure could be explored. On the other hand, the $\overline{\rho}$ measure shows a strong agreement between the two rankings ($\tau = 0.64$) which implies that the measure of expected density may be used to indicate whether or not the structure of a gold standard contains a well defined structure. We would like to emphasize that the calculation was performed over corpora of different kind and, therefore, our conclusions are general and not exclusively for narrow domain short-text corpora.

Table 4.15: Ranking of corpus structure computed with *Dunn* (rank correlation value τ=−0.09)

Corpus	*Dunn*	**Automatic ranking**	**Manual ranking**
CICLing-2002	0.9991	1	5
WebKb-Training	0.9985	2	8
WebKb-Test	0.9977	3	7
WSI-SemEval	0.9958	4	6
20Newsgroups-Test	0.9921	5	9
20Newsgroups-Training	0.9879	6	10
R8-Test	0.9693	7	1
R8-Training	0.9605	8	3
hep-ex	0.9463	9	11
R52-Test	0.8207	10	2
R52-Training	0.7601	11	4

Each evaluation measure presents a simple final value. The aim was to easily evaluate the features of each corpus. However, this final value relies on more than one calculations which may be useful to understand the obtained value as result. For instance, with respect to the stylometric measure, the simple value is obtained by means of the Kullback-Leibler distance between the Zipfian distribution and the corpus term frequencies distribution.

We may use a graphical representation in order to observe the particular behaviour of each corpus. For example, see Figure 4.3, where we show the term frequencies distribution of two different corpora with respect to its stylometry. Figure 4.3(a) shows the *CICLing-2002* corpus term frequencies distributions, whereas Figure 4.3(b)

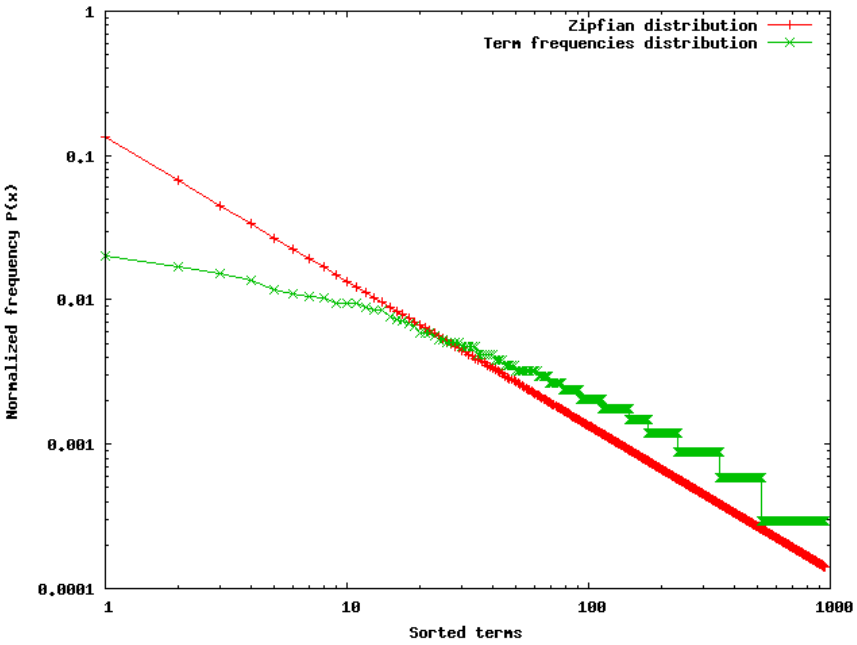
Table 4.16: Ranking of corpus structure computed with $\overline{\rho}$ (rank correlation value τ=0.64)

Corpus	$\overline{\rho}$	**Automatic ranking**	**Manual ranking**
R52-Test	1.2196	1	2
R8-Test	1.1895	2	1
R52-Training	1.1881	3	4
R8-Training	1.1680	4	3
WSI-SemEval	0.9875	5	6
hep-ex	0.9433	6	11
CICLing-2002	0.8724	7	5
WebKb-Training	0.7370	8	8
WebKb-Test	0.7098	9	7
20Newsgroups-Training	0.6192	10	10
20Newsgroups-Test	0.6046	11	9

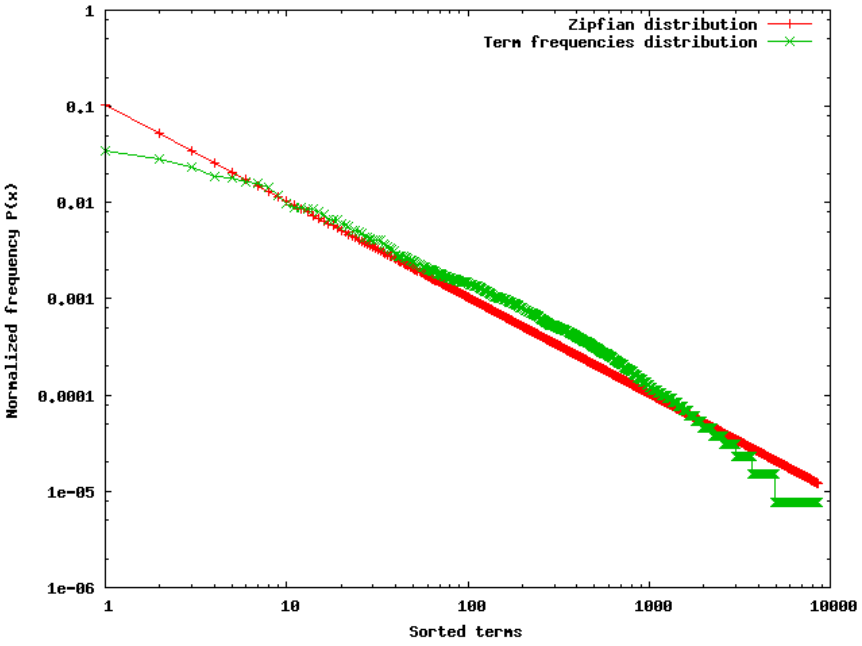
the same for the *R8-Reuters test* corpus. The difference between the term frequencies distribution and the Zipfian distribution is quite evident for the *CICLing-2002* corpus, whereas the *R8-Reuters* corpus correlates very well with its corresponding Zipfian distribution. The stylometric-related figures for all the evaluated corpora are shown in Appendix B, where we present three different representations of the term frequency distribution of a given corpus: 1) using all the term frequencies, 2) using only the unrepeated frequencies (by range) and, 3) using all the term frequencies but in a cumulative way.

In Figure 4.4 we may see the number of documents per category of the *CICLing-2002*(a) and *hep-ex*(b) corpora. The arithmetic mean and the standard deviation are also shown in order to see the way the final category balance value was calculated. It may be easily observed that the *CICLing-2002* corpus is quite balanced, at difference of the high unbalanced *hep-ex* corpus.

The above comparison was presented only as a simple example. If the reader is interested in seeing the characteristics of all the evaluated corpora, the complete results are presented in the Appendix B.

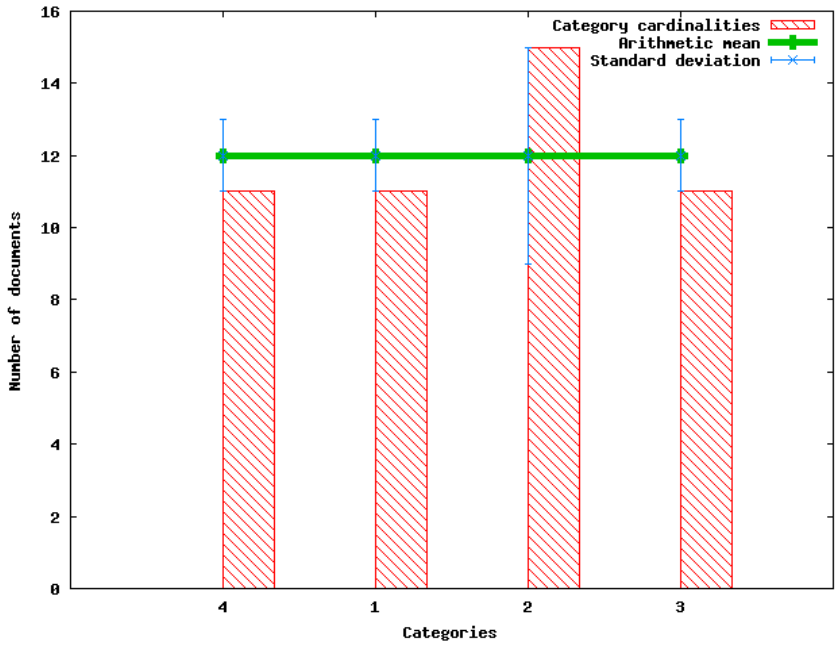


(a) The *CICLing-2002* corpus

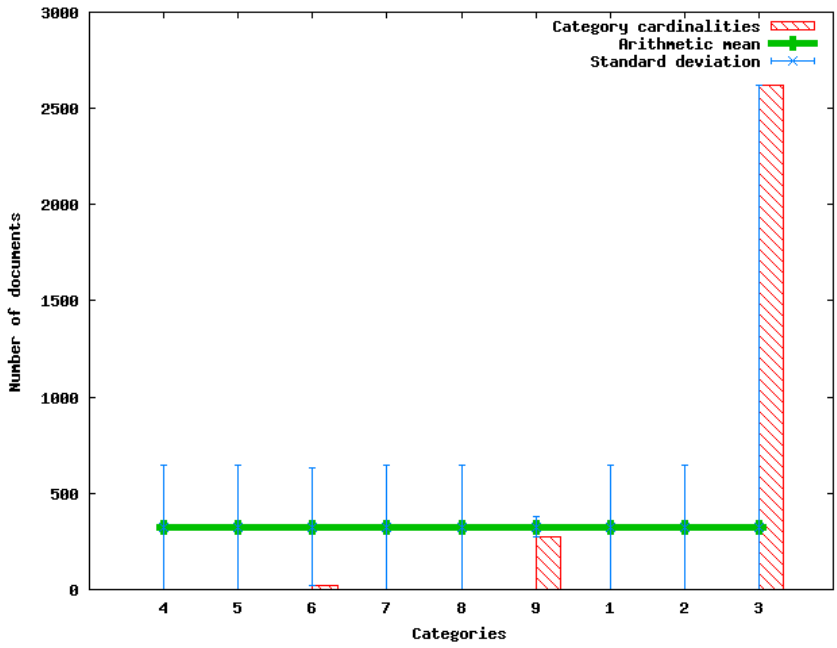


(b) The *R8-Reuters test* corpus

Figure 4.3: Graphical representation of stylometry-based characteristics



(a) The *CICLing-2002* corpus



(b) The *hep-ex* corpus

Figure 4.4: Graphical representation of the category balance degree

4.7 WaCOS: The Watermarking Corpus On-line System

This section briefly presents the web-based system developed for easily evaluating corpora features that we discussed in this chapter – broadness, shortness, balance, stylometric and structure. In Figure 4.5 we may see the main screen of WaCOS. When considering all the evaluation measures, it is expected to input two files, the corpus gold standard and the corpus itself. The *gold standard* file is used for the *supervised* measures, whereas the *unsupervised* ones only make use of the *corpus* file. The system will then upload the files to a server and, thereafter, it will execute all the scripts in order to present in a very naïve manner the obtained evaluation values.

Figures 4.6 to 4.12 show different snapshots of the watermarking corpus on-line system. Document cardinalities, balance per category, perplexity per category, Zipfian distribution, etc., are some of the graphical representations that may be obtained using the web-based system.

We have emphasized the use of this tool for assessing the quality of narrow domain sort-text corpora. However, the measures presented in this chapter may also be used for analysing the features of any corpora to be categorized. Morevoer, due to the unsupervised nature of some of the assessment measures, they may be also be useful with other kind of corpora. In general, we believe that the WaCOS system would be of high benefit for the linguistics and computational linguistics research community.

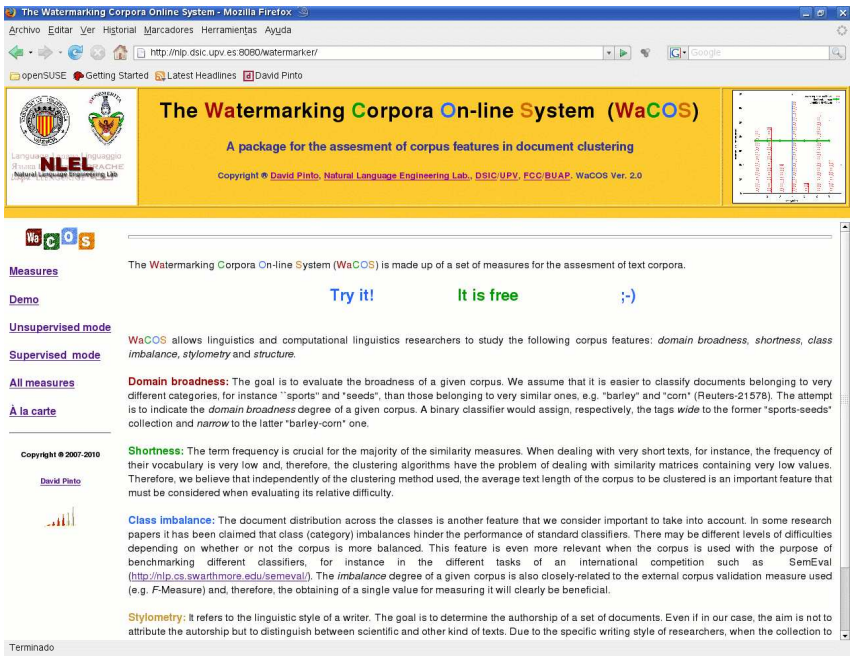


Figure 4.5: Snapshot of the WaCOS web site

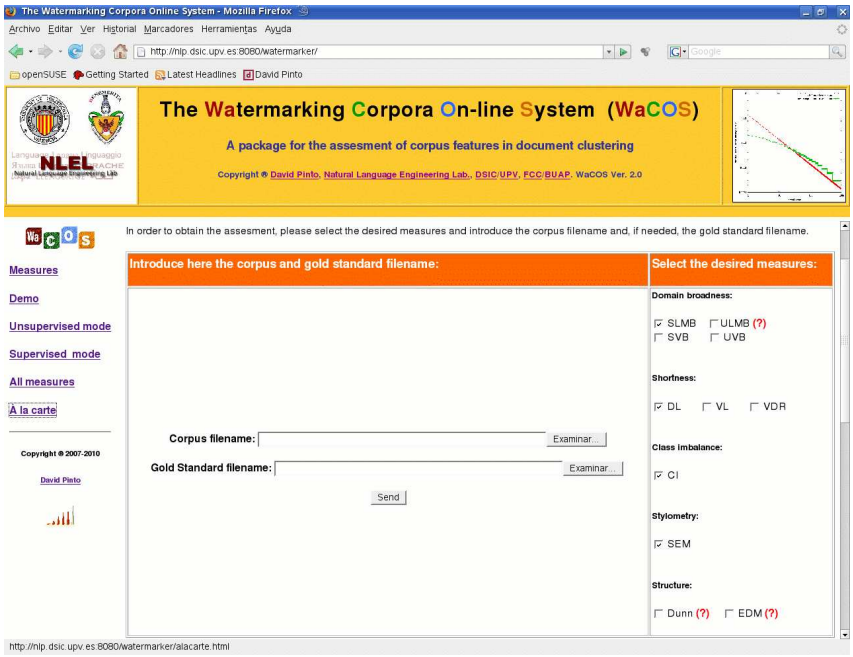


Figure 4.6: Personalised selection of measures (all, supervised, unsupervised, à la carte)

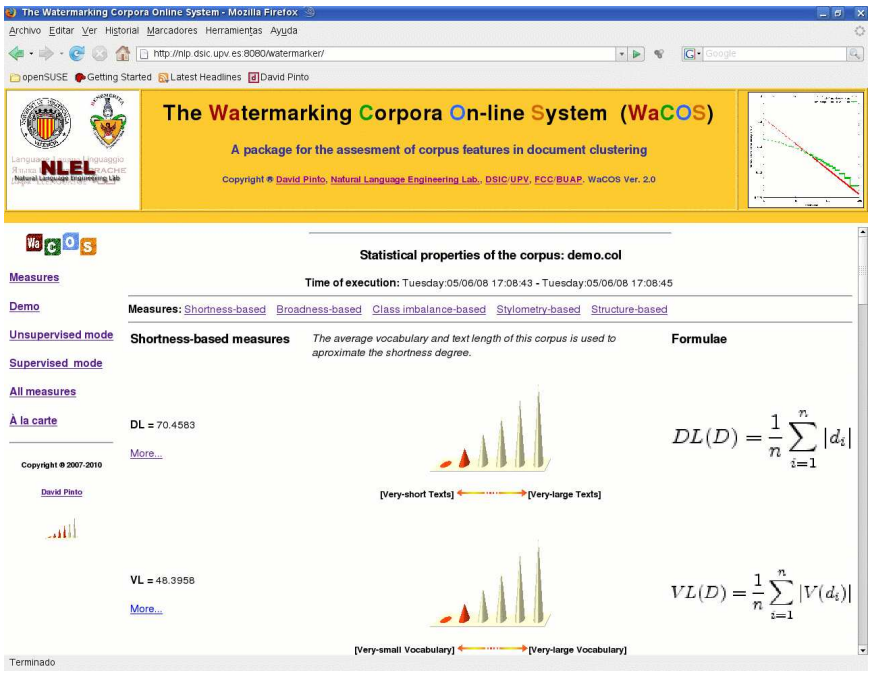


Figure 4.7: Naïve representation of the final evaluation values

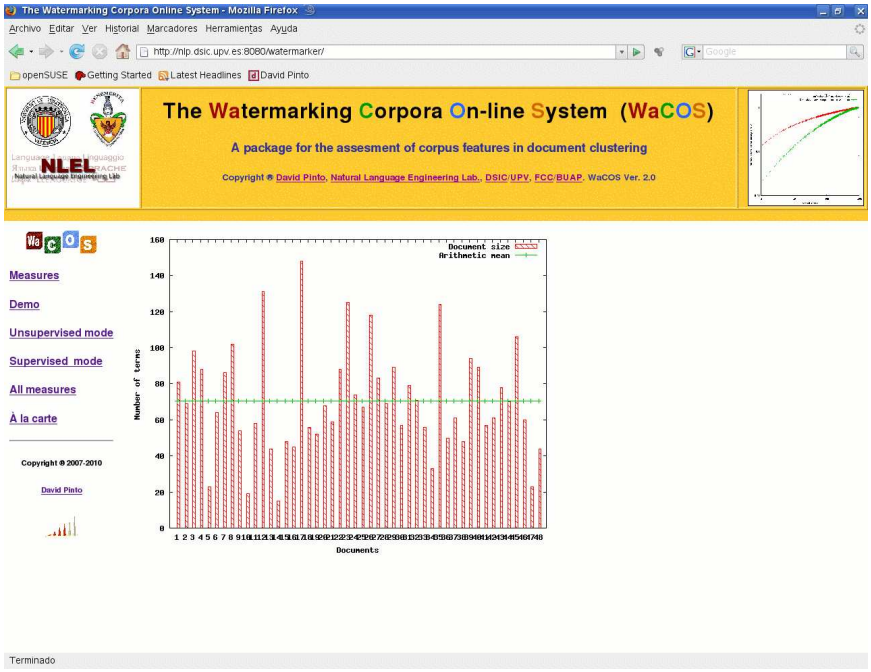


Figure 4.8: Document cardinalities

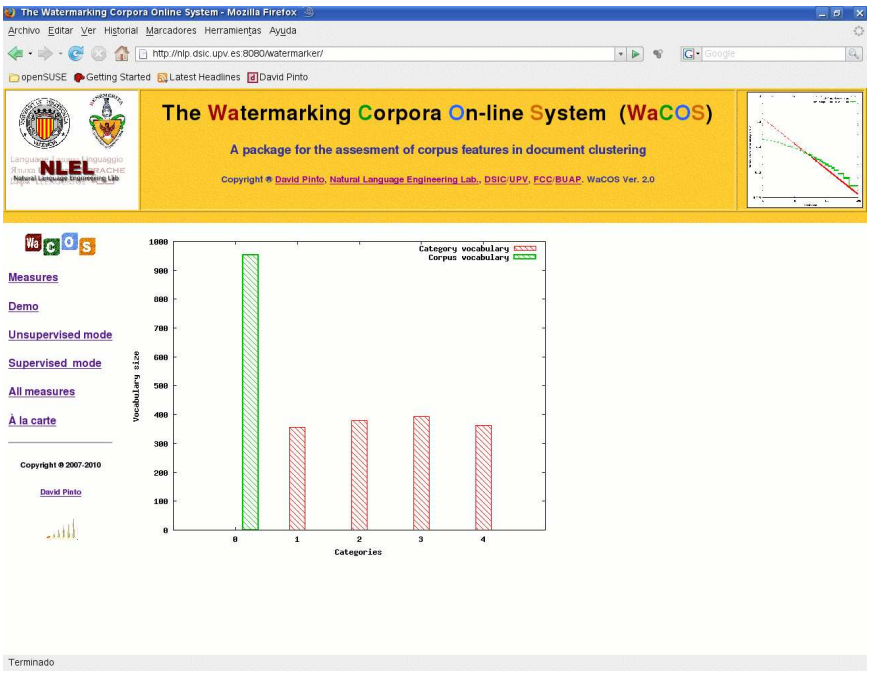


Figure 4.9: Corpus vocabulary vs. category vocabulary

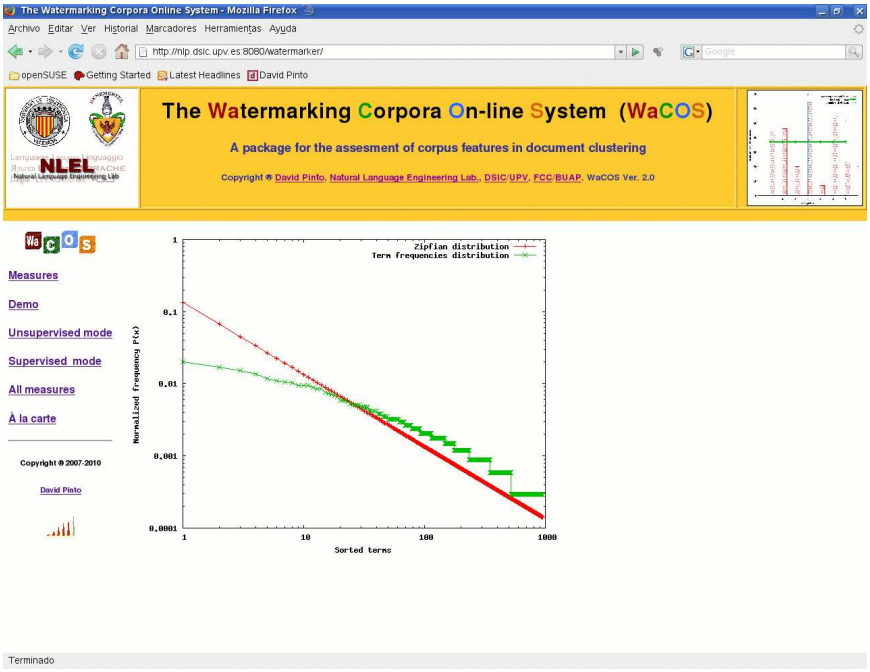


Figure 4.10: Zipfian vs corpus term frequency distribution

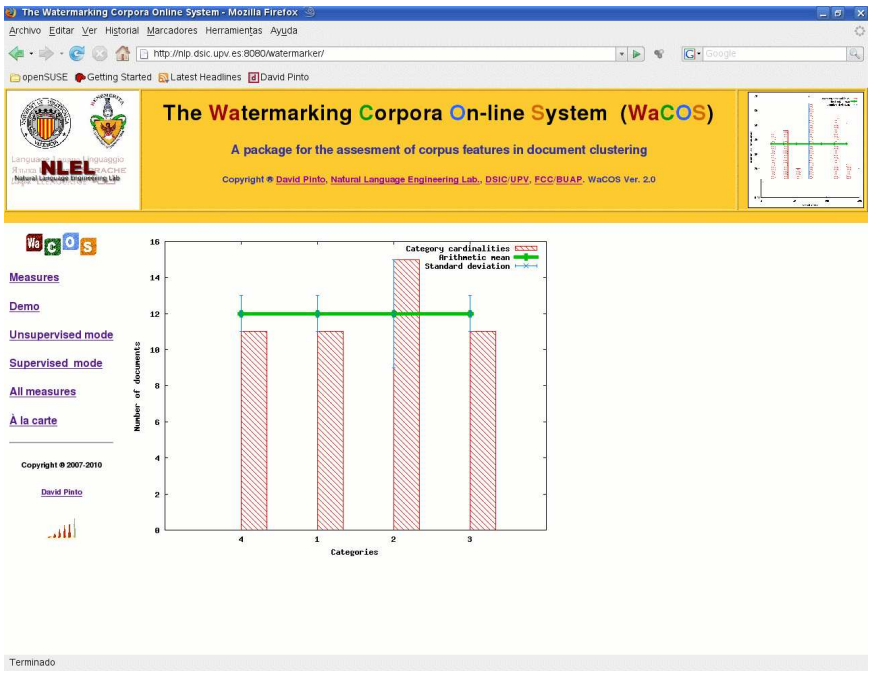


Figure 4.11: Graphical view of the class imbalance (per categories)

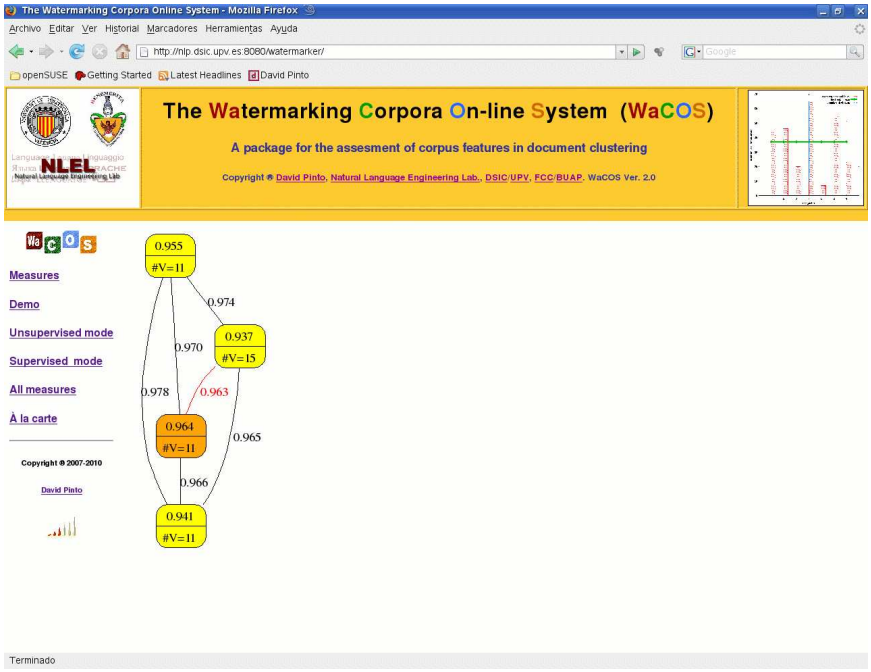


Figure 4.12: A graph-based representation of the corpus categories

4.8 Concluding remarks

In this chapter we have presented a set of corpus evaluation measures that can be used to either evaluate the proposed gold standard or to make decisions a priori when, for instance, clustering particular types of text collections such as narrow domain short-text corpora.

The evaluation measures were categorized into five different categories: *domain broadness*, *shortness*, *class imbalance*, *stylometry* and *structure*. All the proposed measures were executed over several corpora in order to determine their evaluation capability. We ranked each corpus according to the evaluation value given by the corresponding measure. We then calculated the Kendall tau correlation coefficient in order to determine the degree of correlation between the automatically obtained and the manually obtained ranking.

The findings were quite interesting. The major evaluation measures obtained a strong correlation with respect to the manual ranking. What also happened was that the micro-averaged RH evaluation measure results did not correlate well with the manually ranking. However, we analysed the obtained behaviour concluding that it would be outperformed by taking into account the standard deviation instead of a simple arithmetic mean over the ranking.

The most successful set of measures were computationally implemented in order to build a tool named Watermarking Corpora On-line System (WaCOS), which will allow researchers in different fields of linguistics and computational linguistics to easily assess their corpora with respect to the aforementioned corpus features.

Finally, the aim of this chapter was to formally investigate the quality of the narrow domain short-text corpora in order to further know about their features and, therefore, to take into account these findings in order to enable an appropriate approach to the clustering task.

Chapter 5

The self-term expansion methodology

In this chapter we present the obtained results after executing a set of experiments that highlight the need of improving text representation for narrow domain short-text corpora. We suggest to improve the representation of short-length documents by using a term enrichment procedure, often called *term expansion*. We consider that a properly enrichment of the target documents in the categorization task will improve the "semantic" similarity hidden behind of the lexical structure.

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 a external knowledge resource which usually helps in obtaining successful results [52, 128, 115].

However, we consider particularly important to attempt to use firstly the intrinsic information of the target dataset itself before using external knowledge. The motivation relies on the fact that the applications that use external resources such as WordNet are domain-dependent since they make use of supervised 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 which works without

the help of any external resource. We called this approach the *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 list calculated from the same target dataset by using the pointwise mutual information (see Eq. (3.5)) in order to determine the co-occurrence between the corpus terms. This list is then used to expand every term of the original corpus. Since the co-ocurrence formula captures relations between related terms, it is possible to hypothesize that the self-term expansion magnifies noise in a lower degree with respect to the meaningful information. Therefore, we consider that employing a clustering algorithm on the self-expanded corpus should allow us to obtain better results. The aim of this chapter is to investigate, whether the considered hypothesis is true or false and, therefore, to accept it or not it according to the obtained experimental results.

The chapter is structured as follows. In Section 5.1 we describe some research works which use external knowledge resources in order to improve text clustering. Section 5.2 introduces, in a formal manner, our novel technique for enriching narrow domain short-text corpora without any use of external knowledge. The proposed technique and the use of term selection techniques (see Section 5.3) allow us to construe a methodology which we have named *self-term expansion*. The experimental results with two narrow domain short-text corpora are presented in Section 5.4. Finally, in Section 5.5 the concluding remarks are given.

5.1 Term expansion using external knowledge

The expansion of short sentences is not new. In information retrieval, for instance, the expansion of query terms is a well researched topic which has shown to improve results in terms of when query expansion is not employed [113, 123, 7, 44, 117].

The availability of Machine Readable Resources (MRR) such as *Dictionaries*, *Thesauri* and *Lexicons* has allowed the application of term expansion to other fields of natural language processing like Word Sense Disambiguation (WSD). In [10] 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 (with respect to the examples of its gloss) is calculated in order to determine the correct sense of the ambiguous word. A variety of approaches have made use of a similar procedure. By using dictionaries, the proposals presented in [72, 146, 55] are ones of the most sucessful in WSD. In contrast, Yarowsky [151] tended used thesauri for his experiments. Finally, in [136, 116, 10] 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.

In [52, 53, 26, 97] authors suggested different ways of improving text clustering results by using ontologies. They have obtained a better similarity intra-documents incorporating background knowledge (using the WordNet ontology [53, 26] and the HowNet ontology [97], as mentioned in Chapter 3) into the document representation. In these papers it has been claimed that this procedure "always" improves performance compared to the best baseline.

In general, we agree with the fact that the enrichment of terms by using external knowledge resources should help in obtaining better results. However, the application of term expansion by using co-related terms will only improve the baseline results if we carefully select the external resource to use (i.e., with a priori knowledge of the domain), if any is available. Even more, we have to be sure that the Lexical Data Base (LDB) has been suitable constructed. Therefore, we consider that the use of a self automatically constructed LDB (using the same test corpus), may be useful. This assumption is based on the intrinsic properties extracted from the corpus itself. Our proposal is somehow related to the approaches presented in [126] and [112], where words are also expanded with co-ocurrence 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.

Avoiding the use of external resources in the process of term expansion with a combination of term selection techniques, as far as we know, has not been investigated previously. We consider that the proposed technique will be particularly useful when dealing with narrow domain short texts due to the fact that the short text corpora to

be clustered have low term frequencies. This does not pose much of a problem when these documents belong to wide domains, since they are easily classified nevertheless the low term frequencies. However, when considering narrow domain corpora, the situation is completely different, since the vocabulary overlapping between documents is generally high and, therefore, the clustering of the documents become to be a very challenging task.

The term expansion may alleviate the above problem, but as previously discussed, researchers generally use external resources to obtain the co-related terms. This approach is very effective when the topic of the domain is known and, of course, there exists a good lexical resource to be used. Unfortunately, the recent use of document clustering is applied to frameworks where both the specific topic and broadness domain are unknown a priori. Clustering may be applied at: blogs, information retrieval results, web people search, advertising, etc. We consider that the proposed external resource-independent approach will be beneficial for many of these tasks.

5.2 The self-term expansion technique

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 $d_i \in D$, it is satisfied two properties: **1)** if $t_j \in d_i$ then $t_j \in d'_i$, and **2)** 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 dataset, then we say that D' is the *self-term expansion* version of D.

The degree of co-occurrence between a pair of terms may be calculated through any co-ocurrence method, since this model is based on the intuition that two words 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 by in [47].

In order to fully appreciate the self term expansion technique, in Table 5.1 we show the co-ocurrence list for some words related with the verb "kill" (*soldier*, *rape*, *grenad*, *death* and *temblor*). The terms presented in the second column of the table

are the best co-ocurrence terms of their corresponding term at the left side of the same table. This list was obtained directly from the target corpus provided in an international competition[1] by using pointwise mutual information as the co-ocurrence method. Since the co-ocurrence list of each word is calculated after pre-processing the corpus, we present the stemmed version of the terms. The general process is graphically presented in Figure 5.1.

Table 5.1: An example of co-ocurrence terms

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

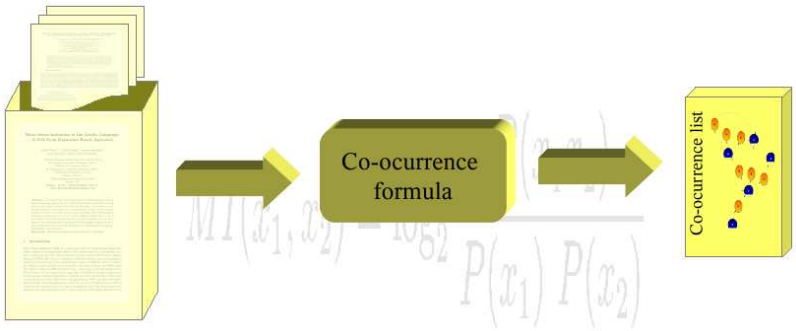


Figure 5.1: The extraction of the co-ocurrence list

Once the co-ocurrence 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. Figure 5.2 shows the graphical representation of the self-term expansion process.

[1]The competition was the task #02 of SemEval 2007

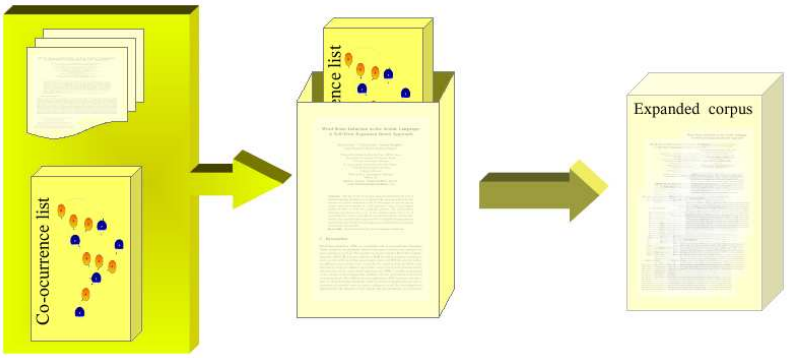


Figure 5.2: Self-expanding the clustering corpus

5.3 Term selection

The feasibility of obtaining co-ocurrence terms in the corpus will lead to a self-term expanded corpus generally richer than the unexpanded one, and with higher term frequencies with respect to the original corpus. However, the expanded corpus size will also be bigger than the unexpanded corpus, which could be a considerable drawback in terms of the time needed for computing, for instance, the clustering task. Thus, in addition to the self-term expansion technique, we propose to use a term selection technique in order to reduce the time needed for the chosen clustering algorithm. We have named this complete process as *self-term expansion methodology*, that is, *self-term expansion technique + term selection technique = self-term expansion methodology.*

The proposed method relies neither on any particular co-ocurrence method nor a specific term selection technique. The onus is on the researcher to decide which technique is better suited to the task in question. In our experiments, we have investigated the performance of two co-ocurrence methods and three term selection techniques in the narrow domain short-text corpora clustering task. The following section presents into detail the findings.

5.4 Experimental results

The aim of these experiments was to investigate the use of the self-term expansion technique in conjuntion with a term selection technique for clustering narrow domain short-text corpora. We use the two corpora that obtained the best agreement (with respect to the two corpus characteristics: *domain wideness* and *shortness*) according to the evaluation of corpora carried out in Chapter 4. *CICLing-2002* and *hep-ex* resulted to be the two corpora we are completely sure they are *narrow domain short-text* featured, and, therefore, the experiments of this chapter will use both of them.

We first observed the behaviour of the application of each TST on the complete collection (named as *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 the two unsupervised clustering methods K-Star and DK-Means, 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 three unsupervised term selection techniques described in Section 2.2 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 five and ten partitions respectively, for the *CICLing-2002* and the *hep-ex* corpus that, as formally investigated in the previous chapter, resulted 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-Measure (see Section 2.1.5).

The performance of each term selection technique with respect to the pre-processed original corpus by using the K-Star clustering method was evaluated in Chapter 3 (Figures 3.7 and 3.8 show the obtained plotting for the *CICLing-2002* and *hep-ex* corpus, respectively).

The experiments of Chapter 3 were carried out in order to have a first idea of how difficult the problem of clustering narrow domain short-text corpora could be. The experimental results shown in this chapter analyse the behaviour of both, the K-Star and DK-Means unsupervised clustering methods, applied to different subsets of the two corpora. The purpose was to analyse a possible bias when using a particular clustering method.

The corpora subsets were constructed by means of the three different term selection techniques; for each TST, we have reduced from 10% to 90% the corpus vocabulary by selecting the most relevant terms of the full corpus according to each TST.

Co-ocurrence methods

In order to determine the correct method for calculating the list of co-ocurrence used in the self-term expansion process, we have tested two different co-occurrence methods with different thresholds: n-grams and Pointwise Mutual Information (PMI). We investigated the behaviour of each term selection technique with respect to the use of the two self-term expansion techniques (n-grams based and PMI based) mentioned above. The automatically constructed co-ocurrence list served as basis to perform term expansion over all the subsets of each corpus, as well as the baseline.

The experimental results showed that it is possible to obtain a considerable improvement when using bigrams of frequency bigger or equal to 4, and pointwise mutual information with a threshold equal to 7. However, since the bigram counting is considered within the PMI formula then it was expected that PMI would outperform the bigrams results. The obtained F-measure values confirmed the previous hypothesis. In fact, in Figure 5.3 we may see how the baseline results calculated over the full *hep-ex* corpus 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 dataset implies an increasing 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 and

this makes possible to improve the original results.

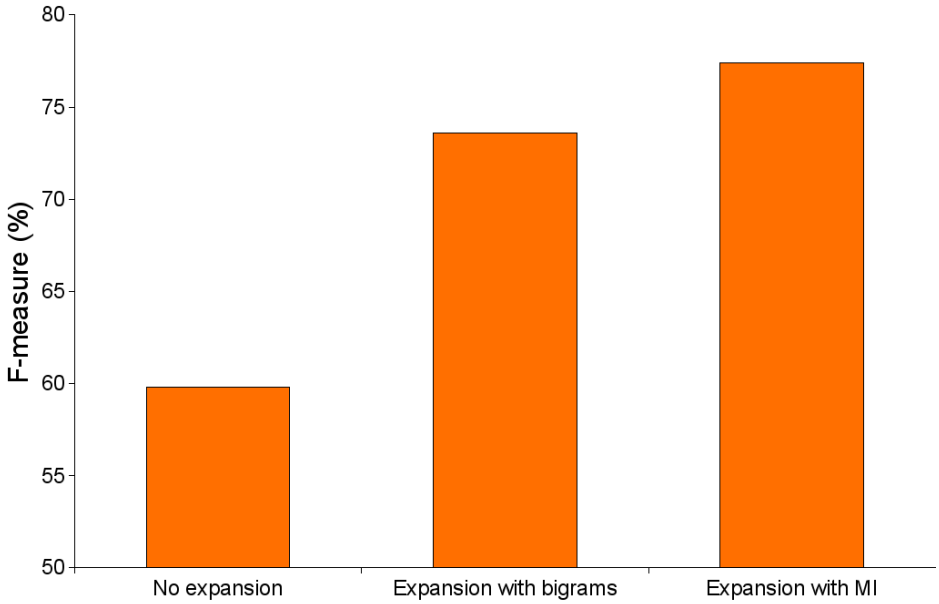
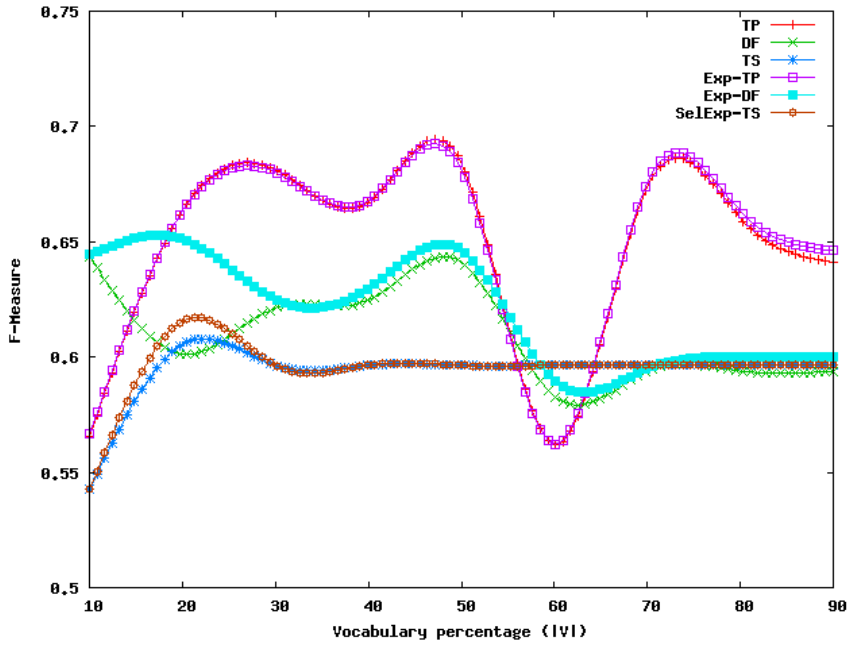


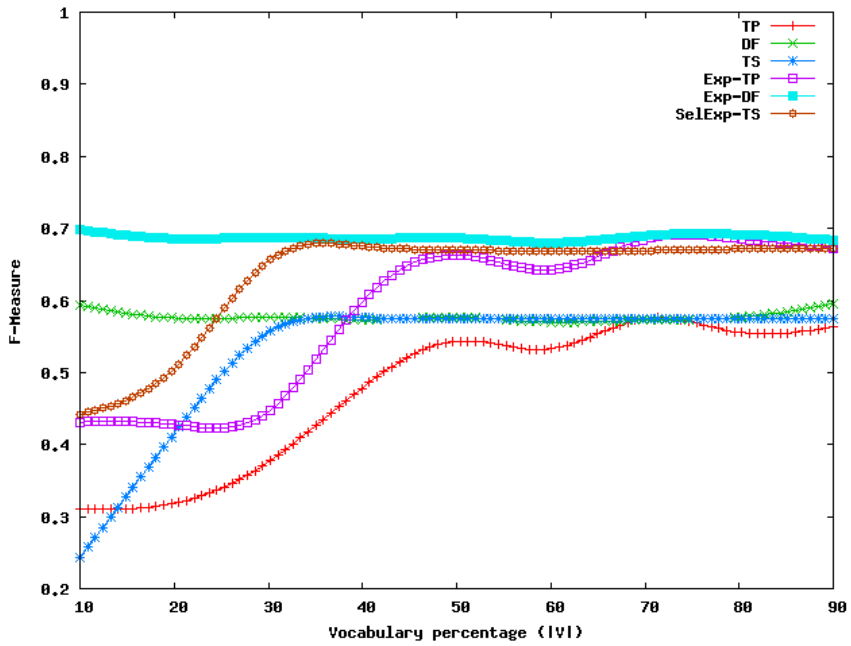
Figure 5.3: Effect of self-term expanding the *hep-ex* narrow domain short-text corpus with two co-ocurrence methods

Once observed that the best of the two co-ocurrence 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. Two possible approaches were tested: (i) we constructed the subsets of each corpora by using the three term selection techniques and, thereafter, we expanded each subset with a co-ocurrence list calculated over the same subset. The co-ocurrence list was calculated employing PMI; (ii) 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 same technique of reduction.

The plotting of the two approaches for both corpora is shown in Figures 5.4 and 5.5. When we applied the corpus reduction process before calculating the self-term expansion, we observed that a small improvement is obtained. It is remarkable that this improvement is almost the same for each corpora subset. Whereas, when we applied the self-term expansion technique before, we obtained different degrees of



(a) The *CICLing-2002* corpus



(b) The *hep-ex* corpus

Figure 5.4: Selection of terms *before* self-term expansion

improvement for each level of vocabulary reduction. We consider that this behaviour is mainly derived from the following fact: the higher is the length of a document the better is the term expansion and, therefore, when the term selection is applied first we are decreasing the possibility of obtaining all the co-ocurrence terms extracted by the self-term selection technique. We may then conclude saying that the best improvement is obtained with the second approach, that is, it is better to expand first the original corpus and, thereafter, to reduce the vocabulary by using a term selection technique.

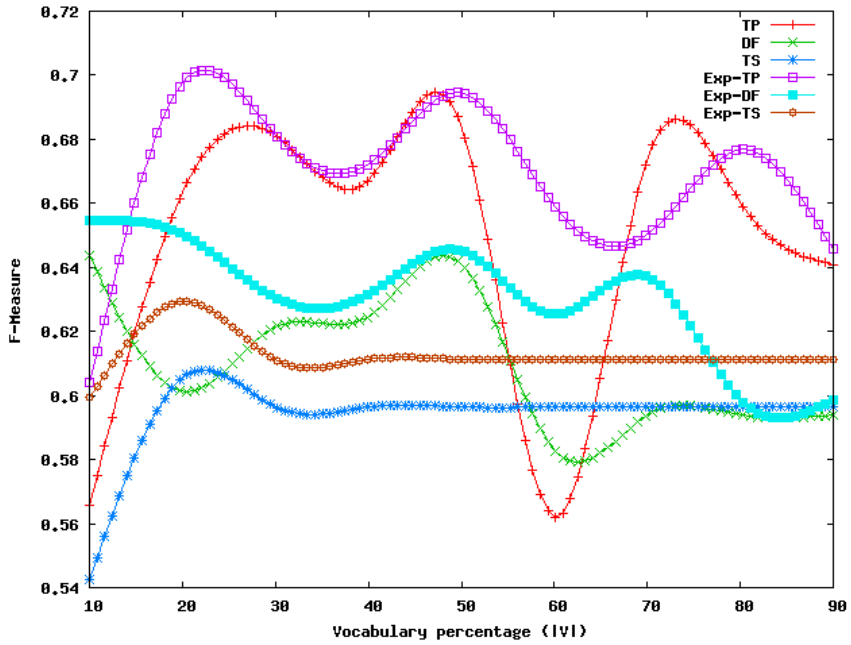
Moreover, when using the self-term expansion before applying the term selection technique, the best results are obtained with a very small size of the vocabulary. The discrimination of noisy terms is well-performed by each TST. For the *hep-ex* corpus, in particular, we have seen that the DF technique is the one which performs better in comparison with the other two TSTs. Besides that the DF technique obtains the best F-Measure results, it also reduces the corpus vocabulary of approximately 90%.

In Figures 5.6 and 5.7 we may see the behaviour of the self-term expansion (before the selection of terms, as in Figure 5.5) methodology, over the *CICLing-2002* and *hep-ex* corpus, respectively. These figures, which were obtained by executing the K-Star clustering method show the behaviour of each term selection technique separately. In both figures, we may see the obtained improvement which is independent of the term selection technique used.

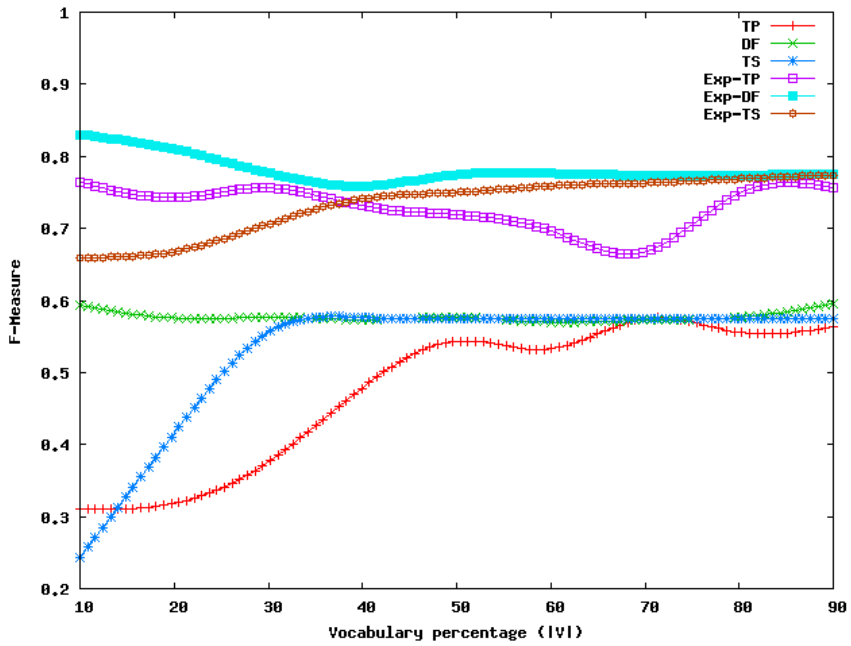
The *hep-ex* corpus, in particular, obtained a very high improvement for each corpus subset, whereas when the K-Star clustering method was applied to the *CICLing-2002* corpus, the DF and TS term selection techniques performed best the vocabulary reduction process with respect to the TP technique. The 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 of the text collection. We analysed that the self-term expansion technique did not have enough contexts to determine the correct co-ocurrence 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 fur-

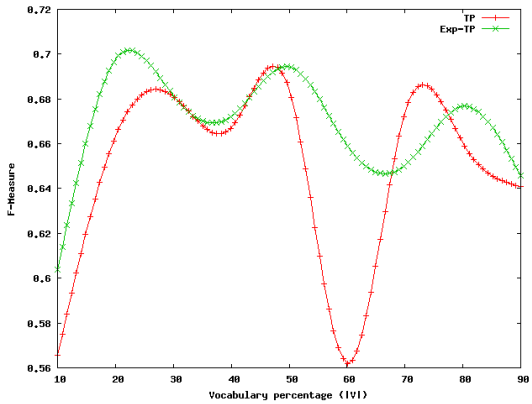


(a) The *CICLing-2002* corpus

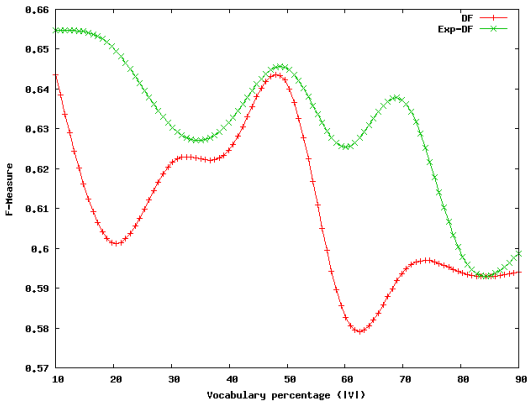


(b) The *hep-ex* corpus

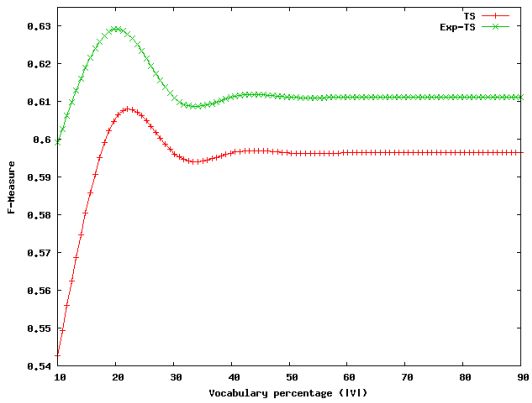
Figure 5.5: Self-term expansion *before* the selection of terms



(a) The TP technique



(b) The DF technique



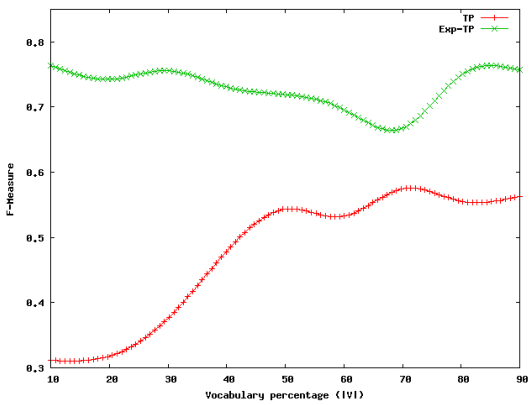
(c) The TS technique

Figure 5.6: Analysis of the self-term expansion methodology by using the K-Star clustering method over the *CICLing-2002* corpus

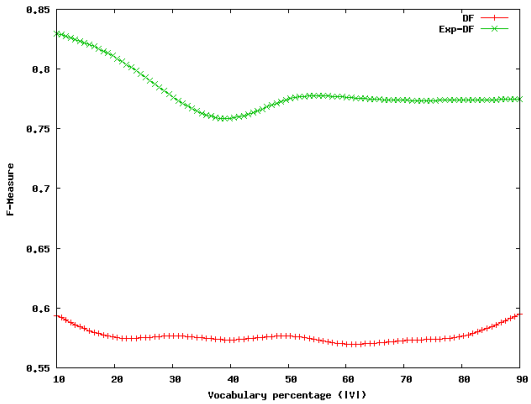
ther experiments employing the DK-Means clustering algorithm. We initialised the clustering method with the final results obtained by the K-Star clustering method (including the clusters and their number). The F-measure values for each corpora are shown in Figure 5.8.

The performance of each term selection technique with the *CICLing-2002* and the *hep-ex* corpora is shown, respectively, in Figures 5.9 and 5.10.

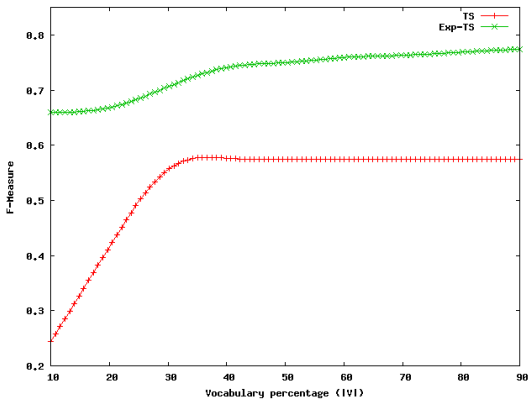
We may conclude that the clustering of narrow domain short-text corpora obtains better results if the original corpus is previously enriched employing the self-term expansion methodology. This is due to the high vocabulary overlapping associated to this kind of corpora which allows to determine co-ocurrence relationships that may



(a) The TP technique



(b) The DF technique

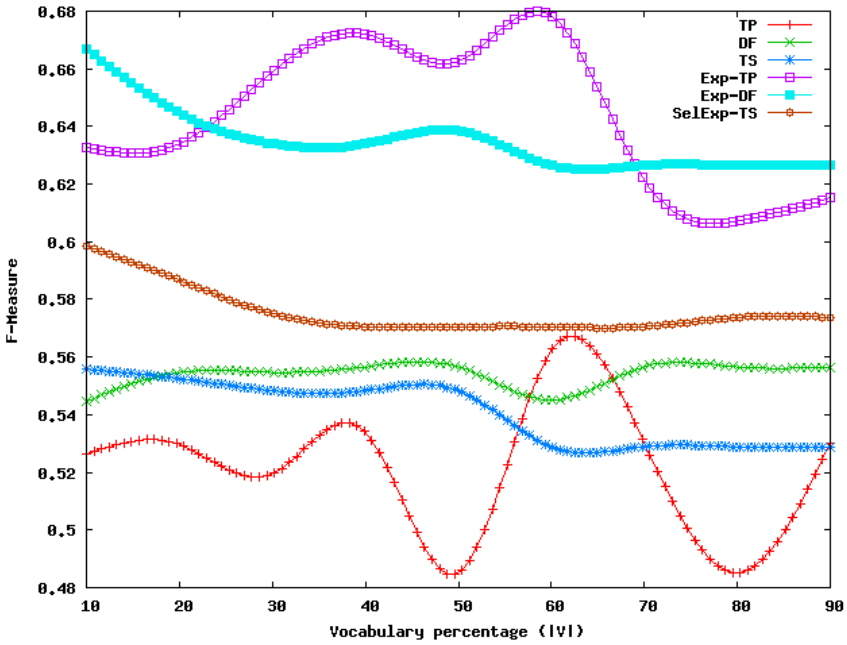


(c) The TS technique

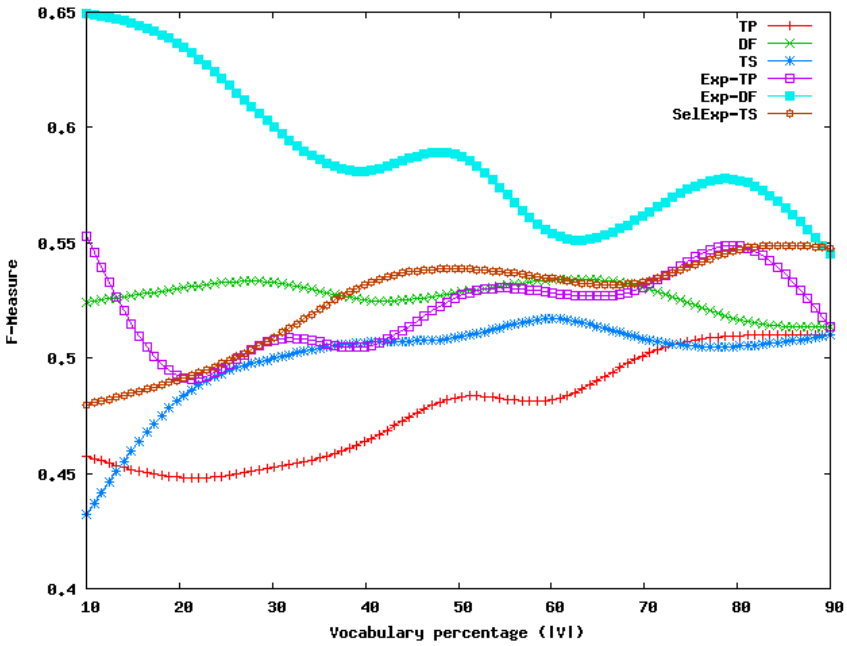
Figure 5.7: Analysis of the self-term expansion methodology by using the K-Star clustering method over the *hep-ex* corpus

be useful when expanding the original terms of the documents. Up to now, we only know that this is true for narrow domain short texts, because its application to other kind of corpora has not been fully investigated.

With respect to the term selection technique 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 Figures 5.6(a), 5.7(b), 5.9(c), and 5.10(c) we may see the comparison of the baseline (unexpanded approach) with the self-term expanded version. In these experiments, we observed that independently of the corpus and clustering method which is used, the DF term selection technique always performed well.

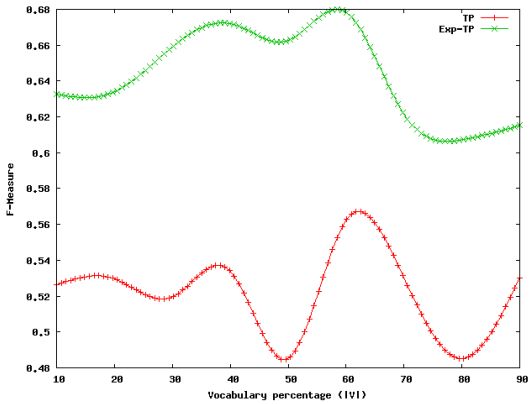


(a) The *CICLing-2002* corpus

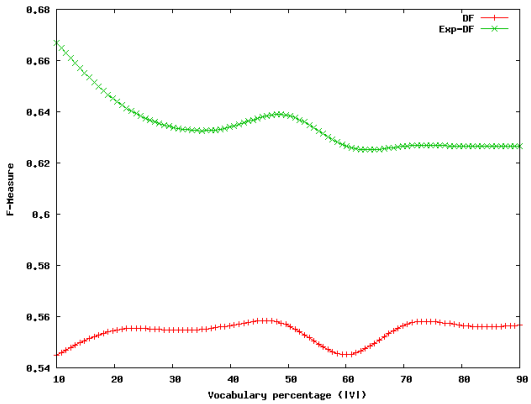


(b) The *hep-ex* corpus

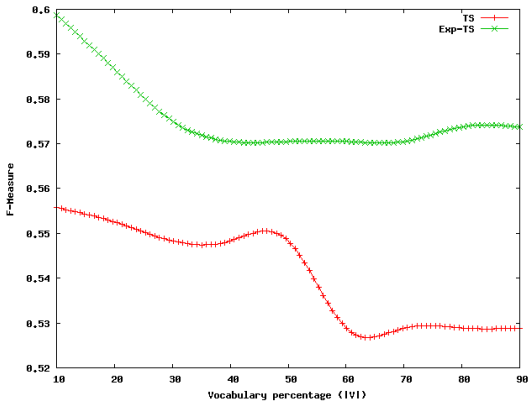
Figure 5.8: Execution of the DK-Means clustering algorithm with the self-term term expansion methodology (all the TSTs)



(a) The TP technique



(b) The DF technique



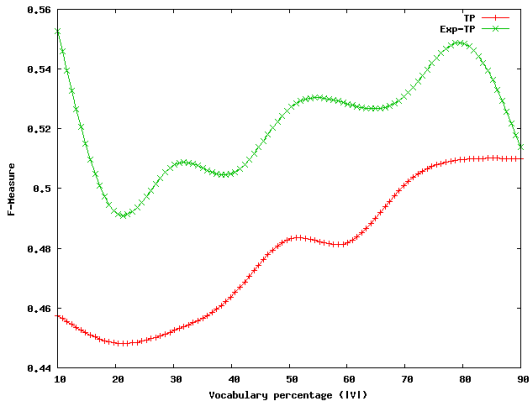
(c) The TS technique

Figure 5.9: 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

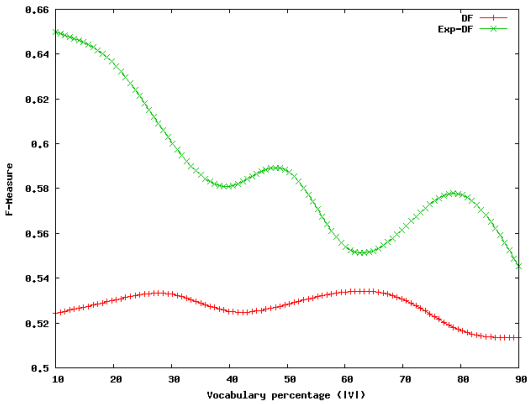
5.5 Concluding remarks

Clustering narrow domain short texts 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 documents of this kind is quite a difficult task.

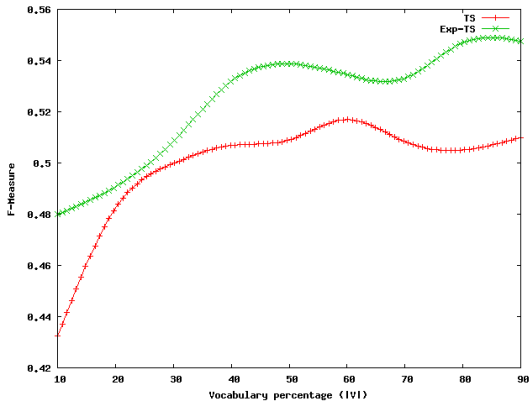
We have introduced a self-term expansion methodology that allows the baseline corpus to be enriched by adding co-related terms from an automatically constructed lexical-knowledge resource obtained from the *same* target dataset (and not from an external resource). This was done by using two different co-occurrence techniques



(a) The TP technique



(b) The DF technique



(c) The TS technique

Figure 5.10: 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

based on bigrams and pointwise mutual information, respectively. The experiments demonstrated 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 then the term selection process. Moreover, the clustering results of the target dataset obtained by just doing the self-term expansion alone are better than those obtained by classical methods of document representation.

The experiments were carried out on two real collections extracted from the CICLing-2000 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 main goal of this study was to boost the performance of clustering narrow-domain short texts by employing the self-term expansion method. This effectively improved the baseline F-Measure by approximately 40%. 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 to the clustering of narrow-domain short texts 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 number of documents is directly proportional to the performance of the proposed methodology.

The application of the method in the specific NLP task of word sense induction is shown in the next chapter.

Chapter 6

Word sense induction

Word Sense Induction (WSI) is a particular narrow domain short-text clustering task of computational linguistics which consists in automatically inducing the correct *sense* for each instance of a given ambiguous word [4].

Such as for Word Sense Disambiguation (WSD) [56], also in WSI the goal is to identify the correct sense of a target ambiguous word in a given sentence but whereas in the case of WSD (a categorization task) a number of possible senses (of a given ontology, e.g. WordNet [39]) of the ambiguous word are given in advance, in WSI we must "discover" the senses without any external knowledge. The use of an ontology such as WordNet is quite typical in WSD. However, those approaches which make use of a general purpose ontology when the corpus domain is specific or narrow, usually fail.

The aim of this chapter is to investigate the behaviour of the self-term expansion methodology (introduced in Chapter 5) in order to discriminate the set of narrow domain (i.e. referring to the same word) short sentences and, therefore, to discover the senses of each word of the WSI dataset.

Due to the magnitude of this collection (composed of 100 sub-collections: each one referring to a different ambiguous word), the characteristics of word sense induction (avoid the use of external knowledge such as an ontology because the sense of a word have to be induce from the text itself), and the importance of this task from a

computational linguistics viewpoint (task 2 of SemEval[1]) we believe that WSI is an important narrow domain short-text benchmark to validate our self-term expansion methodology with.

In comparison to WSI, WSD may be considered a relatively simple task for human beings. However, from an automatic viewpoint could become quite difficult. Let us take, for instance, the sentences presented in Table 6.1, and the different senses associated with the word *bank* (taken from the WordNet 3.0 ontology) which are shown in Table 6.2. We may easily see that the correct sense of the word *bank* in sentences 1, 2, 3 and 4 corresponds to the senses 02343056, 02343252, 09213565 and 01234793, respectively.

Table 6.1: Example of sentences with the ambiguous word *bank*

Num	**Sentence**
1	Arthur Wood is in the *banking* business
2	Arthur Wood acts as the *bank* when we play poker
3	Arthur Wood pulled the canoe up on the *bank*
4	Arthur Wood is in charge of *banking* the fire

The automatic correct association between sentences and senses is a complex task that has been dealt with over a number of years. In fact, word sense disambiguation is one of the oldest problems in computational linguistics, introduced for the first time in [143], given its usefulness in tasks such as machine translation [31].

Typically, the most of the 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 approaches use the ambiguous word context and the information extracted from external knowledge resources (e.g. ontologies such as WordNet) in order to disambiguate the different senses of a word. For instance, in [27] a knowledge-based approach which uses the conceptual density technique is

[1] `http://nlp.cs.swarthmore.edu/semeval/tasks/task02/summary.shtml`

Table 6.2: The WordNet senses for the ambiguous word *bank*

Sense	**POS**	**WordNet gloss**
00169305	noun	a flight maneuver
02787772	noun	a building in which the business of banking transacted
08462066	noun	an arrangement of similar objects in a row or in tiers
09213434	noun	a long ridge or pile
09213565	noun	sloping land (especially the slope beside a body of water)
09213828	noun	a slope in the turn of a road or track
13356402	noun	the dealer or the funds held by a gambling house
13368318	noun	a supply or stock held in reserve for future use
01234793	verb	cover with ashes so to control the rate of burning
01587705	verb	enclose with a bank
02039413	verb	tip laterally
02343056	verb	be in the banking business
02343252	verb	act as the banker in a game or in gambling
02343374	verb	do business with a bank or keep an account at a bank

presented.

A more detailed study of WSD goes beyond the scope of this Ph.D. thesis. For further details see [56] and [3].

The problem of WSI may be described as follows. Given a set of ambiguous words, each one with a set of instances, the aim is to discriminate among all the instances and automatically discover the sense each instance belongs to [107]. Word sense induction implies to perform first the task of word sense discrimination and, thereafter, the induction of senses. Word sense discrimination consists in the clustering of sentences with similar "sense" and, therefore, it may be formally expressed as follows: given a document (sentences) collection $D = \{d_1, d_2, \cdots, d_{|D|}\}$, a clustering of senses of D is a partition into k subsets $\mathcal{C} = \{C_1, C_2, \cdots, C_k | C_i \subseteq D\}$, such that $\bigcup_{i=1}^{k} C_i = D$. Additionally, the sense induction phase is defined as introduced in [82]: for a categorization $\mathcal{C}$ let $W_d = \{w_{d1}, \cdots w_{dn}\}$ denote the word set for a document (sentence) d, and let $W = \bigcup_{d \in D} W_d$ denote the entire word set underlying D, the sense induction

means the construction of a function $\tau : \mathcal{C} \rightarrow 2^W$ that assigns to each element $C \in \mathcal{C}$ a set $W_C \subset W$. As expressed in [82], the following properties are ideal constraints which generally are desired for the labeling function τ: unique, summarizing, expressive, discriminating, contiguous, hierarchically consistent, irredundant. However, these properties may only be approximated in real world tasks.

Usually, the WSI problem is tackled by using clustering algorithms. From the different approaches that exist in literature, the most relevant work with respect to the proposed methodology in this thesis is the one presented in [112] and [126]. In these papers, the authors expanded each term of the target corpus with a set of co-related terms just as we propose to do in the self-term expansion methodology. The difference with respect to our approach is twofold: a) The calculation of the co-ocurrence list in their proposal was done from a training corpus, whereas we *do not* use any external resource; in our particular case the list is calculated from the *same* corpus to be clustered; b) The order of term-selection and term-expansion is done inversely; in our particular case, we have analysed that the best performance is obtained when the term-expansion is carried out before the term selection (see Chapter 5).

The rest of this chapter is structured as follows. The following section outlines the importance of using the *WSI-SemEval* collection in the task of clustering and inducing sense contexts with the help of the self-term expansion methodology. The implemented WSI system is shown in Section 6.2. The obtained results of the experiments carried out are shown in Section 6.3. Particularly, in Section 6.3.1 we will see a comparison between our approach and the one presented in [112] in the international competition of WSI organized by the Association for Computational Linguistics at SemEval. Moreover, in Section 6.3.2 we also aimed to test our self-term expansion approach on the dataset of the equivalent Arabic task of SemEval[2]. Finally, the concluding remarks of this chapter are given in Section 6.4.

[2] `http://nlp.cs.swarthmore.edu/semeval/tasks/task18/description.shtml`

Table 6.3: Assessment values for the *WSI-SemEval* collection

Assessment measure	**Corpus characteristic**	**Value**	**Automatic ranking**	**Manual ranking**
$SLMB$	Domain broadness	195.02	2	3
$ULMB$	Domain broadness	130.62	3	3
SVB	Domain broadness	164.04	5	3
UVB	Domain broadness	153.89	5	3
SEM	Stylometry	0.4477	10	8
DL	Shortness	59.58	2	2
VL	Shortness	50.30	7	2
VDR	Shortness	0.9586	1	2
CI	Class imbalance	0.2263	9	9

6.1 Peculiarities of the *WSI-SemEval* data collection

In Section 2.3.1 we presented general statistics about the *WSI-SemEval* collection. However, further assesment related with its *domain broadness*, *shortness*, *class imbalance*, *stylometry* and *structure* was carried out in Chapter 4. In this section we emphasize the importance of experimenting with this particular set of sub-collections with respect to the aforementioned corpus features.

Table 6.3 shows the average of the values calculated with each of the 100 sub-collections of *WSI-SemEval*. We present both, the manually and automatically obtained ranking for each measure. The ranking values range from 1 to 10. DL and VL indicate that the collection is composed of short texts. Moreover, the VDR measure (not recommended alone for measuring shortness) indicates the highest ratio (between document length and document vocabulary size) that was obtained when analysing all the corpora in Chapter 4. Therefore, we may confirm that the *WSI-SemEval* collection may be categorized as *short-text* collection.

The four formulae that assess the degree of domain broadness agree on the fact

that the corpus is not wide domain. The obtained values are low with respect to the widest domain corpus (*20Newsgroups*) and, therefore, the collection could be considered *narrow domain*, although the degree of narrowness is not exactly defined.

From the ranking given by the SEM measure it is clear that the collection was written with a particular writing style. The term frequency distribution is quite far from the expected Zipfian distribution. Finally, CI may help us to infer that the collection is completely unbalanced.

The self-term expansion methodology is perfectly suitable for this particular task since the expected evaluation corpus may be considered narrow domain and composed of short texts. The last assumption is based on the fact that the fine granularity in WordNet of the senses of a word may imply a relatively high overlapping of vocabulary between the different glosses. This hypothesis may be not be true for all the words evaluated in the task, but in general it holds.

6.2 The proposed word sense induction system

The WSI system we participated at SemEval with is based on the self-term expansion methodology that we described in the previous chapter of this Ph.D. thesis. In Figure 6.1 we may see a diagram of the described self-term expansion process applied to the WSI task. We have used the same format provided for the "Evaluating Word Sense Induction and Discrimination Systems" task of the SemEval 2007 workshop. The ambiguous words are enclosed by two tags: <lexelt> and </lexelt>. Each instance of a given ambiguous word, which contains a unique identifier (*id*), is enclosed by the tags <head> and </head> in its corresponding paragraph. In the figure, the ambiguous word "construct" and a set of instances of it feed the WSI system, which outputs a set of discovered senses.

The developed WSI system is composed of different modules which are illustrated in Figure 6.2. Three are the basic components: the self-term expansion technique, the term selection technique, and the clustering method. The former contains two basic sub-modules: the co-ocurrence list constructor which uses the pointwise mutual information, and the sub-module which expands the terms of the input data. The

latter module employed clusters a reduced version of the expanded corpus, which is downsized by means of a term selection technique.

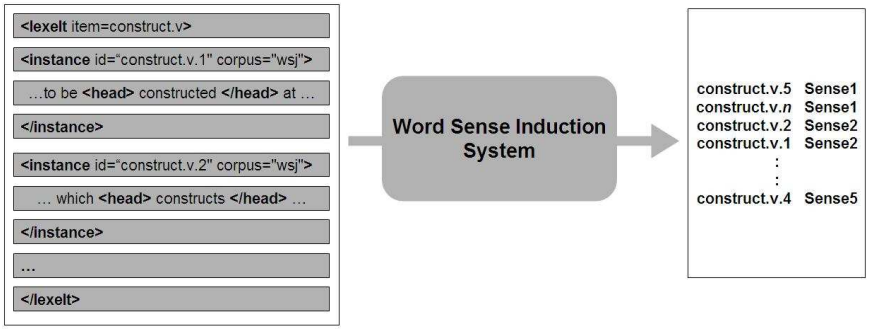


Figure 6.1: The UPV-SI word Sense Induction system

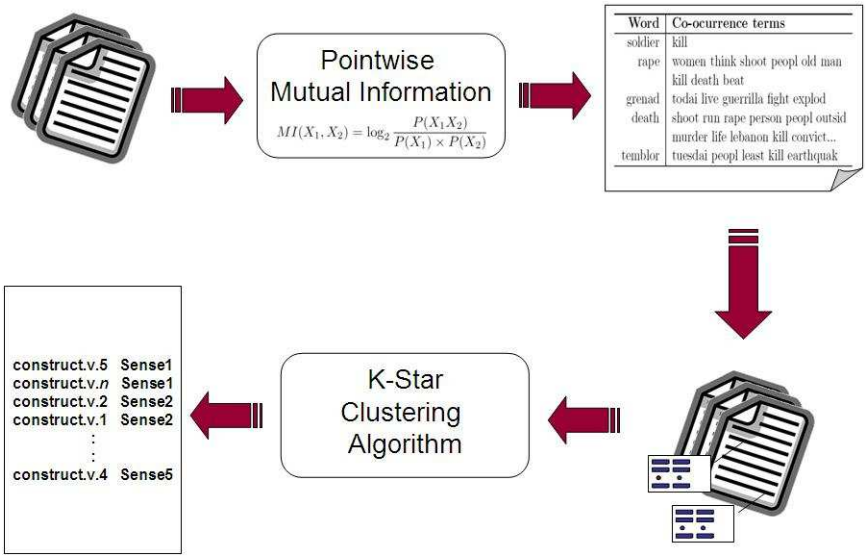


Figure 6.2: The main components of the proposed WSI system

We employed the K-Star unsupervised clustering method for the competition and the further experiments we carried out. We defined the average of similarities among all the sentences for a given ambiguous word as the stop criterion of the clustering process.

Since we know in advance that the corpus of the WSI competition contains short texts (sentences), the expected frequency of the terms is low, and, therefore, we

calculated the input similarity matrix for the clustering method by using the Jaccard coefficient. We have observed in previous experiments that when the frequencies are low there are not significant differences between the performance of other similarity measures with respect to Jaccard.

The Jaccard coefficient between two documents (see Section 2.1.1) gets a normalised value between zero and one. A value of one is obtained when the two given documents contain the same set of terms (words), whereas a value of zero is obtained when no identical words at all are shared by those documents.

6.3 Experimental results

Our interest was to be able to compare the proposed self-term enriching methodology with other approaches in the framework of an international forum such as SemEval, organised by the Association for Computational Linguistics, in order to have a valuable feedback.

We basically were insterested on comparing the proposed methodology with other approaches. Moreover, we consider the methodology to be language-independent and, therefore, we carried out a simple experiment in a completely different language such as the Arabic. In the following sub-section the performance of the WSI system on the two different English and Arabic corpora is described.

6.3.1 Word sense induction in the English language

The results presented in this sub-section are twofold. On the one hand, we show the obtained results in the "Evaluating Word Sense Induction and Discrimination Systems" task of the SemEval 2007 workshop [4, 107]. The evaluation gives an overview of the behaviour of the proposed WSI system with respect to the other systems that participated in the same task. The data collection used in the experiments is made up of one hundred corpora. Each corpus contains sentences related to some ambiguous word. The data collection is composed by 100 ambiguous words distributed in 35 nouns and 65 verbs (see Section 2.3). We pre-processed this set of 100 corpora by

eliminating stopwords and, thereafter, applying the Porter stemmer.

On the other hand, we present a further analysis of our WSI system with three term selection techniques. Moreover, we investigate the use of an another self-term expansion technique which takes advantage of the a priori known morphosyntactical variations of each ambiguous word in the evaluated corpus.

Since the collection used in the experiments is composed of 100 corpora, we show the average in all the results presented in this chapter. However, we have included one appendix with each corpus analysis for the case when the self-term expansion technique is applied (see Appendix C).

Most of the experiments we have carried out made use of the self-term expansion method described in Chapter 5. That is, we replaced each term of the target corpus with a set of co-related terms calculated by using the pointwise mutual information. We determined that the single ocurrence of each term should be at least three, as suggested in [80], whereas the threshold for the co-ocurrence formula should be seven. We employed 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. Since the competition's organizers only accepted one run per team, we decided to submit the results over the expanded version of the corpus, i.e., we did not apply any term selection to the self-expanded data collection.

In order to fully appreciate the self-term expansion technique, in Table 6.4 we show the co-ocurrence list for some words related with the verb "kill", which were obtained from the English language corpus used in the WSI task of the SemEval 2007 workshop. The pointwise mutual information is calculated after preprocessing the corpus and, therefore, we only present the stemmed version of the terms.

The task organizers decided to use two different measures for evaluating the runs submitted to the task. Since the proposed measures gave conflicting information [4], we decided to report both of them. The first measure is called unsupervised, and it is based on the Fscore measure (F-Measure), whereas the second measure is called supervised recall. For further information about how these measures were calculated

Table 6.4: An example of co-ocurrence terms

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

you can refer to [1, 2] or to see Section 2.1.5 of this document.

In Table 6.5 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 are also included. The upper baseline (Baseline1) assumes that each ambiguous word has only one sense, whereas the lower baseline (Baseline2) is a random assignation of senses. We are ranked as third place and our results are second only to those of the best team, due to the fact that all the teams obtained lower values than the upper baseline (Baseline1). This aspect highlights the difficulty of the WSI clustering task. However, given the similar values with the "Baseline1", we may assume that probably the best team presented one cluster per ambiguous word as the Baseline1 did, whereas our UPV-SI system obtained instead 9.03 senses per ambiguous word on average.

In Table 6.6 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 of the two baseline are also presented (the other baseline obtained the same Fscore). In this case, the baseline approach tagged each test instance with the most frequent sense obtained from a training split. We are ranked again in the third place and our score is slightly above the baseline.

Table 6.5: Unsupervised evaluation (F-Measure performance)

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 6.6: Supervised evaluation

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

Analysis of the self-term expansion methodology

We consider that the evaluation of the proposed methodology in the "Evaluating Word Sense Induction and Discrimination Systems" of the SemEval 2007 workshop is far to be complete. We submitted a run which clustered a self-term expanded version of the original corpus. All the terms were enriched with their co-related terms. However, it would be interesting to know the behaviour of the term selection techniques we referred in the previous chapters of this thesis. Moreover, since we know in advance the ambiguous words (they are tagged in the test corpus), we would like to investigate the behaviour of the self-term expansion methodology when expanding only the ambiguous words and its corresponding morphosyntactical variations.

Therefore, we present three different ways of clustering the WSI-SemEval corpus. The first approach clusters every subset of the original corpus obtained by mean

of a term selection technique (vocabulary percentage from 20% to 90%). We have named this approach as NETS (No-Expanded with Term Selection). The other two approaches are named JAWETS and AETS, for Just Ambiguous Words Expanded with Term Selection and All words Expanded with Term Selection. In Figure 6.3 we may see the NETS approach with three different term selection techniques: Transition Point, Document Frequency and Term Strength. For the analysis presented in this section we will assume the values obtained employing just the TSTs as the baseline for the self-term expansion methodology.

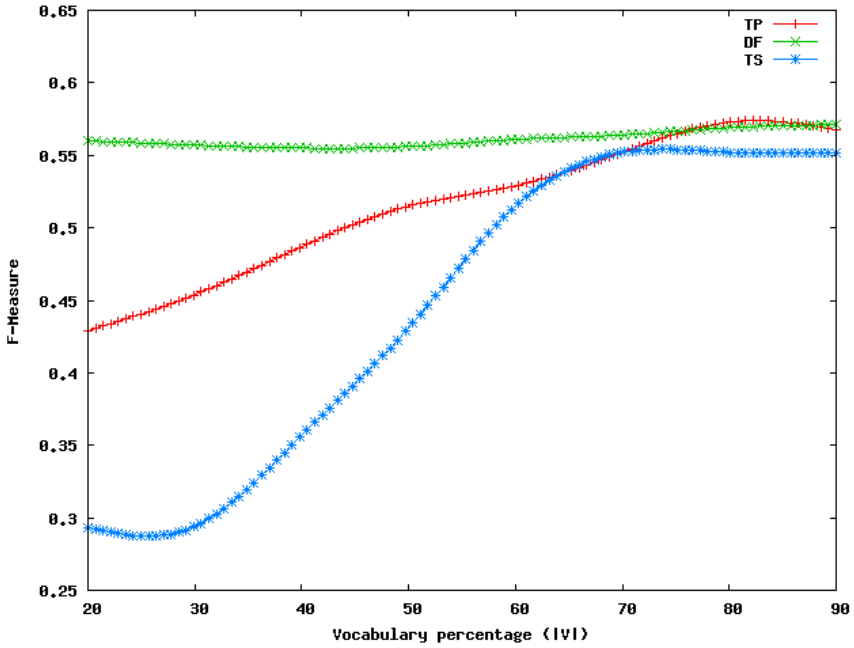


Figure 6.3: Behaviour of the term selection techniques over the WSI-Semeval corpus

In Figure 6.4 we may see the behaviour of the methodology when using the DF term selection technique. The figure shows the JAWETS and AETS approaches. Both versions obtained a high improvement over the unexpanded version. The plotting shows that the use of some kind of a priori knowledge could be very useful. JAWETS obtained comparable results with respect to the AETS approach in almost all the subsets of the corpus. Moreover, the time required JAWETS for the expansion, term selection and clustering was significantly smaller than the one for the AETS approach. Unfortunately, the JAWETS approach may only be used in clustering tasks where

the most significant words are known, such as in this particular case the ambiguous words are.

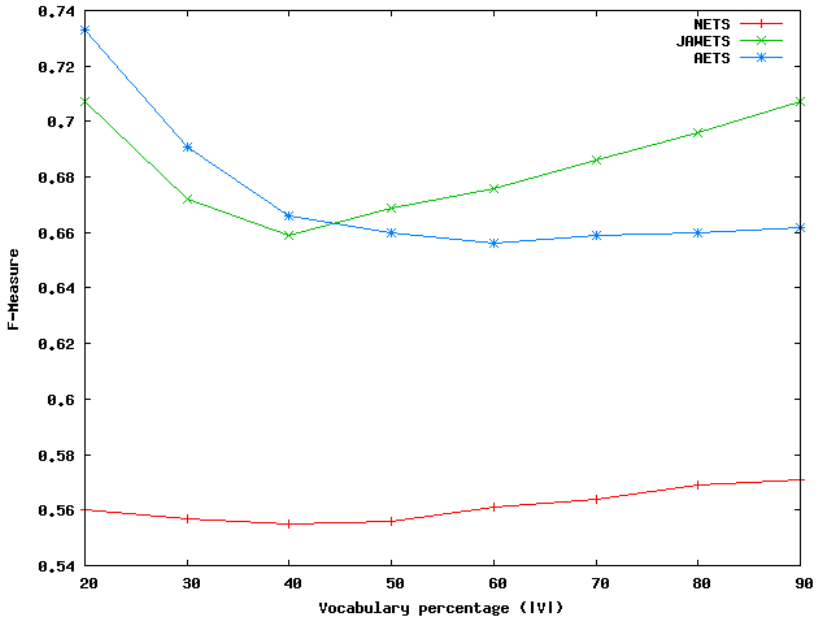


Figure 6.4: Behaviour of the DF term selection technique with three different approaches: NETS, JAWETS and AETS

Figures 6.5 and 6.6 show the behaviour of the TP and the TS term selection techniques, respectively. In the two cases, the AETS approach outperformed both, the unexpanded version and the JAWETS approach.

The fact that we still are below the baseline indicates the difficulty of improving this threshold in an unsupervised way. The obtained results are an indicative of the effect of the class imbalance problem for this particular text collection. However, even if the most of the corpora that compose the *WSI-SemEval* collection are unbalanced, we may confirm again that the self-term expansion method highly improves the results in comparison to when no expansion is done.

6.3.2 Word sense induction in the Arabic language

The self-term expansion approach already tested on the English-written dataset [107] was executed on another corpus written on Arabic language. The aim was to

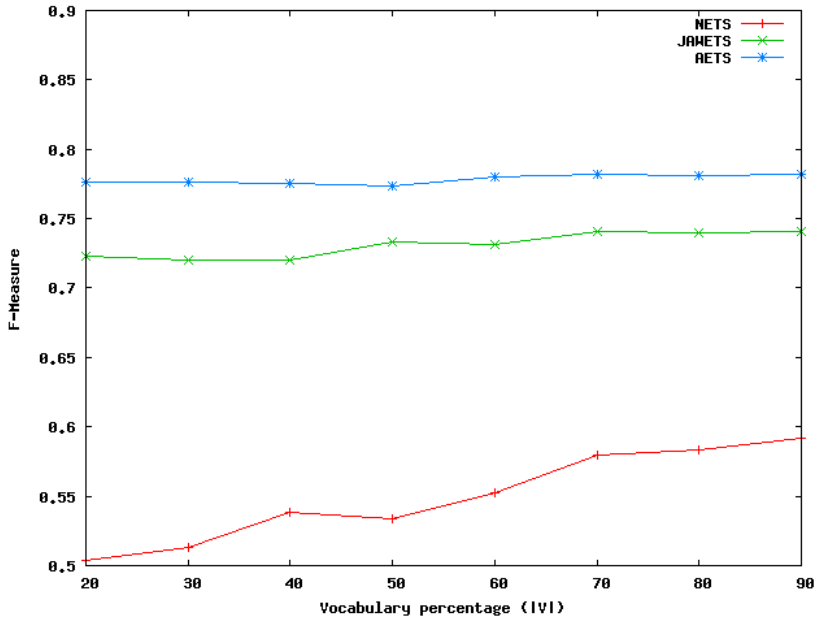


Figure 6.5: Behaviour of the TP term selection technique with three different approaches: NETS, JAWETS and AETS

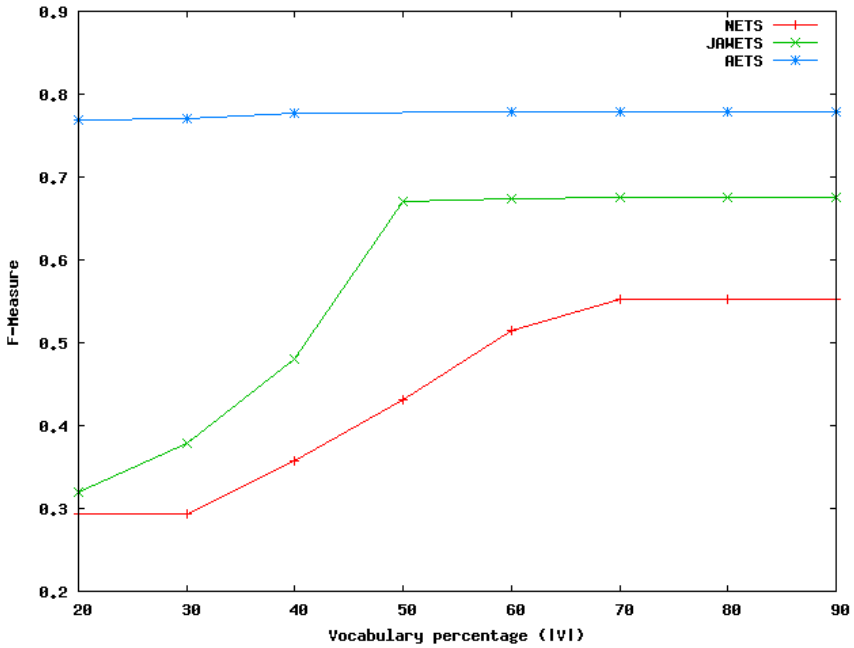


Figure 6.6: Behaviour of the TS term selection technique with three different approaches: NETS, JAWETS and AETS

analyse the behaviour of the aforementioned methodology in a completely different language.

For the experiments carried out in this research work, we have used the dataset prepared for the Arabic task of the SemEval workshop. A set of 509 ambiguous words (379 nouns and 130 verbs) were provided. We preprocessed this original dataset by eliminating punctuation symbols and Arabic stopwords. The experiments were carried out by using a tokenized (with segmentation) version of the target corpus. That is, we found the root morpheme and the affixes of each word separated. The complete characteristics of the used corpus are described in Table 6.7.

Table 6.7: Characteristics of the Arabic corpus used in the WSI experiment

Characteristic	**Value**
Size	343 Kbytes
Ambiguous words	509
Nouns	379
Verbs	130
Instances (in average)	1,025

In order to determine the efficacy of the approach presented [106] we have carried out a manual evaluation based on the judgment of a native Arabic speaker, since it was impossible to calculate the precision and recall of our approach because the gold standard file has not been released.

The examples showed in this section are given first in Buckwalter transliterated characters. The Buckwalter transliteration was developed by Tim Buckwalter for practical storage, display and email transmission of Arabic text in environments where the display of genuine Arabic characters is not possible or convenient [24]. The Arabic and the translated English sentences follow the Buckwalter transliterated ones.

Some examples show that the approach performed quite well on Arabic data. For instance, the Arabic word "كل" ("kl" in the Buckwalter transliterated characters) in the data has two different meanings: the first one is "all" and the second is "every". The obtained sense discrimination for this particular ambiguous word is shown as

follows.

Sense "all (kl)" with Buckwater transliteration:

1) w >wDHt An " AltEAwn AlHAly byn Albldyn ysyr b xTY Hvyvp w fy kl AlmjAlAt " .

2) w qAl AlsA}H AlbryTAny jwrj dwd , m$yrAF AlY Alfndq AlAnyq AlmTl ElY AlbHr : " nElm >n hm yqymwn fy h*A Alfndq w lA ymknn ny Alqwl <n nA n$Er b {rtyAH mE wjwd hm w kl h*A AlEdd mn rjAl Al$rTp Hwl nA " .

3) w qAl AlsA}H AlAlmAny bwl hwfmAn , mtHdvAF En AlHrAs Almtmrkzyn ElY sTH Alfndq : " AEtqd An mn Algryb AHATp AlmkAn b kl AjrA'At AlAmn h*h .

4) HtY lA ysA' tfsyr EbArp " rfD AlEnf b kl >$kAl h " AlwArdp fy byAn ...

Sense "all (كل)" in Arabic language:

واوضحت ان التعاون الحالي بين البلدين يسير بخطى حثيثة وفي كل المجالات
وقال السائح البريطاني جورج داود مشيرا الى الفندق الانيق المطل على البحر
نعلم انهم يقيمون في هذا الفندق ولا يمكنني القول اننا نشعر بارتياح مع وجودهم
وكل هذا العدد من رجال الشرطة حولنا
وقال السائح الالماني بول هوفمان متحدثا عن الحراس المتمركزين على سطح الفندق
اعتقد ان من الغريب احاطة المكان بكل اجراءات الامن هذه
حتى لايساء تفسير عبارة رفض العنف بكل اشكاله الواردة في بيان

Sense "all (كل)" translated into the English language:

1. and it is clear that the collaboration between the two countries is going slowly and in all domains
2. and a British tourist George David pointing out the elegant hotel with the sea view, "we know that they live in this hotel and I cannot say that we feel confortable with their presence, and all this policeman around us"

3. and the German tourist Paul Hoffman talking about the guardians on the hotels attic "I think that it is extrange to surround the place with all these security measures"
4. in order to not misunderstand the expression "refusal of violence with all its forms" seen in the report

Sense "every (kl)" with Buckwater transliteration:

```
1) -LRB- . . . -RRB- f  Alkl  hnA tHt AlmrAqbp  h*h AlAyAm  " .

2) w zAr  Aldwry  klA  mn AlArdn w lbnAn w swryA fy ATAr
jwlp  tsthdf H$d  AlmEArDp  AlErbyp l >y  Hmlp  qd t$n  hA
AlwlAyAt  AlmtHdp ElY AlErAq .
```

Sense "every (كل)" in Arabic language:

فالكل هنا تحت المراقبة هذه الايام
وزار الدوري كلا من الاردن ولبنان وسورية في اطار جولة تستهدف حشد المعارضة
العربية لاي حملة قد تشنها الولايات المتحدة الامركية

Sense "every (كل)" translated into the English language:

1. so every one of them here is being monitored these days.
2. and Al-Dory visited every one of the following countries, Jordan, Lebanon and Syria aiming at encouraging the Arabic opposition against any campaign which the US might lead

We noticed that sometimes our method tends to discriminate several senses of a word even if all the instances of the word mean the same. In the following example, for instance, all the samples of the Arabic word "جندي" (soldier - jndy) have the same sense. However, our method discriminates the first instance as having a different sense mainly because it appears in a quite different context.

Word "soldier (jndy)" with Buckwater transliteration:

1) w >fAd ms&wlwn hnwd An jndyyn lqyA Htf hmA w >n vlAvp |xryn jrHwA xlAl ATlAq nAr fy mstwdE l *xyrp Aljy$ fy wlAyp jAmw w k$myr Alhndyp .

2) w qAl DAbT $rTp An vwArA dhmwA AlmstwdE fy mqATEp bwn$ ElY msAfp 250 kylwmtrA $mAl jAmw AlEASmp Al$twyp l AlwlAyp w >lqwA AlnAr f qtl jndyAn fwrAF w >Syb vlAvp qbl frAr AlmqAtlyn .

3) w nfY ms&wl |xr fy AlHkwmp HSwl hjwm l Almt$ddyn w qAl An jndyAF hw Al*y >lq AlnAr ElY zmlA} h .

Word "soldier (جندي)" in Arabic language:

واقاد مسؤولون هنود ان جنديين لقيا حتفهما وان ثلاثة اخرين جرحوا خلال اطلاق نار في مستودع لدخيرة الجيش في ولاية جامو وكشمير الهندية
وقال ضابط شرطة ان ثوارا دهموا المستودع في مقاطعة بونش على مسافة 250 كيلومترا شمال جامو العاصمة الشتوية للولاية والقوا النار فقتل جنديان فورا واصيب ثلاثة قبل فرار المقاتلين
ونفى مسؤول اخر في الحكومة حصول هجوم للمتشددين وقال ان جنديا هو الذي القى النار على زملائه

Word "soldier (جندي)" translated into the English language:

1. and some indian responsables have declared that two soldiers died and three others were wounded during the shooting which happen in the Jammu and Cachemira army repository
2. and a police officer said that some rebelds attacked the repository in province of Bunsh which is located to 150 kilometers north of the Jammu capital shooted and killed two soldiers inmediately and three were wounded before the murders escaped
3. and another responsible of the goverment denied any atack of extremists and declared that a soldier open fired on his colleages

After analysing our results, we have observed that the more instances we have for a given ambiguous word (and, therefore more contextual information), the better our

method discriminates the different senses of the given instances. In Figure 6.7 we may see the example of the noun "President" used in four different sentences. In the first sentence the word president is used to mean the "Prime Minister" which is said in Arabic "Ministers President", whereas in the other three sentences "president" mean "head of nation or country". In this case the approach that we have used succeeded in discriminating the two mentioned senses.

و قال زعيم آخر في المؤتمر هو عمر فاروق , ان الكرة الان في الملعب الهندي , و ان على **رئيس** الوزراء الهندي أتال بيهاري فاجباي الان ان يرد على مبادرة
And another leader of the conference was Umar Farooq, he said that the ball is now in India's stadium and that the Indian **Prime Minister** Atal Bihari Vajpayee now had "to respond to the initiative

و رحبت " جبهة تحرير جامو و كشمير " الانفصالية ب خطاب **الرئيس** الباكستاني و خصوصا التأكيد مجددا ل الدعم السياسي و المعنوي و الديبلوماسي لالكشميريين
And Liberation Front "the Jammu and Kashmir separatist" welcomed the Pakistani **President** speech and especially the reiteration to support for the political, moral and diplomatic the Koshemurien

واصلت نيودلهي الضغط على اسلام اباد لقمع الجماعات الاسلامية المتشددة , فأعلن وزير الدفاع الهندي جورج فرناندز ان القوات الهندية المحتشدة على الحدود مع باكستان لن تسحب اذا لم يترجم **الرئيس** الباكستاني برويز مشرف تعهداته كيح جماح الاسلاميين الى افعال
New Delhi continues to put pressure on Islamabad to suppress extremist Islamic groups, and the Indian Defense Minister George Fernandes announced that Indian forces mobilized on the borders with Pakistan would only withdraw if the Pakistani **President** Pervez Musharraf will curb the Islamists as he promised.

و اكد نائب **الرئيس** العراقي طه ياسين رمضان الاحد الماضي ان بلاده لن تسمح بعودة مفتشي الاسلحة
And Iraq's Vice **President** Taha Yassin Ramadan said last Sunday that the country will not allow the return of weapons inspectors

Figure 6.7: Samples of the noun "President"

To sum up, we can say that when we have poor context information (few sentences using the ambiguous word) our method is not able to discriminate the different senses of the word and tends to cluster all the senses in the same group. For instance, the use of the verb "to see" in the first sentence of the example given in Figure 6.8 means the opinion of somebody about something, and in the second one the verb "to see" is used to mean what a man sees with his eyes. However, there are very few sentences which contain the verb "to see" in the Arabic SemEval corpus and, therefore, our system did not obtain enough information about the contexts in which this verb could possibly appear.

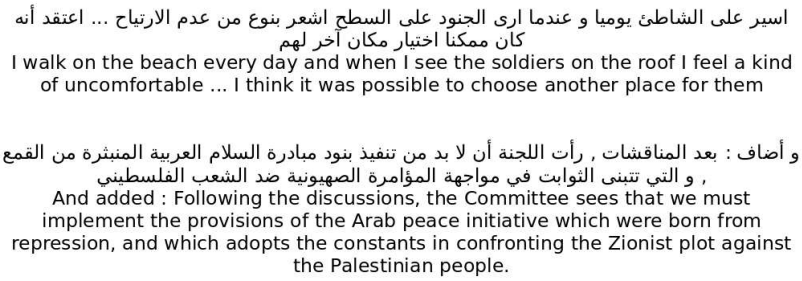


Figure 6.8: Samples of the verb "to see"

See [106] for further details about the experiments we carried out on the Arabic data set of the SemEval task.

6.4 Concluding remarks

The self-term expansion methodology is explicitly designed for narrow domain short-text corpora. It was applied to the word sense induction task which consists of distinguishing sentences with an ambiguous word from other sentences that have the same ambiguous word but with a different meaning. The results with a corpus written in English showed that the technique employed obtained a better performance than the baseline, especially a baseline which had chosen the most frequent meaning. In fact, we outperformed every other unsupervised approach. Having obtained third place in the rankings at the SemEval competition [107] highlights how valuable this simple technique can be in the clustering process.

We confirmed that the self-term expansion technique improves the clustering of the unexpanded corpus no matter which term selection technique is used when enriching the corpus subsets. Moreover, we observed that when some kind of crucial information is known a priori, such as ambiguous words, the method may even improve on the results simply by expanding only the most important terms of the corpus instead of each one of them.

The evaluation with the WSI-SemEval corpus of the "Evaluating Word Sense Induction and Discrimination Systems" task of the SemEval 2007 workshop showed that

expanding only the ambiguous terms is the best approach for word sense induction.

We also studied the language-independent characteristic of the self-term expansion methodology for the word sense induction/discrimination task. A set of preliminary experiments also showed good performance in the Arabic language. The tokenization performed on the Arabic corpus of the SemEval workshop by the task organisers was only partial since they kept, for instance, the Arabic definite article "Al" joined to the words. Even though this partial tokenization might be positive for other natural language processing tasks, we consider that the method presented in this research work would have performed better if the tokenization used had taken into consideration the definite article.

We consider that the evaluation of the proposed methodology on a real task, which was performed in an international forum, has been truly positive and measures its performance fairly. The evaluation has also provided us the opportunity to detect points for improvement. Our aim is to study the behavior of the self-term expansion methodology in other application areas.

Chapter 7

Evaluation of clustering validity measures in short-text corpora

Text clustering consists in the assignment of documents to unknown categories. This task is more difficult than supervised text categorization [127, 87] because the information about categories and correctly categorized documents is not provided in advance. An important consequence of this lack of information is that clustering results cannot be evaluated with typical external measures like F-Measure (see Section 2.1.5) and, therefore, the quality of the obtained groups is evaluated with respect to *structural properties* or *internal measures*. Classical internal measures used as cluster validity measures include the *Dunn* and *Davies-Bouldin* indexes, new graph-based measures like *Expected Density Measure* (EDM) and Λ-*Measure* as well as some measures based on the corpus vocabulary overlapping (see Section 4.5 and 4.1.4 for a description of the aforementioned clustering validity measures).

When clustering techniques are applied to collections containing *very short* documents, additional difficulties are introduced given the low frequencies of the document terms. Research work on "short-text clustering" is particularly relevant, if we consider the current and future trend of the use of 'small-language', e.g. blogs, text-messaging, snippets, etc. Potential applications in different areas of natural language processing may include re-ranking of snippets in information retrieval, and automatic clustering of scientific texts available on the Web [102, 100].

In order to obtain a better understanding of the complexity of clustering short-text corpora, a deeper analysis of the main factors that have a direct impact on the obtained results is required. Specifically, we are interested in studying whether or not the internal clustering validity measures are good estimators of the usability of the results from a user viewpoint. For this reason, several short-text corpora are considered. Since the information about the correct categories of the documents is available, then the quality of the clustering results evaluated accordingly to the internal measures may be compared with external ones, such as with F-Measure.

Our study is closely related to the work presented in [132] where different internal cluster validity measures are used to predict the quality of clustering results in experiments with samples of the RCV1 Reuters collection [120]. The predicted quality in this case is compared with the real quality expressed by the F-measure values obtained from a manual categorization. In our case, we study *very short-text corpora*. The aim of the presented research work (published in [57]) was to determine the correlation degree between internal and external clustering validity measures.

In the following section we explain how the experiments were carried out and we show the results obtained. Thereafter, we draw some conclusions and we discuss the possible future work with respect to the topic of this chapter.

7.1 Correlation between internal and external clustering validity measures

The aim of this research work was to investigate the possible correlation between the external F-Measure and some Internal Clustering Validity Measures (ICVM). Cluster validity may be seen as a measure of goodness for the results obtained by clustering algorithms. There exist two types of cluster validy measures: external and internal [82]. The difference relies, respectively, on the use or not of a pre-specified structure of the data which is imposed usually by an expert, such as a corpus gold standard.

In the experiments we are presenting in this chapter, we made use of five different

internal clustering validity measures, namely *Dunn* index [38], *Davies-Bouldin* index [34], *Expected Density Measure* [132], Λ-*Measure* [132], and macro-average Relative Hardness (MRH) [105]. Other internal validity measures (such as the Silhouette coefficient [122], correlation, cophenetic distance [95], Neill's conditional entropy [93] and Newman's Q-Measure [92]) could have been explored. For instance, relative closeness, relative interconnectivity were introduced in [64] in the framework of dynamic modeling for hierarchical clustering. However, we consider that the analysis of all of them would be out of the scope of this chapter.

7.1.1 Datasets and subcorpora generation

For the experiments of this chapter, we used the following short-text corpora: *CICLing-2002* and *WSI-SemEval* (*narrow domain*), and *R8-Reuters* (*wide domain*). As mentioned before, the WSI data collection is made up by 100 corpora and, therefore, the correlation between the ICVMs and the F-Measure may be quite confidently determined, since we may assume that all the corpora keep the same or at least similar intrinsic characteristics such as domain broadness, shortness, class imbalance and stylometry. However, when we used the other two corpora we proposed to generate subsets of them based on the categories given in the gold standard. This was made with the purpose of analysing the correlation between the investigated internal clustering validity measures and the F-Measure under the same possible circumstances, i.e., with similar underlying corpus characteristics. For this purpose, we have generated subsets for the *CICLing-2002* and the *R8-Reuters* corpora in the following manner. We considered all the possible combinations equal or more than two categories of each corpus and for each of them we calculated its ICVM value. Therefore, for a corpus of n categories, a number of $2^n - (n + 1)$ possible subcorpora were obtained. Thus, for the *R8-Reuters* corpus (eight categories) we obtained 247 subsets, whereas for the *CICLing-2002* corpus (four categories) we obtained 11 subsets.

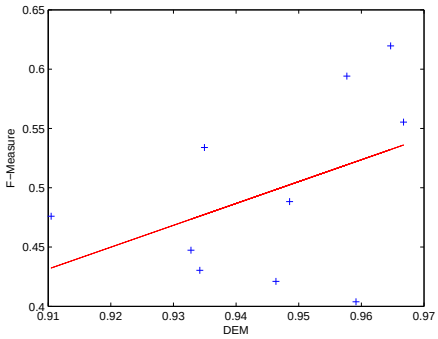
7.1.2 Experimental results

The results of the experiments we carried out were plotted showing the F-Measure as a function of the corresponding internal clustering validity measure, where both measures were evaluated with the clusters obtained by the K-Star clustering method [130]. In particular, Figures 7.1, 7.2, 7.3 and 7.4 show the obtained correlation results for each corpus and the EDM, Λ-Measure, Davies-Bouldin and Dunn clustering validity measures were considered. The x-axis corresponds to the different ICVM, whereas the y-axis corresponds to the F-Measure. In order to easily visualise the correlation between all the ICVM and F-Measure, we plotted the polynomial approximation of degree one. A desirable correlation would show dots that start in the left corner (low values of F-Measure) and grows monotonically (high values of F-Measure). In this sense, for a better readability we changed the sign of the Davies-Bouldin index, which is the only measure to be minimised. Therefore, in this way the results are directly comparable. This modification was not done in Figures 7.5, 7.6 and 7.7, where we present the obtained results of two introduced internal clustering validity measures (MRH-J and MRH-C), since we wanted to emphasize the specific behaviour of these new measures in the framework of validating the clustering of short-text corpora.

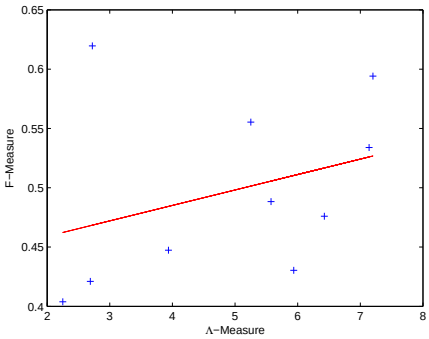
We observed that EDM is the only measure analysed that keeps the expected direct correlation in all the corpora. This behaviour suggests a certain robustness of this measure. Specifically, when it is evaluated on the *WSI-SemEval* corpora, it appears to have a lineal correlation with the F-Measure.

The Λ-Measure obtains an "acceptable" correlation with the *CICLing-2002* and R8 corpora. However, it is remarkable that the correlation obtained with the *WSI-SemEval* corpus is inverse. It seems that this ICVM is not adequate in general for short texts. One important finding is that if a clustering algorithm is designed in a way that attempts to optimise the Λ-Measure, then it will be negatively affected when using short-text corpora.

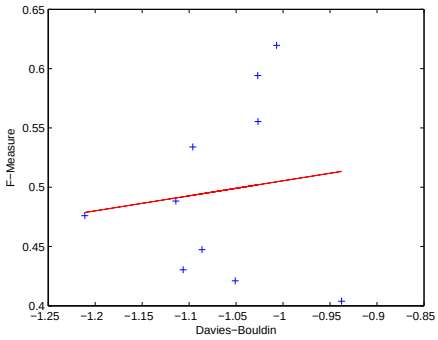
The Davies-Bouldin index correlates very well with the F-Measure in the *WSI-SemEval* collection, acceptably in the *CICLing-2002* corpus and quite badly in the R8 dataset.



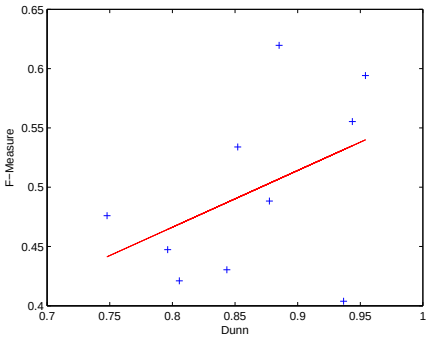
(a) EDM vs F-Measure



(b) Λ-Measure vs F-Measure

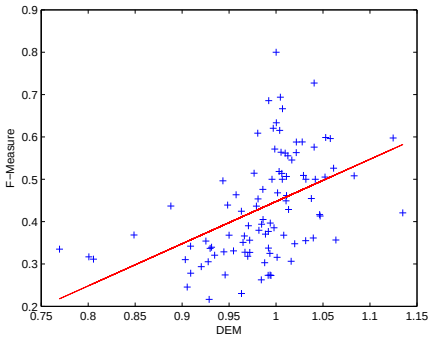


(c) Davies-Bouldin vs F-Measure

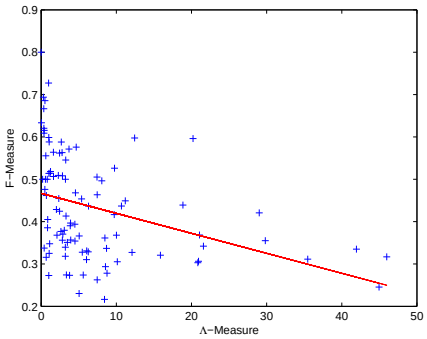


(d) Dunn Index vs F-Measure

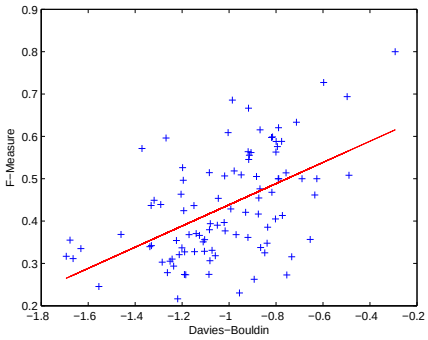
Figure 7.1: Correlation of validity measures for the CICLing-2002 corpus



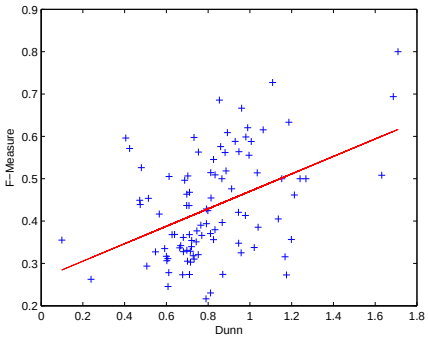
(a) EDM vs F-Measure



(b) Λ-Measure vs F-Measure

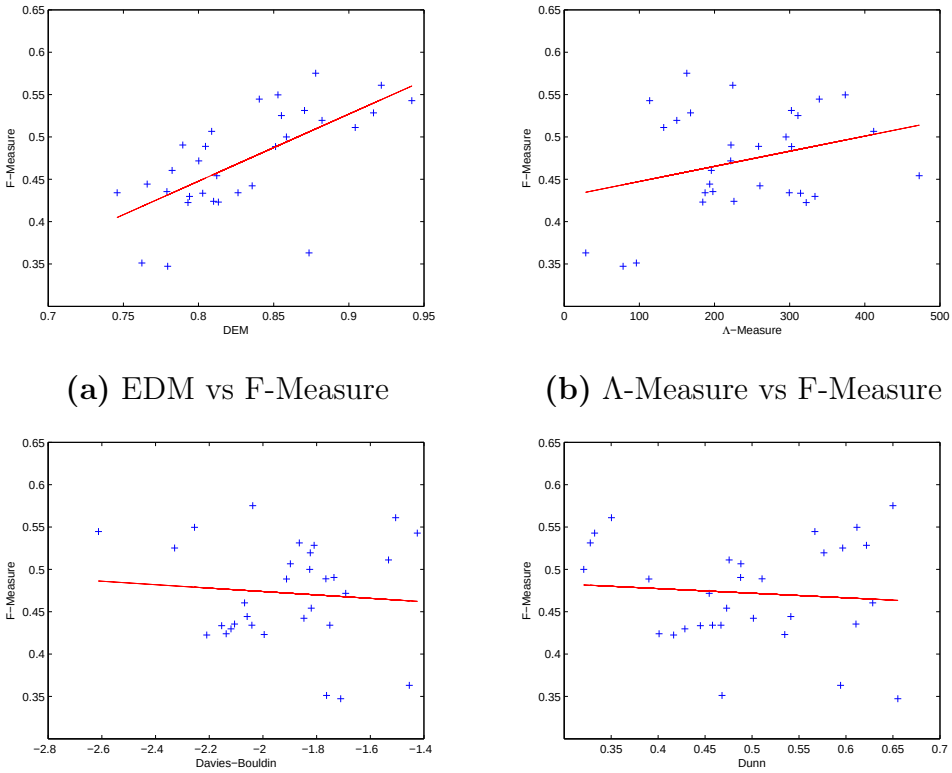


(c) Davies-Bouldin vs F-Measure



(d) Dunn Index vs F-Measure

Figure 7.2: Correlation of validity measures for the *WSI-SemEval* collection



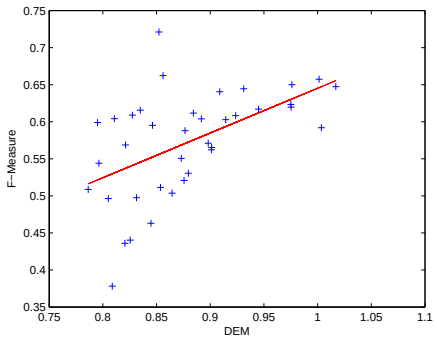
(a) EDM vs F-Measure

(b) Λ-Measure vs F-Measure

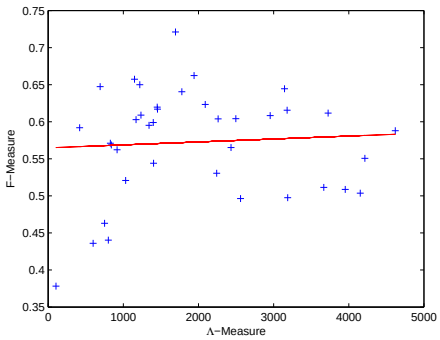
(c) Davies-Bouldin vs F-Measure

(d) Dunn Index vs F-Measure

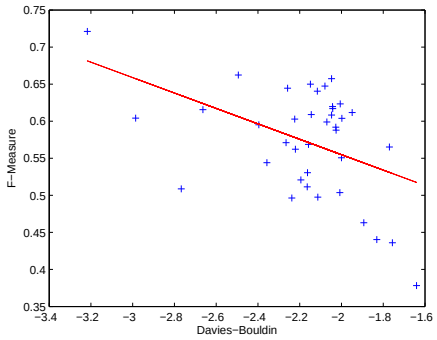
Figure 7.3: Correlation of validity measures for the R8 test corpus



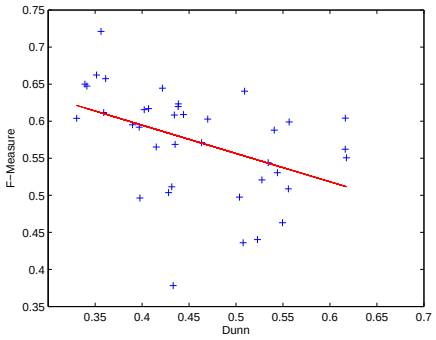
(a) EDM vs F-Measure



(b) Λ-Measure vs F-Measure

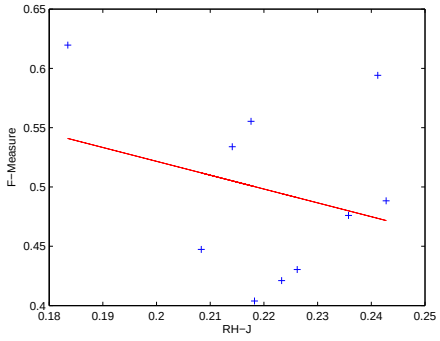


(c) Davies-Bouldin vs F-Measure

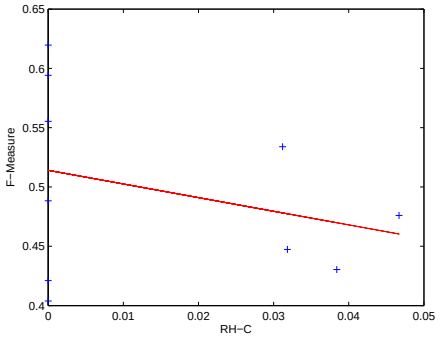


(d) Dunn Index vs F-Measure

Figure 7.4: Correlation of validity measures for the R8 train corpus

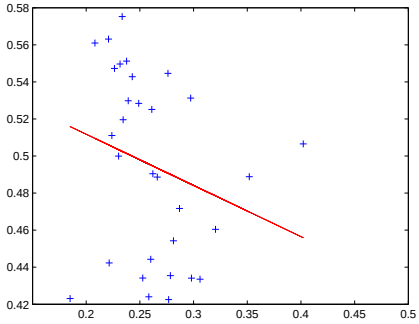


(a) MRH-J

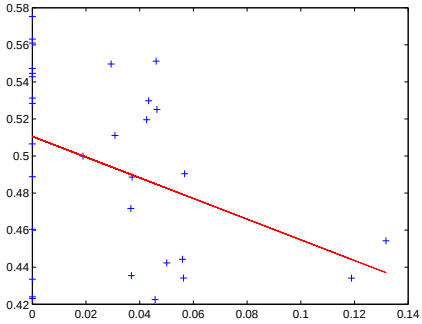


(b) MRH-C

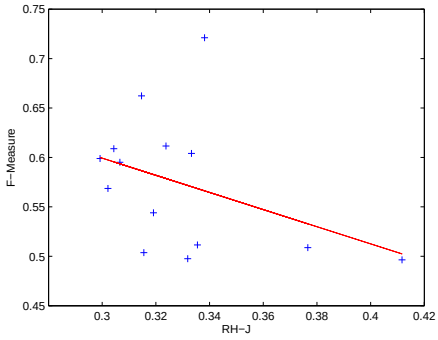
Figure 7.5: Evaluation of the CICLing-2002 corpus with the MRH formulae based on the Jaccard coefficient and the cosine measure



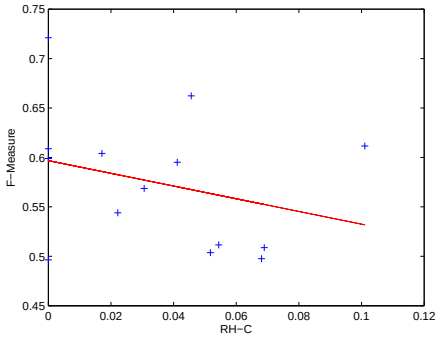
(a) MRH-J with the R8 test corpus



(b) MRH-C with the R8 test corpus

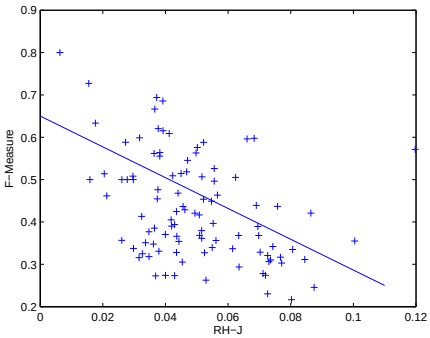


(a) MRH-J with the R8 train corpus

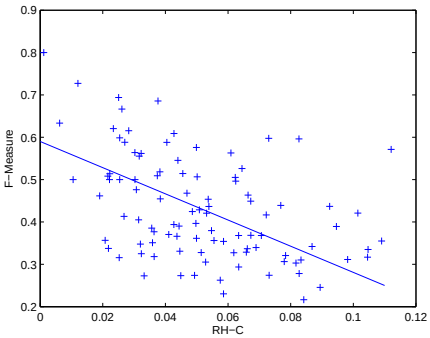


(b) MRH-C with the R8 train corpus

Figure 7.6: Evaluation of the R8 test and train corpora with the MRH formulae based on the Jaccard coefficient and the cosine measure



(a) MRH-J



(b) MRH-C

Figure 7.7: Evaluation of the *WSI-SemEval* collection with the MRH formulae based on the Jaccard coefficient and the cosine measure

Finally, the Dunn measure behaves well with both, the *CICLing-2002* and *WSI-SemEval* corpora, but it did not obtain a good correlation in the R8 dataset. We observed that the Davies-Bouldin and the Dunn indices have obtained similar results. With respect to the relative hardness (both similarity measures based on the Jaccard and the cosine similarity), it obtained good results in all the corpora.

From a corpora viewpoint, we may see that in the *CICLing-2002* corpus all the ICVM measures showed a good behaviour. In R8 all the results were consistent when evaluated in the test and train versions of this corpus; EDM, Λ-Measure and MRH correlated very well with F-Measure, but Davies-Bouldin and Dunn obtained an inverse correlation. In *WSI-SemEval* we obtained very good results for almost all ICVMs (except Λ-Measure). The reader should pay attention that this collection consists of 100 corpora and, therefore, it makes sense to have obtained more stable results.

7.2 The relative hardness of clustering corpora

Reuters (21578, 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 or not the current clustering corpora are easy clustering instances.

We are interested in investigating two aspects: a set of possible features hypothetically related with the hardness of the clustering task, and 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 texts to be clusterd,

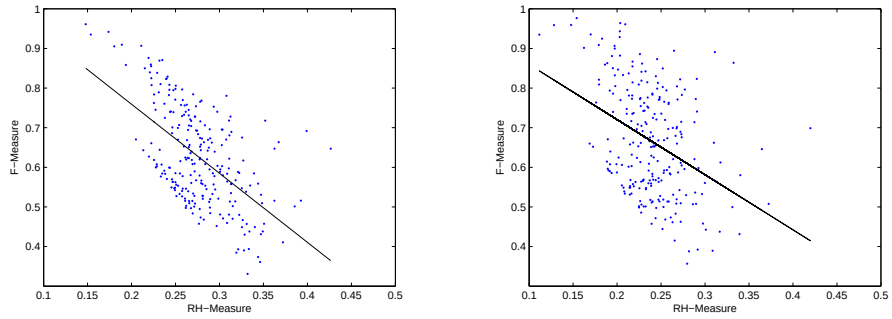
2. the broadness of the corpora domain, and

3. whether the documents are single or multi categorized.

In this study we analyse the impact of the domain broadness in the relative hardness of clustering. The measure relies on the calculus of the vocabulary overlapping. These preliminary experiments were carried out by using the three following 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 preprocessed each obtained corpus by eliminating punctuation symbols, stopwords and, thereafter, by applying the Porter stemmer. The characteristics of each corpus after the pre-processing step were given in Tables 2.13, 2.14, 2.18 and 2.17 of Section 2.3. In order to calculate the correlation for several corpora with similar characteristics we have constructed subsets of each corpus by using the technique described in Section 7.1.1.

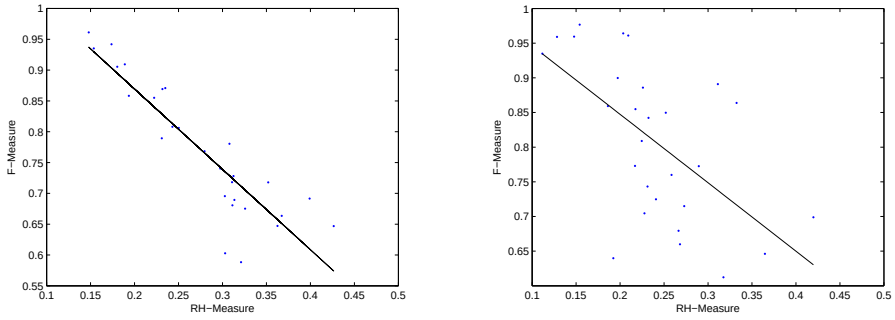
In order to determine the relative hardness of a given corpus, we have considered the vocabulary overlapping among the texts of the corpus. In our experiments, we have used the Jaccard-based overlapping measure described in Section 4.1.4. We have carried out an unsupervised clustering of all the documents of each subcorpus obtained for each dataset. We have chosen the MajorClust clustering algorithm [134] 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 the MRH measure, we have used the tf-idf formula for calculating the input similarity matrix for MajorClust. A better explanation of the tf-idf formula was given in Section 2.1.1, whereas the description of the MajorClust clustering algorithm was presented in Section 2.1.4. Each evaluation was performed with the F-Measure formula which was calculated as shown in Section 2.1.5.

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 7.8 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. We can appreciate for both corpora that the smaller is the value of MRH (x-axis) the higher is the obtained F-Measure (y-axis) and viceversa. The correlation of the relative hardness vs. the F-Measure was calculated for all the possible sub-corpora variants



(a) Train (b) Test

Figure 7.8: Evaluation of all R8 subcorpora (more than two categories per corpus)



(a) Train (b) Test

Figure 7.9: Evaluation of single pairs of the *R8-Reuters* categories

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 7.9 we may see the possible correlation between the relative hardness of each pair of categories of the *R8-Reuters* collection and the F-Measure, which was obtained again by using the MajorClust clustering algorithm. The same conclusion may be drawn: the smaller is the value of MRH (x-axis) the higher is the obtained F-Measure (y-axis) and viceversa.

In order to fully appreciate the MRH formula, the most and least related pairs of categories for the *R8-Reuters* dataset are presented in Tables 7.1 and 7.2, respectively.

Table 7.1: The most related categories of the *R8-Reuters* collection

RH value	**Category**	**Category**
0.426	trade	monex-fx
1 0.399	monex-fx	interest
2 0.367	trade	crude
3 0.362	monex-fx	crude
4 0.352	trade	interest
5		

(a) Train

RH value	**Category**	**Category**
0.419	monex-fx	interest
2 0.364	trade	monex-fx
1 0.332	trade	interest
5 0.317	trade	crude
3 0.311	monex-fx	crude
4		

(b) Test

Table 7.2: The least related categories of the *R8-Reuters* collection

RH value	**Category**	**Category**
0.188	interest	earn
0.180	acq	ship
0.173	ship	earn
0.153	grain	acq
0.147	grain	earn

(a) Train

RH value	**Category**	**Category**
0.186	interest	acq
0.154	ship	earn
0.147	acq	ship
0.128	grain	earn
0.111	grain	acq

(b) Test

The MRH value associated with each pair was calculated with the same formula introduced in Section 4.1.4. There exists a high agreement when executing the MRH formula over both, the training and test dataset of R8. In fact, the Kendall tau coefficient value asocciated to the correlation between the training and test set is 0.4. According to the relative hardness measure proposed, the *trade*, *monex-fx*, and *interest* categories are the most related ones and, therefore, those that would cause more difficulty when clustering the R8 dataset. On the contrary, the *grain*, *acq*, and *earn* categories are the least related ones and, therefore, the easiest ones to be discriminated.

Some preliminary experiments were carried out also on the *Mini20Newsgroups* dataset and the most and least related pairs of categories are shown in Tables 7.3 and 7.4, respectively. In these tables it is easier to analyse (from a subjective pointview) the performance of the unsupervised relative hardness measure by inspecting the

Table 7.3: 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

category names.

7.3 Concluding remarks

Short texts clustering is one of the most difficult tasks in natural language processing given the low frequencies of the document terms. In this chapter we presented the evaluation of different internal clustering validity measures over narrow domain short-text corpora. The aim was to determine the possible correlation between these measures and F-Measure, a well-known external clustering measure used to calculate the performance of clustering algorithms. In the experiments carried out, we used several corpora (358). The correlation obtained with a particular set of internal validity measures allows us to conclude that some of them can be used to improve the performance of clustering algorithms when they have to deal with short texts.

presented the evaluation of different internal clustering validity measures in order to determine the possible correlation between these measures and F-Measure, a well-known external clustering measure used to calculate the performance of clustering

Table 7.4: 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

algorithms. In the experiments carried out we have used several short-text corpora (358). The obtained correlation with a particular set of internal validity measures allow us to conclude that some of them may be used to improve the performance of clustering algorithms when they have to deal with short texts.

Our findings indicate that the EDM and the MRH measures are those that obtain the best results. However, it should be investigated whether the other ICVMs are related to specific kinds of corpora (for instance, narrow or wide domains) or whether they may be used to calculate the relative hardness of them. We have had some insights of the RH study with two widely used categorization datasets (Reuters and 20 Newsgroups). In the preliminary experiments, we studied the possible relationship between the degree of vocabulary overlapping of a given text corpus with the F-Measure which was obtained using the MajorClust clustering algorithm. We have observed that it is possible to determine the relative hardness of a corpus by using a measure based on vocabulary overlapping. The obtained results show that there is a correlation between the F-Measure and the RH formula. With respect to the analysis which was carried out in [35], the formula introduced in our research work

relies only on vocabulary overlapping and does not use any classifier. In fact, we use the MajorClust clustering algorithm only to evaluate the quality of the proposed formula by using the F-Measure. Therefore, we believe that the RH formula presented can be efficiently used to determine the relative hardness of corpora to be clustered.

Chapter 8

Conclusions and further work

In this chapter we draw the conclusions of the research that we have carried out.

The clustering of narrow domain short-text corpora is one of the most difficult tasks of unsupervised data analysis. The high overlapping of vocabularies among the texts in narrow domain corpora (with a consequent specific terminology), and the low term frequency of short texts lead us to investigate novel techniques to tackle both problems.

We have addresed the above problems by studying three lines of research:

1. The study of methods and techniques for improving clustering of narrow domain short-text corpora.

2. The determination of classifier-independent corpus features and the assessment of each of them.

3. The applications of the proposed methods and techniques in different areas of natural language processing.

8.1 Findings and research directions

Due to the high number of experiments performed, the findings are summarized in the following section. We also discuss some interesting research directions, which

are derived from the obtained results of this Ph.D. thesis and which we consider to be useful for future work.

8.1.1 Behaviour of unsupervised term selection techniques

Findings: A first experiment with a real collection of scientific abstracts (*hep-ex*) of the *High Energy Physics* domain [100] motivated this Ph.D. thesis. We analysed the behaviour of three unsupervised term selection techniques (DF, TS and TP) in the framework of clustering narrow domain short texts. The TP technique outperformed the other two techniques when using a subset of *hep-ex*. However, when the full document collection was used, the new TPMI term selection technique had to be developed in order to improve the previous unstable results obtained by the TP technique. TPMI takes advantage of a dictionary of related terms, which is constructed over the *same* collection, by using pointwise mutual information since common or general-purpose dictionaries are not very useful (due to the very specialised nature of narrow domain vocabularies). After the calculation of a baseline in both the full corpus and a subset of it, the experiments that were carried out allowed us to verify that the TPMI technique outperformed the other approaches.

Research direction: Due to the instability of TP, we carried out an analysis to understand its behaviour and to be able to determine the number of terms needed in the task. We observed that it does not seem possible to determine the number of terms that a term selection technique must choose in order to carry out the clustering task. A new research direction has arisen from this analysis: the automatic determination of cut-off points in vocabulary reduction.

8.1.2 The novel symmetric Kullback-Leibler distance

Findings: We studied 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 were few differences in the use of any of the symmetric KLDs analysed. We evaluated the proposed approach with three different narrow domain short-text corpora, and our findings indicated that

it is possible to use this measure to tackle this problem. We obtained results that were comparable to those that use the Jaccard similarity measure. Nevertheless, due to the fact that the KLD distance measure is computationally more expensive than the Jaccard one, the fastest measure was used in the majority of experiments we carried out.

Research direction: Even if we implemented the KLD to use it for clustering narrow domain short texts, we consider that this distance measure could also be employed for clustering more general domain and large size text corpora. The use of a smooth procedure should be useful since the vocabulary of each document is more similar to the corpus vocabulary. We consider that a performance improvement could be obtained by using a term expansion method before calculating the similarity matrix with the analysed KLD.

8.1.3 The impact of term selection techniques on word sense induction

Findings: We studied the impact of the term selection techniques in the standard *WSI-SemEval* data collection. We compared these results with those reported in [4]. The TP and DF term selection techniques outperformed two of the six systems with the additional advantage of vocabulary reduction.

8.1.4 Watermarking Corpora: some novel corpus evaluation measures

Findings: We presented a set of corpora evaluation measures that can be used to either evaluate the proposed gold standard or to make decisions a priori when, for instance, clustering particular types of text collections such as narrow domain short text corpora.

The evaluation measures were classified into five different categories: *domain broadness*, *shortness*, *class imbalance*, *stylometry* and *structure*. All the proposed measures were executed over several corpora in order to determine their evaluation

capability.

We introduced (un)supervised measures in order to assess these features. The supervised ones were used both to evaluate the corpus features and, more relevantly, to assess the gold standard provided by experts for the corpus to be clustered. The unsupervised measures directly evaluate the document collections, i.e., without any gold standard. Therefore, they can also be used for other purposes, for instance, to adjust clustering methods while being executed in order to improve clustering results.

The most successful set of measures were compiled in a tool named Watermarking Corpora On-line System (WaCOS), which will allow researchers in different fields of linguistics and computational linguistics to easily assess their corpora with respect to the aforementioned corpus features.

We ranked each corpus according to the evaluation value given by the corresponding measure. We then calculated the Kendall tau correlation coefficient in order to determine the degree of correlation between the automatically obtained and the manually obtained ranking. The findings are that the major evaluation measures obtained a very strong correlation with respect to the manual ranking.

Research direction: Our intention in this work was to include or exclude clustering corpora in order to concentrate our efforts on the most challenging clustering datasets, i.e., narrow domain short-text corpora. Even though we successfully categorized the analysed corpora, manual ranking of the values for each evaluation measure had to be done. In future works, it would be interesting to apply machine learning techniques to automatically fix the specific thresholds that must be used in the categorization of corpora.

The WaCOS system is a completely functional prototype that could be improved in the future by adding, for instance, different variants for the already implemented corpora evaluation measures.

8.1.5 The self-term expansion methodology

Findings: We have introduced a self-term expansion methodology that allows the baseline corpus to be enriched by adding co-related terms from an automatically

constructed lexical-knowledge resource obtained from the *same* target dataset (and not from an external resource). This was done by using two different co-occurrence techniques based on bigrams and pointwise mutual information, respectively. The experiments demostrated that the PMI outperforms the bigrams co-occurrence technique due to the fact 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 then the term selection process. Moreover, the clustering results of the target dataset obtained by just doing the self-term expansion alone are better than those obtained by classical methods of document representation.

The experiments were carried out on two real collections extracted from the CICLing-2000 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 main goal of this study was to boost the performance of clustering narrow domain short texts by employing the self-term expansion method. This sucessfully improved the baseline F-Measure by approximately 40%. 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 to the clustering of narrow domain short texts 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 number of documents is directly proportional to the performance of the proposed methodology.

8.1.6 The impact of the self-term expansion technique on word sense induction

Findings: The self-term expansion methodology is explicitly designed for narrow domain short-text corpora. It was applied to the word sense induction task which consists of distinguishing sentences with an ambiguous word from other sentences that have the same ambiguous word but with a different sense. The results with a corpus

written in English showed that the technique employed obtained a better performance than the baseline, especially the baseline that had chosen the most frequent sense. In fact, we outperformed every other unsupervised approach. The third place that we obtained at the SemEval competition [107] highlights how valuable this simple technique can be in the clustering process.

We confirmed that the self-term expansion technique improves the clustering of the unexpanded corpus no matter which term selection technique is used when enriching the corpus subsets. Moreover, we observed that when some kind of important information is known a priori, such as ambiguous words, the method may even improve the results by just expanding only the most important terms of the corpus instead of each one of them.

The evaluation with the WSI-SemEval corpus of the "Evaluating Word Sense Induction and Discrimination Systems" task of the SemEval 2007 workshop showed that expanding only the ambiguous terms is the best approach for word sense induction.

We also studied the language-independent characteristic of the self-term expansion methodology for the word sense induction/discrimination task. A set of preliminary experiments also showed good performance in the Arabic language. The tokenization performed on the Arabic corpus of the SemEval workshop by the task organisers was only partial since they kept, for instance, the Arabic definite article "Al" joined to the words. Even though this partial tokenization might be positive for other natural language processing tasks, we consider that the method presented in this research work would have performed better if the tokenization used had taken into consideration the definite article.

We consider that the evaluation of the proposed methodology on a real task, which was performed in an international forum has been really positive and measures its performance fairly. The evaluation also has provided us the opportunity to detect points for improvement. Our aim is to study the behaviour of the self-term expansion methodology in other areas of application.

Research direction: The language-independent characteristic of the proposed methodology has not yet been fully proved. Further experiments with corpora in other language that are different from English will confirm this hypothesis.

There are other potential applications where this methodology should be tested such as automatic summary generation, clustering of snippets, homonymy discrimination, etc.

8.1.7 The evaluation of internal clustering validity measures

Findings: We presented the evaluation of different internal clustering validity measures over narrow domain short-text corpora. The aim was to determine the possible correlation between these measures and F-Measure, a well-known external clustering measure used to calculate the performance of clustering algorithms.

In the experiments carried out, we used several corpora (358). The obtained correlation with a particular set of internal validity measures allows us to conclude that some of them can be used to improve the performance of clustering algorithms when they have to deal with short texts. We specifically observed that the two best correlated measures are the ones based on expected density and vocabulary overlapping (RH), respectively.

We have had some insights of the RH study with two widely used categorization datasets (Reuters and 20 Newsgroups). In the preliminary experiments, we studied the possible relationship between the degree of vocabulary overlapping of a given text corpus with the F-Measure which was obtained using the MajorClust clustering algorithm. We have observed that it is possible to determine the relative hardness of a corpus by using of a measure based on vocabulary overlapping. The obtained results show that there is a correlation between the F-Measure and the RH formula.

With respect to the analysis which was carried out in [35], the formula introduced in our research work relies only on vocabulary overlapping and does not use any classifier. In fact, we use the MajorClust clustering algorithm only to evaluate the quality of the proposed formula by using the F-Measure. Therefore, we believe that the RH formula presented can be efficiently used to determine the relative hardness of corpora to be clustered.

In the next section, we summarise the major contributions of this research work.

8.2 Major contributions

The major contributions of this research work are enumerated as follows:

1. The study and introduction of evaluation measures to analyse the following features of a corpus: *shortness*, *domain broadness*, *class imbalance*, *stylometry* and *structure*.

2. The development of the Watermarking Corpora On-line System, named WaCOS, for the assessment of corpus features.

3. A new unsupervised methodology (which does not use any external knowledge resource) for dealing with narrow domain short-text corpora. This methodology suggests first applying self-term expansion and then term selection.

8.3 Further work

There are other experiments that we consider to be important to future research. It would be interesting to observe the possible relationship that the clustering of narrow domain short-text corpora may have with summarization and viceversa. The idea would be to integrate the proposed self-term expansion technique into the summarization task and to determine whether or not the added methodology improves the classical summarization approach. When we talk about the classical approach, we refer to a summarization system that does not use the self-term expansion nor any other term selection techniques. Until now, the summarization task has fully placed its focus on question answering[1] since people are interested in obtaining a summary from the global content of a data collection. Therefore, it would be important to experiment with a simple technique that integrates at least the following areas of natural language processing: information retrieval, clustering and summarization.

Among the 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,

[1] `http://www.nist.gov/tac/`

there is no exact boundary among different clusters. Therefore, introducing a fuzzy version of the clustering methods used (for instance, the MajorClust algorithm [73]) would extend the anaylisis carried out in this Ph.D. thesis. The clustering method will assign documents to more than one cluster by taking into account a membership function for both the edges and nodes of the input similarity matrix for this clustering algorithm. Thus, the clustering problem will be formulated in terms of weighted fuzzy graphs. The fuzzy approach will decrease some of the negative effects that appear in the clustering of large-sized corpora with noisy data.

After implementing the fuzzy version of the clustering algorithm, its performance should be tested against other fuzzy clustering algorithms for the specific problem we have studied: the clustering of narrow domain short-text corpora.

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 P. Edmonds. *Word Sense Disambiguation: Algorithms and Applications*. Springer, 2007.

[4] 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.

[5] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. of the very large databases conference - VLDB'96*, pages 487–99, 1994.

[6] 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.

[7] R. Baeza-Yates and B. Ribeiro-Neto. *Modern information retrieval*. New York: ACM Press; Addison-Wesley, 1999.

[8] A. Bagga and B. Baldwin. Algorithms for scoring coreference chains. In *Proc. of the Linguistic Coreference Workshop at the First International Conference on Language Resources and Evaluation (LREC'98)*, pages 563–566, 1998.

[9] L. R. Bahl, E Jelinek, and R. L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):179–190, 1983.

[10] 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.

[11] S. Banerjee, K. Ramanathan, and A. Gupta. Clustering short texts using wikipedia. In *SIRIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 787–788. ACM Press, 2007.

[12] Y. Benajiba and P. Rosso. Towards a measure for arabic corpora quality. In *Proc. of the International Colloquium on Arabic Language Processing - CITALA-2007*, pages 213–221, 2007.

[13] C. H. Bennett, P. Gács, M. Li, P. Vitányi, and W. Zurek. Information distance. *IEEE Transactions on Information Theory*, 44(4):1407–1423, 1998.

[14] 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.

[15] 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.

[16] 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.

[17] B. Bigi, R. d. Mori, M. El-Bèze, and T. Spriet. A fuzzy decision strategy for topic identication and dynamic selection of language models. *Special Issue on Fuzzy Logic in Signal Processing, Signal Processing Journal*, 80(6):1085–1097, 2000.

[18] 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.

[19] A. D. Booth. A law of occurrences for words of low frequency. *Information and control*, 10(4):386–393, 1967.

[20] T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean. Large language models in machine translation. In *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, 2007.

[21] C. H. Brooks and N. Montanez. An analysis of the effectiveness of tagging in blogs. Technical Report SS-06-03, Integrated Intelligent Knowledge Management. In Computational Approaches to Analyzing Weblogs: Papers from the 2006 Spring Symposium, ed. N. Nicolov, F. Salvetti, M. Liberman, and J. H. Martin, 24-31. American Association for Artificial Intelligence, Menlo Park, California, 2006.

[22] P. F. Brown, V. J. Della Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.

[23] C. Buckley and A. F. Lewit. Optimizations of inverted vector searches. In *Proc. of the 8th annual international ACM SIGIR conference on Research and development in information retrieval-SIGIR'85*, pages 97–110. Association for Computing Machinery (ACM), 1985.

[24] T. Buckwalter. Issues in arabic orthography and morphology analysis. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, COLING 2004*, Geneva, Italy, 2004.

[25] 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.

[26] D. Buscaldi, A. Juan, P. Rosso, and M. Alexandrov. Sense cluster-based categorization and clustering of abstracts. In *Proc. of the CICLing 2006 Conference*, volume 3878 of *Lecture Notes in Computer Science*, pages 547–550. Springer-Verlag, 2006.

[27] D. Buscaldi, P. Rosso, and F. Masulli. The upv-unige-CIAOSENSO WSD system. In *Proc. of the Senseval-3 Workshop*, pages 77–82. Association for Computational Linguistics, 2004.

[28] F. Can and J. M. Patton. Change of writing style with time. *Computers and the Humanities*, 38(1):61–82, 2004.

[29] A. Cardoso-Cachopo and A. Oliveira. Combining LSI with other classifiers to improve accuracy of single-label text categorization. In *First European Workshop on Latent Semantic Analysis in Technology Enhanced Learning - EWLSATEL 2007*, 2007.

[30] 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.

[31] M. Carpuat and D. Wu. Improving statistical machine translation using word sense disambiguation. In *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 61–72. Association for Computational Linguistics, 2007.

[32] R. d. Mori. *Spoken Dialogues with Computers*. Academic Press, 1998.

[33] I. Dagan, L. Lee, and F. Pereira. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1–3):43–69, 1999.

[34] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227, 1979.

[35] 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.

[36] J. Diederich, J. Kindermann, E. Leopold, and G. Paass. Authorship attribution with support vector machines. *Applied Intelligence*, 19(1–2):109–123, 2004.

[37] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classication, (2nd Edition)*. Wiley-Interscience, 2000.

[38] J. C. Dunn. Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4:95–104, 1974.

[39] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[40] R. Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. PhD dissertation, University of California, Irvine, 2000.

[41] 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.

[42] E. B. Fowlkes, R. Gnanadesikan, and J. R. Kettenring. Variable selection in clustering. *Journal of Classification*, 5:205–228, 1988.

[43] 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.

[44] G. Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic, 1994.

[45] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. Cluster validity methods: Part i. *SIGMOD Record*, 31(2):40–45, 2002.

[46] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. Clustering validity checking methods: Part ii. *SIGMOD Record*, 31(3):19–27, 2002.

[47] Z. Harris. Distributional structure. *Word*, 10(23):146–162, 1954.

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

[49] G. Herdan. *Quantatative Linguistics*. London: Butterworth, 1964.

[50] D. L. Hoover. Another perspective on vocabulary richness. *Computers and the Humanities*, 37(2):151–178, 2004.

[51] D. L. Hoover. Corpus stylistics, stylometry, and the styles of henry james. *Style*, 41(2):174–203, 2007.

[52] A. Hotho, S. Staab, and G. Stumme. Ontologies improve text document clustering. In *Proc. of the Third IEEE International Conference on Data Mining -ICDM03*, pages 1–4, 2003.

[53] A. Hotho, S. Staab, and G. Stumme. WordNet improves text document clustering. In *Proc. of the Semantic Web Workshop at SIGIR-2003, 26th Annual International ACM SIGIR Conference*, 2003.

[54] J. Hynek, K. Jezek, and O. Rohlik. Short document categorization - itemsets method. In *Proc. of the 4th. European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2000. `http://textmining.zcu.cz/?section=publication&id=21`.

[55] 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.

[56] N. Ide and J. Véronis. Word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40, 1998.

[57] D. Ingaramo, D. Pinto, P. Rosso, and M. Errecalde. Evaluation of internal validity measures in short-text corpora. In *Proc. of the CICLing 2008 Conference*, volume 4919 of *Lecture Notes in Computer Science, Springer-Verlag*, pages 555–567, 2008.

[58] D. A. Ingaramo, M. L. Errecalde, and P. Rosso. Medidas internas y externas en el agrupamiento de resúmenes científicos de dominios reducidos (in spanish). *Procesamiento del Lenguaje Natural*, 39(1):55–62, 2007.

[59] N. Japkowicz. The class imbalance problem: Significance and strategies. In *Proc. of the 2000 International Conference on Artificial Intelligence (IC-AI'2000)*, volume 1, pages 111–117, 2000.

[60] H. Jiménez, D. Pinto, and P. Rosso. Selección de términos no supervisada para agrupamiento de resúmenes (in spanish). In *Proc. of the Human Language Workshop - ENC05*, pages 86–91, 2005.

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

[62] S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32:241–254, 1967.

[63] 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.

[64] G. Karypis, E.-H. Han, and K. Vipin. Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75, 1999.

[65] M. Kendall. A new measure of rank correlation. *Biometrika*, 30:81–89, 1938.

[66] B. W. Kernighan and S. Lin. An efficient heuristic procedure for partitioning graphs. *Bell Systems Technical Journal*, 49(2):291–308, 1970.

[67] G. Kowalski. *Information Retrieval Systems Theory and Implementation*. Kluwer Academic Publishers, 1997.

[68] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[69] 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.

[70] 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.

[71] E. L. Lehmann and H. J. M. D'Abrera. *Nonparametrics: Statistical Methods Based on Ranks*. Prentice-Hall, 1998.

[72] 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.

[73] 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*, volume 4827 of *Lecture Notes in Artificial Intelligence*, pages 821–830. Springer-Verlag, 2007.

[74] 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.

[75] D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[76] 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.

[77] 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.

[78] P. Makagonov, M. Alexandrov, and K. Sboychakov. Keyword-based technology for clustering short documents. *Selected Papers. Computing Research*, 2:105–114, 2000.

[79] D. C. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, 1999.

[80] D. C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 2003. Revised version May 1999.

[81] L. Màrquez and L. Padró. A flexible POS tagger using an automatically acquired language model. In *Proc. of the 35th annual meeting on Association for Computational Linguistics*, pages 238–245, 1997.

[82] S. Meyer zu Eissen. *On information need and categorizing search*. PhD dissertation, University of Paderborn, Germany, Feb 2007.

[83] 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.

[84] G. W. Milligan. A validation study of a variable weighting algorithm for cluster analysis. *Journal of Classification*, 6:53–71, 1989.

[85] B. G. Mirkin. *Mathematical Classification and Clustering*. Springer, 1996.

[86] T. M. Mitchell. *Machine Learning*. McGraw-Hill Higher Education, 1997.

[87] A. Montejo and L. A. Ureña. Binary classifiers versus AdaBoost for labeling of digital documents. *Procesamiento del Lenguaje Natural*, 37(1):319–326, 2006.

[88] A. Montejo-Ráez. *Automatic Text Categorization of Documents in the High Energy Physics Domain*. PhD dissertation, Granada University, Spain, Feb 2006.

[89] 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.

[90] 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.

[91] 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.

[92] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69:026113, 2004. `http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0308217`.

[93] R. V. O'Neill, A. R. Johnson, and A. W. King. Graininess and entropy. *J. Opt. Soc. Am.*, 48(1):945–947, 1958.

[94] C. Ordonez and E. Omiecinski. Accelerating EM clustering to find high-quality solutions. *Knowledge and Information Systems*, 7(2):135–157, 2005.

[95] V. Parker-Lessig. Comparing cluster analyses with cophenetic correlation. *Journal of Marketing Research*, 9(1):82–84, 1972.

[96] 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*, pages 799–805, 2004.

[97] J. Peng, D.-Q. Yang, J.-W. Wang, M.-Q. Wu, and J.-G. Wang. A clustering algorithm for short documents based on concept similarity. In *Proc. IEEE Pacific Rim Conference on Communications, Computers and Signal Processing - PACRIM'07*, pages 42–45. IEEE, 2007.

[98] D. Pinto. *Analysis of narrow-domain short texts clustering*. Research report for "Diploma de Estudios Avanzados (DEA)", Department of Information Systems and Computation, UPV, 2007.

[99] 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.

[100] 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.

[101] 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.

[102] D. Pinto, A. Juan, and P. Rosso. Using query-relevant documents pairs for cross-lingual information retrieval. In *Proc. of the Text, Speech and Dialogue 2007 Conference - TSD07*, volume 4629 of *Lecture Notes in Artificial Intelligence*, pages 630–637. Springer-Verlag, 2007.

[103] 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.

[104] 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.

[105] 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.

[106] D. Pinto, P. Rosso, Y. Benajiba, A. Ahachad, and H. Jiménez-Salazar. Word sense induction in the arabic language: A self-term expansion based approach. In *Proc. 7th Conference on Language Engineering of the Egyptian Society of Language Engineering - ESOLE-2007*, pages 235–245, 2007.

[107] 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.

[108] 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.

[109] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Research and Development in Information Retrieval*, pages 275–281, 1998.

[110] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3), 1980.

[111] Q. Pu and G.-W. Yang. Short-text classification based on ICA and LSA. In *Proc. of the 3rd. International Symposium on Neural Networks - ISNN 2006*, Advances in Neural Networks, pages 265–270. Springer, 2006.

[112] 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.

[113] 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.

[114] L. Qun and L. Shu-jian. Word similarity computing based on HowNet. *Computational Linguistics and Chinese Language Processing*, 7(2):59–76, 2002.

[115] D. Reforgiato-Recupero. A new unsupervised method for document clustering by using WordNet lexical and conceptual relations. *Information Retrieval*, 10(6):563–579, 2007.

[116] 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.

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

[118] 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.

[119] 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.

[120] T. G. Rose, M. Stevenson, and M. Whitehead. The Reuters Corpus volume 1 - from yesterday's news to tomorrow's language resources. In *Proc. of the 3rd*

International Conference on Language Resources and Evaluation - LREC02, pages 827–832, 2002.

[121] S. Roukos. *Language Representation. In Survey of the State of the Art in Human Language Technology*. Cambridge University Press, 1997. Eds. G. B. Varile and A. Zampolli.

[122] P. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65, 1987.

[123] G. Ruge. Experiments on linguistically-based term associations. *Information Processing & Management*, 28(3):317–332, 1992.

[124] G. Salton, A. Wong, and C. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[125] Y. Santiesteban and A. Pons-Porrata. LEX: a new algorithm for the calculus of typical testors. *Mathematics Sciences Journal*, 21(1):85–95, 2003.

[126] H. Schütze. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123, 1998.

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

[128] J. Sedding and D. Kazakov. WordNet-based text document clustering. In *COLING 2004 3rd Workshop on Robust Methods in Analysis of Natural Language Data*, pages 104–113. COLING, 2004.

[129] X. Sevillano, G. Cobo, F. Alías, and J. C. Socoró. Robust document clustering by exploiting feature diversity in cluster ensembles. *Procesamiento del Lenguaje Natural*, 37:169–176, 2006.

[130] 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.

[131] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish. Clustering speakers by their voices. In *Proc. of the ICASSP98 Conference*, pages 757–760, 1998.

[132] B. Stein, S. Meyer, and F. Wißbrock. On cluster validity and the information need of users. In *Proceedings of the 3rd IASTED*, pages 216–221. ACTA Press, 2003.

[133] B. Stein and S. Meyer zu Eissen. Automatic document categorization. In *Proc. of Advances in Artificial Intelligence - KI 2003*, pages 254–266, 2003.

[134] 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.

[135] A. Stolcke. SRILM – an extensible language modeling toolkit, 2002.

[136] M. Sussna. Word sense disambiguation for free-test indexing using a massive semantic network. In *Proc. of the 2nd International Conference on Information and Knowledge Management*, pages 67—74, 1993.

[137] F. J. Tweedie and R. H. Baayen. How variable may a constant be?: Measures of lexical richness in perspective. *Computers and the Humanities*, 32(5):323–352, 1998.

[138] A. R. Urbizagástegui. Las posibilidades de la ley de Zipf en la indización automática. Technical report, Universidad de California, Riverside, 1999.

[139] M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. A model-theoretic coreference scoring scheme. In *Proc. of the Sixth Message Understanding Conference (MUC-6)*, pages 45–52, 1995.

[140] L. Wang, L. Tian, Y. Jia, and W. Han. A hybrid algorithm for web document clustering based on frequent term sets and k-means. In *Advances in Web and Network Technologies, and Information Management*, pages 198–203, 2007.

[141] Y. Wang, Y. Jia, and S. Yang. Parallel mining of top-k frequent itemsets in very large text database. In *Advances in Web-Age Information Management*, pages 706–712, 2005.

[142] Y. Wang, Y. Jia, and S. Yang. Short documents clustering in very large text databases. In *Proc. of the Web Information Systems WISE 2006 Workshops*, pages 83–93, 2006.

[143] W. Weaver. *Translation.* Mimeographed, 12 pp., July 15, 1949. Reprinted in Locke, William N. and Booth, A. Donald (1955) (Eds.), Machine translation of languages. John Wiley & Sons, New York, 15-23, 1949.

[144] W. Wibowo and H. E. Williams. On using hierarchies for document classification. In *Proc. of the Australian Document Computing Symposium*, pages 31–37, 1999.

[145] J. W. Wilbur and K. Sirotkin. The automatic identification of stopwords. *Journal of Information Science*, 18:45–55, 1997.

[146] 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.

[147] I. H. Witten and T. C. Bell. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37(4):1085–1094, 1991.

[148] I. H. Witten and E. Frank. *Data mining: Practical machine learning tools and techniques with Java implementations.* Morgan Kaufmann, 2000.

[149] 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.

[150] 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.

[151] 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.

[152] 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/`.

[153] S. Zelikovitz and H. Hirsh. Transductive LSI for short text classification problems. In *Proc. of the 17th International Conference on Machine Learning - ICML2000*, pages 1183–1190, 2000.

[154] S. Zelikovitz and F. Marquez. Transductive learning for short-text classification problems using latent semantic indexing. *International Journal of Pattern Recognition and Artificial Intelligence*, 19(2):143–163, 2005.

[155] Y. Zhao and G. Karypis. *Criterion Functions for Document Clustering: Experiments and Analysis*. technical report, Univ. of Minnesota, Dept. of Computer Science, Minneapolis, 2002.

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

[157] 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.

Appendix A

Other external clustering validity measures

In this appendix we are refering some possible measures, apart from F-Measure which could be used to validate the results obtained by clustering algorithms with respect to a given gold standard. First, we present the following formal definition of clustering and gold standard.

Given a document collection $D = \{d_1, d_2, \cdots, d_{|D|}\}$, a *clustering* of D is a partition into k subsets $\mathcal{C} = \{C_1, C_2, \cdots, C_k | C_i \subseteq D\}$, such that $\bigcup_{i=1}^{k} C_i = D$, whereas the *gold standard* of D is a partition into l subsets $\mathcal{C}^* = \{C_1^*, C_2^*, \cdots, C_l^* | C_i^* \subseteq D\}$ constructed by using human criteria.

A.1 Pairwise Precision/Recall/Accuracy

Precision, recall, and accuracy are defined as shown in Eq. (A.1), (A.2), and (A.3), respectively. For every pair of documents, a True Positive (TP) or True Negative (TN) is defined when the pair is coreferent or non-coreferent in both the obtained clusters and the gold standard. False Positives (FP) or False Negatives (FN) are defined when

there exists a disagreement on whether the documents are coreferent or not.

$$Precision = \frac{TP}{(TP + FP)} \tag{A.1}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{A.2}$$

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{A.3}$$

A.2 MUC Precision/Recall

MUC is the scoring defined by Vilain et al. [139] and used in the Message Understanding Conferences. This model calculates the number of clusters minus the number of missing links, divided by the number of documents minus the number of clusters (switching classes and clusters for recall).

A.2.1 MUC Precision

$$\frac{|D| - \sum_{C \in \mathcal{C}} \left| \left\{ C^* \in \mathcal{C}^* | C \cap C^* \neq \emptyset \right\} \right|}{|D| - |\mathcal{C}|} \tag{A.4}$$

$$\frac{\sum_{C \in \mathcal{C}} |C| - \left| \left\{ C^* \in \mathcal{C}^* | C \cap C^* \neq \emptyset \right\} \right|}{\sum_{C \in \mathcal{C}} |C| - 1} \tag{A.5}$$

A.2.2 MUC Recall

$$\frac{|D| - \sum_{C^* \in \mathcal{C}^*} \left| \left\{ C \in \mathcal{C} | C \cap C^* \neq \emptyset \right\} \right|}{|D| - |\mathcal{C}^*|} \tag{A.6}$$

$$\frac{\sum_{C^* \in \mathcal{C}^*} |C^*| - \left| \left\{ C \in \mathcal{C} | C \cap C^* \neq \emptyset \right\} \right|}{\sum_{C^* \in \mathcal{C}^*} |C^*| - 1} \tag{A.7}$$

A.3 B-Cubed Precision/Recall

The B-Cubed (B^3) metric of Bagga and Baldwin [8] is the precision computed and averaged for each document individually with its corresponding generated cluster and gold standard class, reversing clusters and classes for recall.

A.3.1 B-Cubed Precision

$$\frac{1}{|D|}\sum_{C\in\mathcal{C}}\sum_{C^*\in\mathcal{C}^*}\frac{|C\cap C^*|^2}{|C|} \tag{A.8}$$

$$\frac{1}{|D|}\sum_{d\in D}\sum_{C\in\mathcal{C}|d\in C}\sum_{C^*\in\mathcal{C}^*|d\in C^*} Precision(C,C^*) \tag{A.9}$$

A.3.2 B-Cubed Recall

$$\frac{1}{|D|}\sum_{C^*\in\mathcal{C}^*}\sum_{C\in\mathcal{C}}\frac{|C\cap C^*|^2}{|C^*|} \tag{A.10}$$

$$\frac{1}{|D|}\sum_{d\in D}\sum_{C^*\in\mathcal{C}^*|d\in C^*}\sum_{C\in\mathcal{C}|d\in C} Precision(C^*,C') \tag{A.11}$$

A.4 Purity/Inverse Purity

This metric maps each obtained clustering to the gold standard class which gives the best precision, and then computes weighted average precision under this mapping, and reverses clusters and classes for inverse purity [131].

A.4.1 Purity

$$\frac{1}{|D|}\sum_{C\in\mathcal{C}}\max_{C^*\in\mathcal{C}^*}|C\cap C^*| \tag{A.12}$$

$$\frac{1}{|D|}\sum_{C\in\mathcal{C}}\max_{C^*\in\mathcal{C}^*}|C|*Precision(C,C^*) \tag{A.13}$$

A.4.2 Inverse Purity

$$\frac{1}{|D|}\sum_{C^*\in\mathcal{C}^*}\max_{C\in\mathcal{C}}|C\cap C^*| \tag{A.14}$$

$$\frac{1}{|D|}\sum_{C^*\in\mathcal{C}^*}\max_{C\in\mathcal{C}}|C^*|*Precision(C^*,C) \tag{A.15}$$

A.5 F-Purity/F-Inverse Purity

Similar to purity and inverse purity, this proposed metric maps each gold standard class to the generated cluster which gives the best harmonic mean of precision and recall, and then computes weighted average F-Measure under this mapping. The difference between this metric and purity/inverse purity is that the maximum is taken from the harmonic mean of precision and recall, rather than just the one being measured. This allows to measure the best matching cluster, rather than just the one which is the most precise or has the highest recall.

A.5.1 F-Purity

$$\frac{1}{|D|}\sum_{C\in\mathcal{C}}\max_{C^*\in\mathcal{C}^*}\frac{2*|C|*|C\cap C^*|}{|C|+|C^*|} \tag{A.16}$$

$$\frac{1}{|D|}\sum_{C\in\mathcal{C}}\max_{C^*\in\mathcal{C}^*}\frac{2*|C|*Precision(C,C^*)*Precision(C^*,C)}{Precision(C,C^*)+Precision(C^*,C)} \tag{A.17}$$

A.5.2 F-Inverse Purity

$$\frac{1}{|D|}\sum_{C^*\in\mathcal{C}^*}\max_{C\in\mathcal{C}}\frac{2*|C^*|*|C\cap C^*|}{|C|+|C^*|} \tag{A.18}$$

$$\frac{1}{|D|}\sum_{C^*\in\mathcal{C}^*}\max_{C\in\mathcal{C}}\frac{2*|C^*|*Precision(C,C^*)*Precision(C^*,C)}{Precision(C,C^*)+Precision(C^*,C)} \tag{A.19}$$

Appendix B

The specific behaviour of the evaluation measures

B.1 The *CICLing-2002* corpus

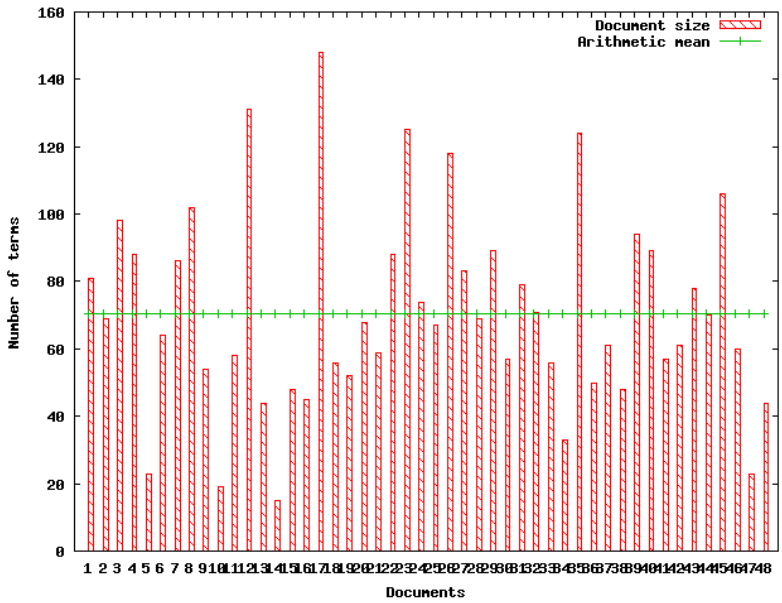


Figure B.1: Document cardinalities of the *CICLing-2002* corpus

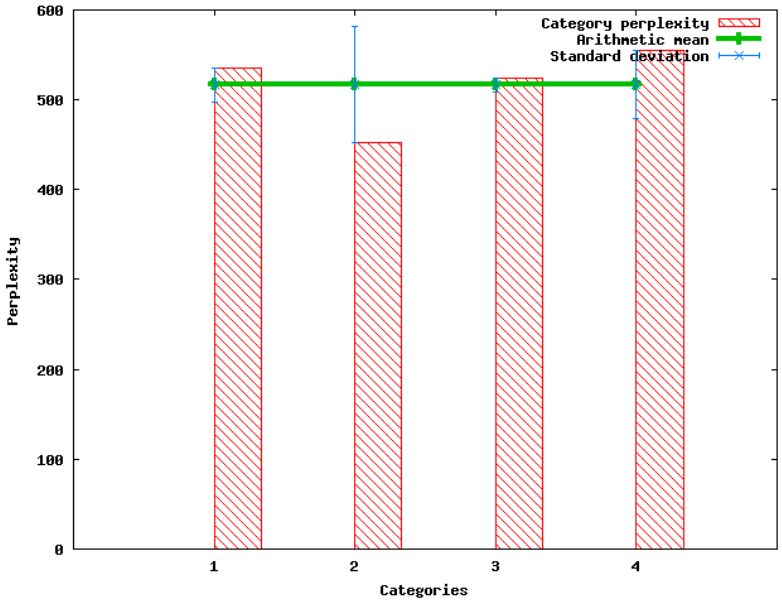


Figure B.2: Perplexity per category of the *CICLing-2002* corpus

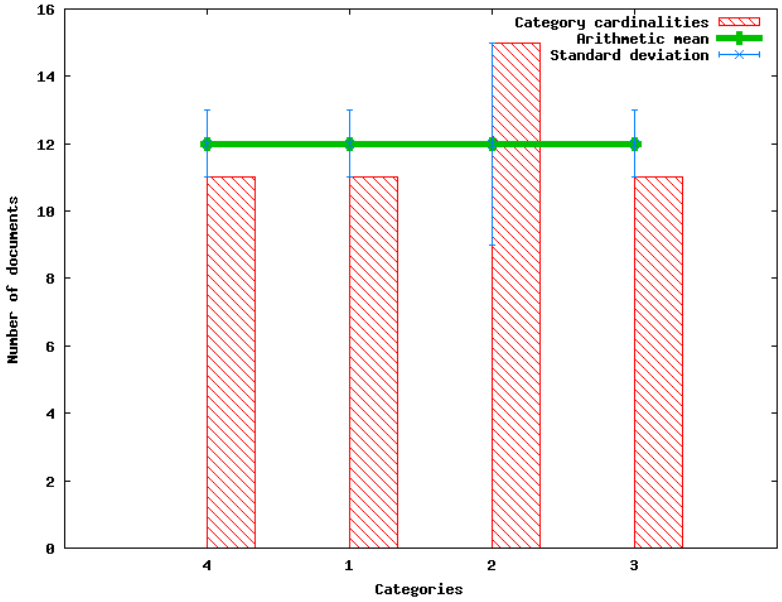


Figure B.3: Imbalance per category of the *CICLing-2002* corpus

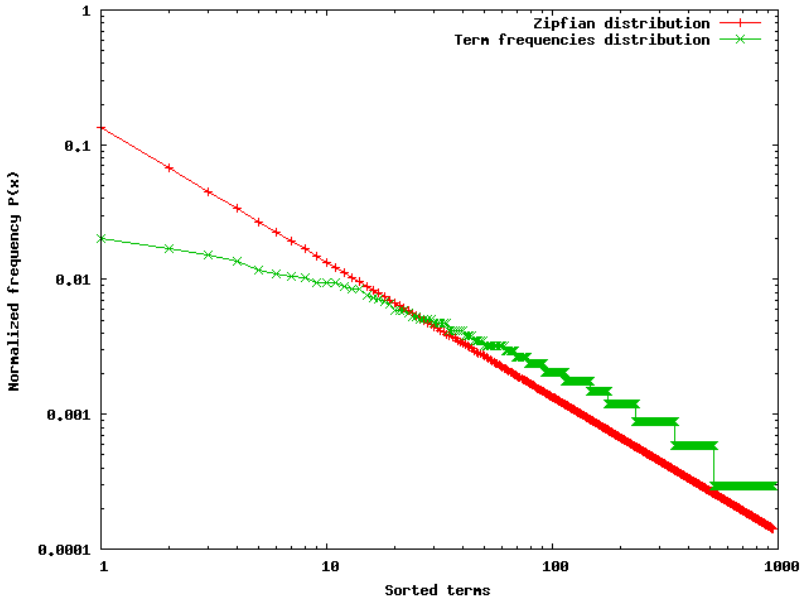


Figure B.4: Stylometry: All term frequency distribution of the *CICLing-2002* corpus

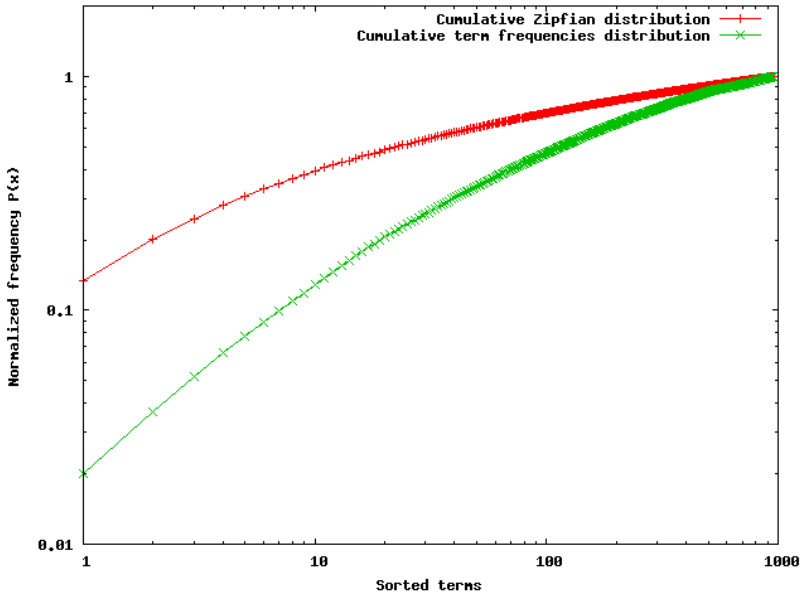


Figure B.5: Stylometry: All term frequency cumulative distribution of the *CICLing-2002* corpus

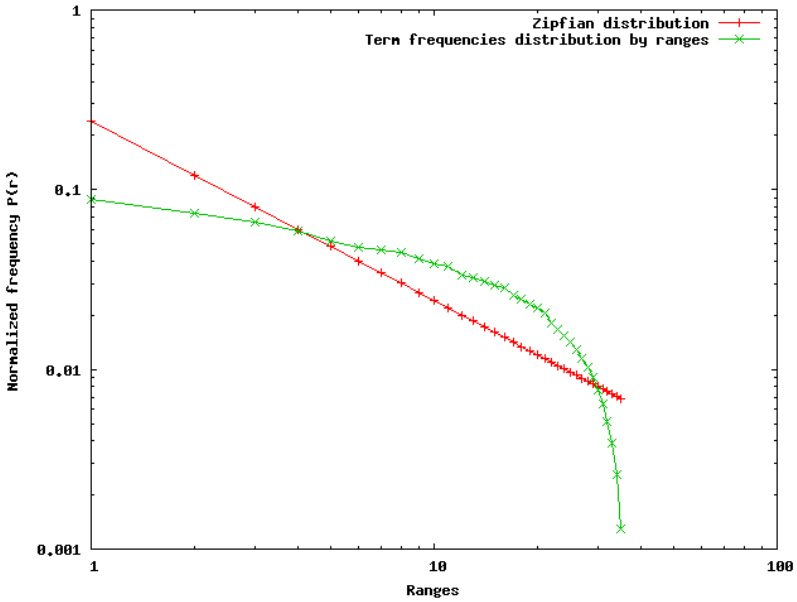


Figure B.6: Stylometry: Range frequency distribution of the *CICLing-2002* corpus

B.2 The *hep-ex* corpus

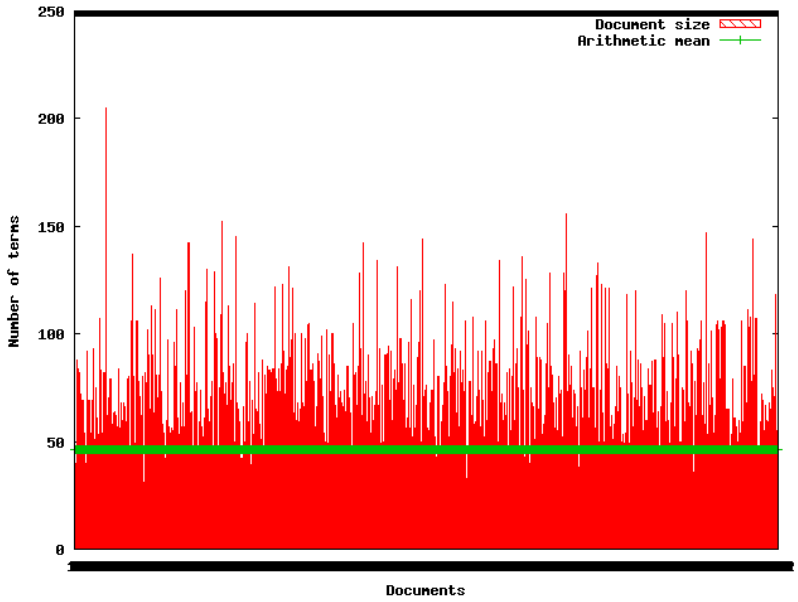


Figure B.7: Document cardinalities of the *hep-ex* corpus

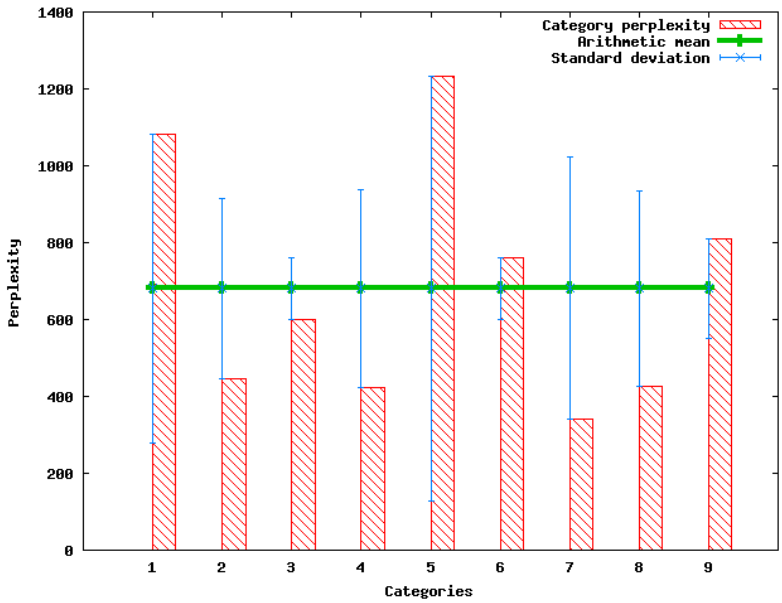


Figure B.8: Perplexity per category of the *hep-ex* corpus

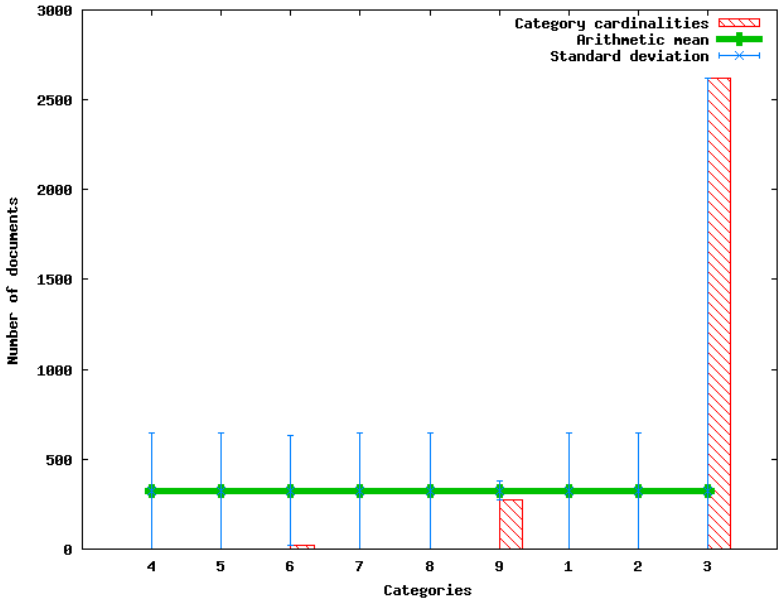


Figure B.9: Imbalance per category of the *hep-ex* corpus

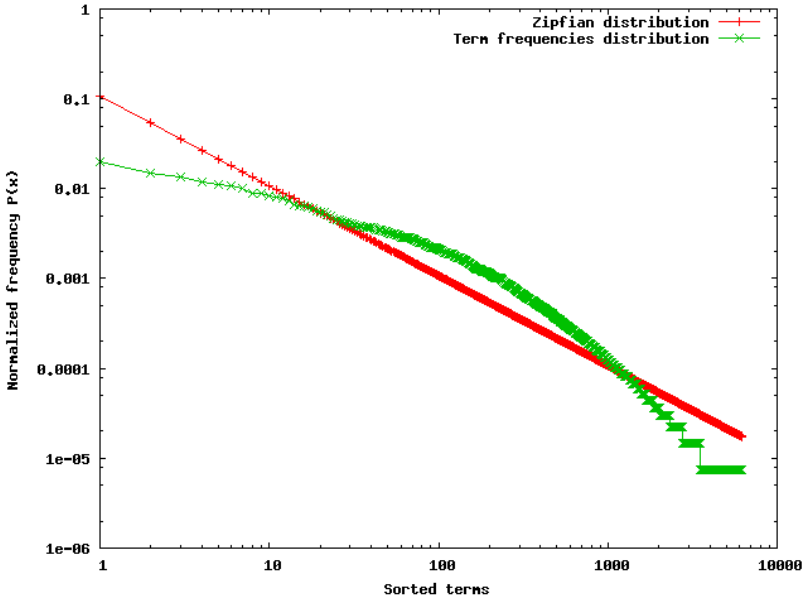


Figure B.10: Stylometry: All term frequency distribution of the *hep-ex* corpus

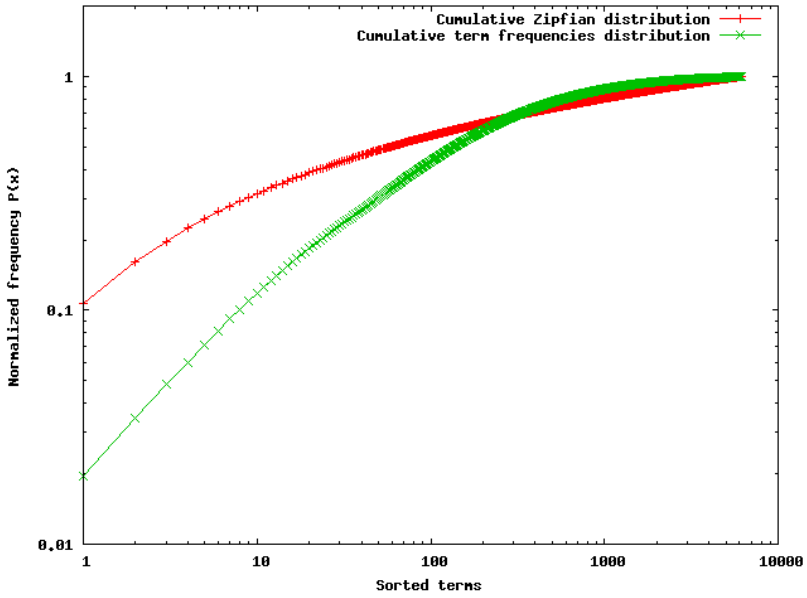


Figure B.11: Stylometry: All term cumulative frequency distribution of the *hep-ex* corpus

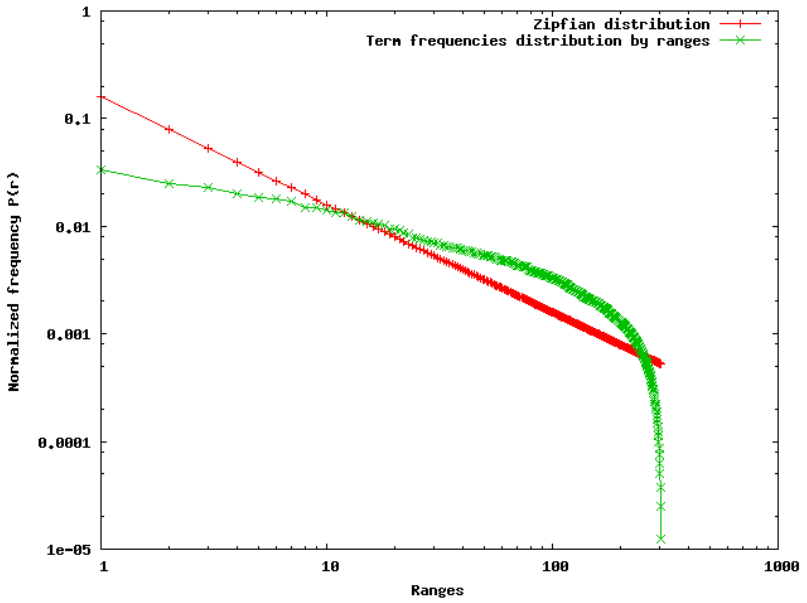


Figure B.12: Stylometry: Range frequency distribution of the *hep-ex* corpus

B.3 The *WebKB train* corpus

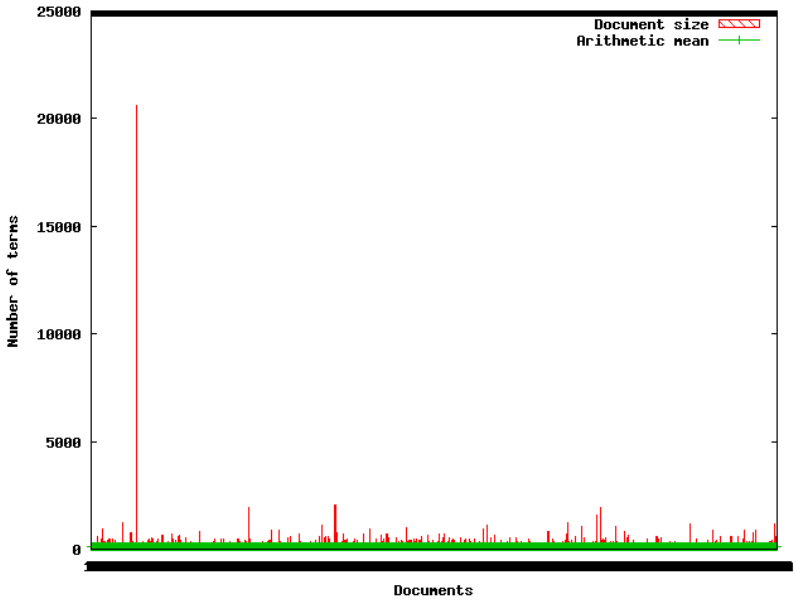


Figure B.13: Document cardinalities of the *WebKB train* corpus

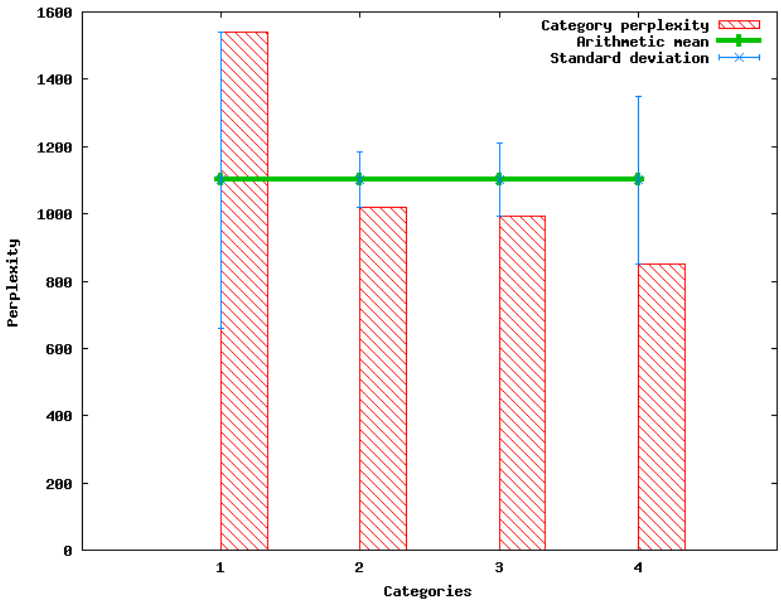


Figure B.14: Perplexity per category of the *WebKB train* corpus

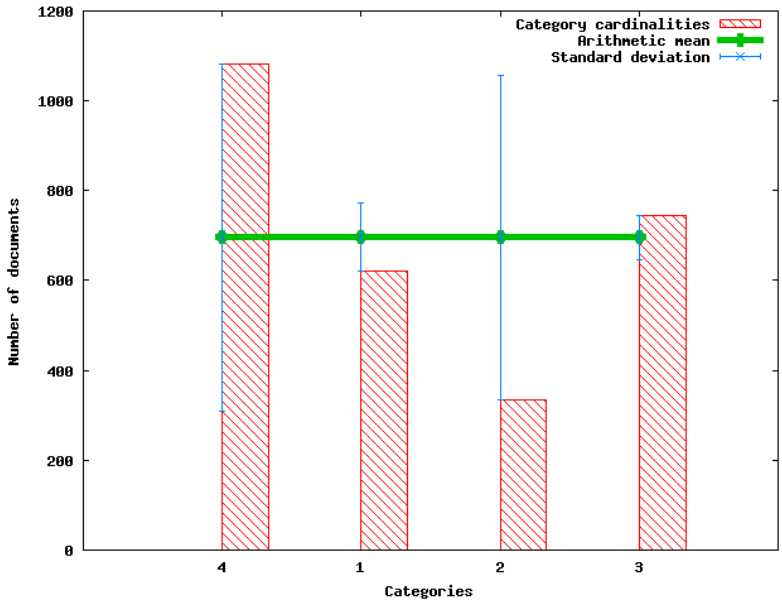


Figure B.15: Imbalance per category of the *WebKB train* corpus

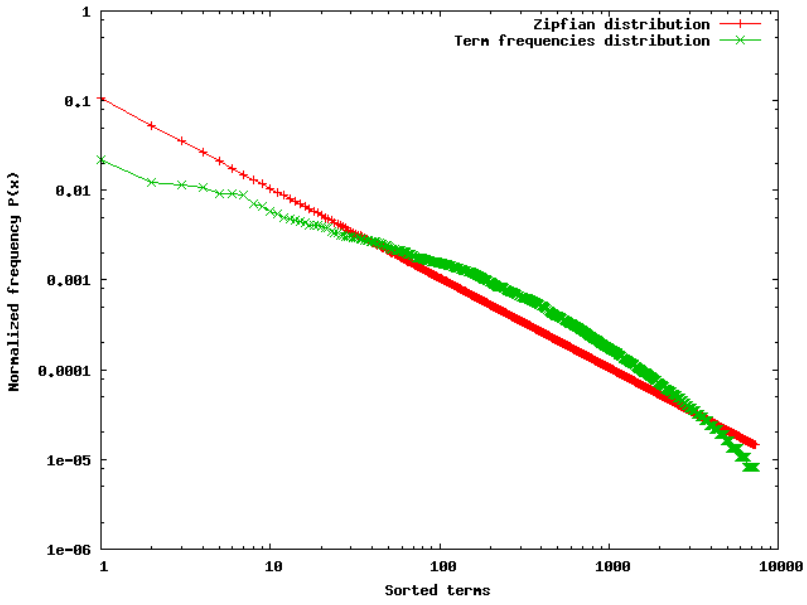


Figure B.16: Stylometry: All term frequency distribution of the *WebKB train* corpus

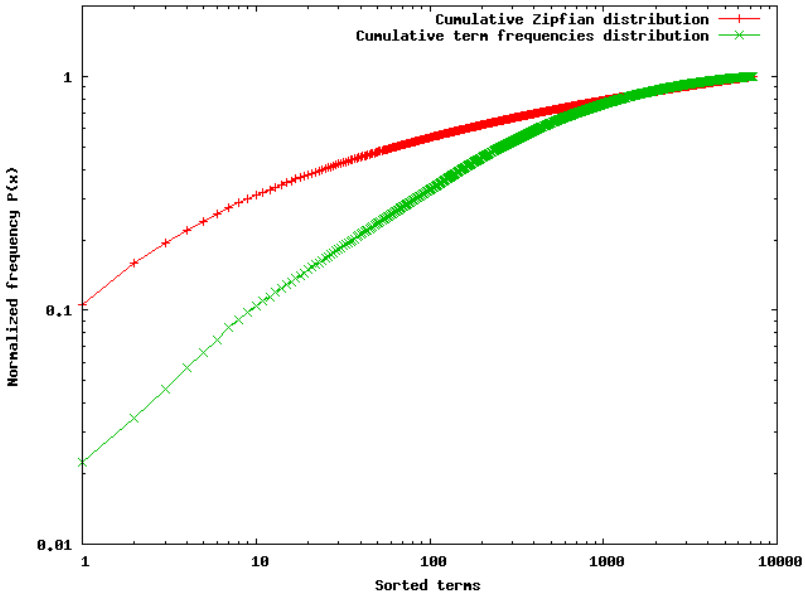


Figure B.17: Stylometry: All term cumulative frequency distribution of the *WebKB train* corpus

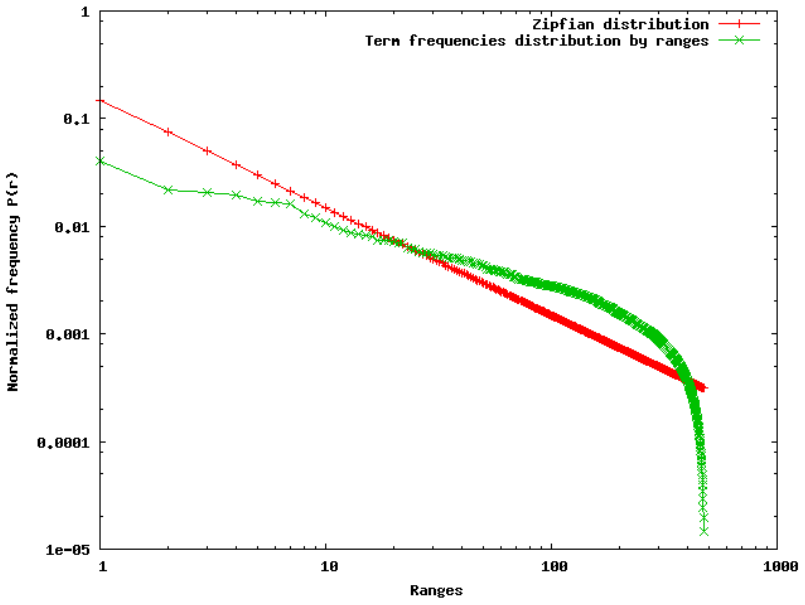


Figure B.18: Stylometry: Range frequency distribution of the *WebKB train* corpus

B.4 The *WebKB test* corpus

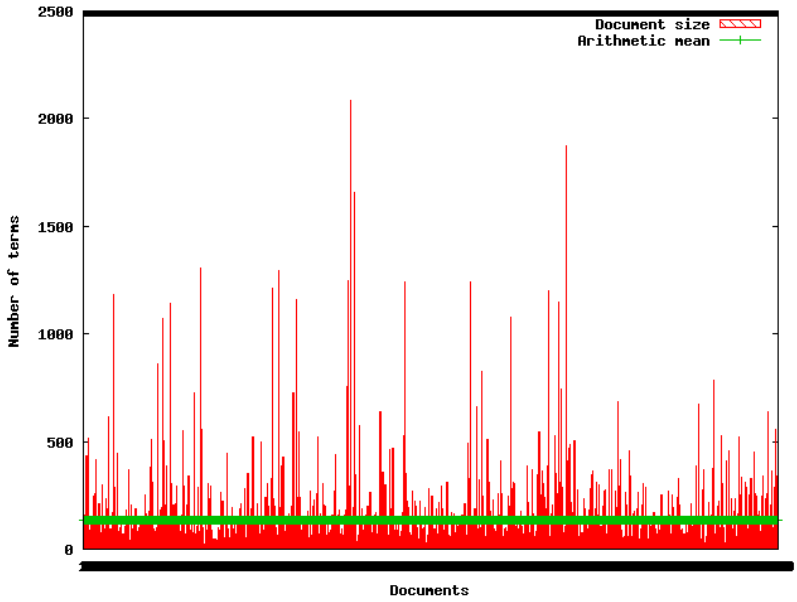


Figure B.19: Document cardinalities of the *WebKB test* corpus

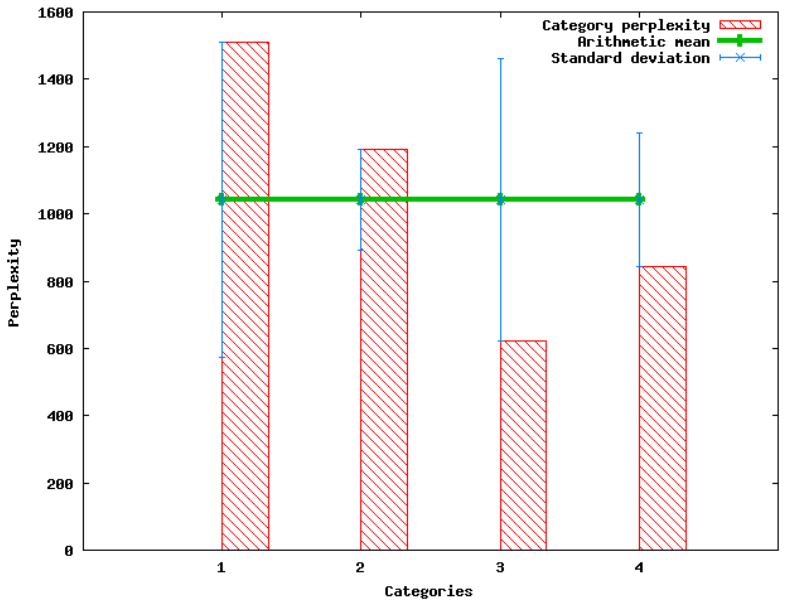


Figure B.20: Perplexity per category of the *WebKB test* corpus

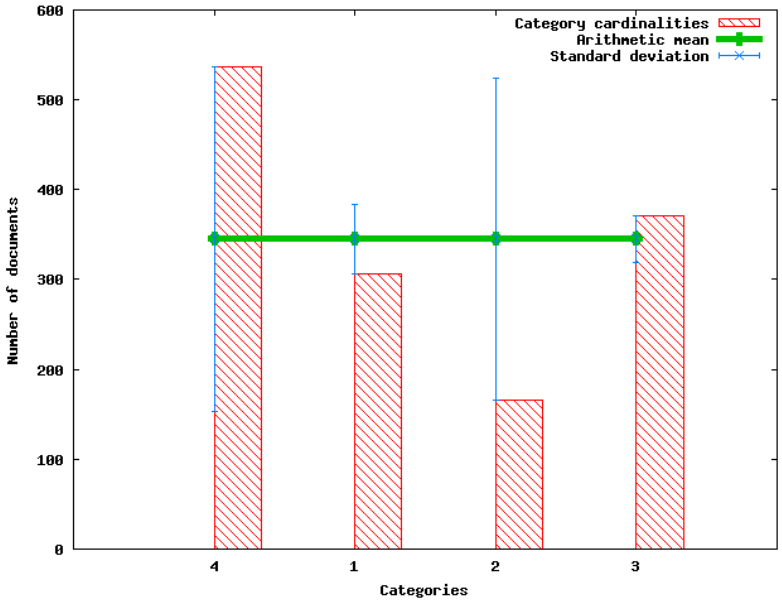


Figure B.21: Imbalance per category of the *WebKB test* corpus

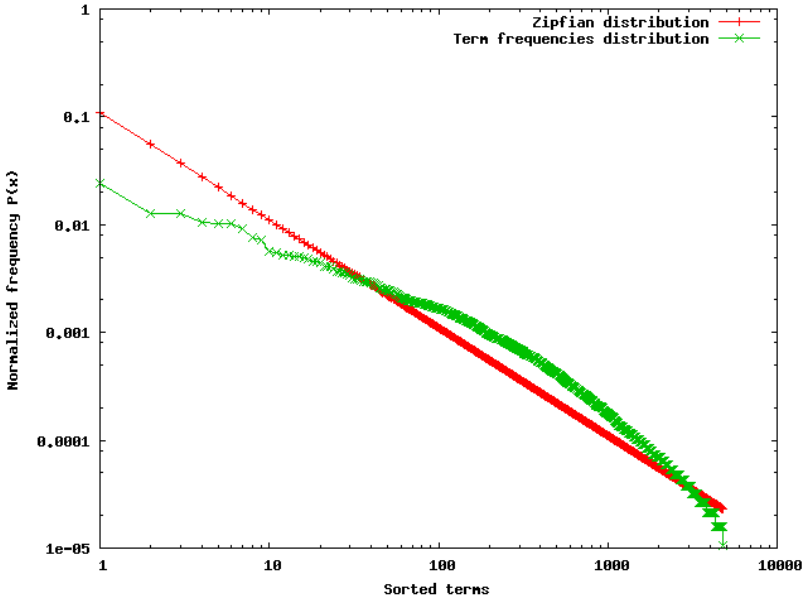


Figure B.22: Stylometry: All term frequency distribution of the *WebKB test* corpus

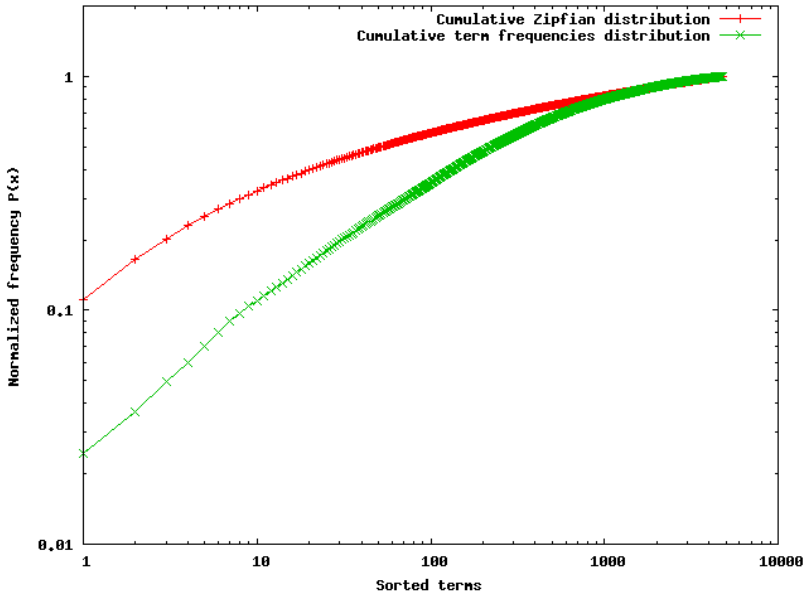


Figure B.23: Stylometry: All term cumulative frequency distribution of the *WebKB test* corpus

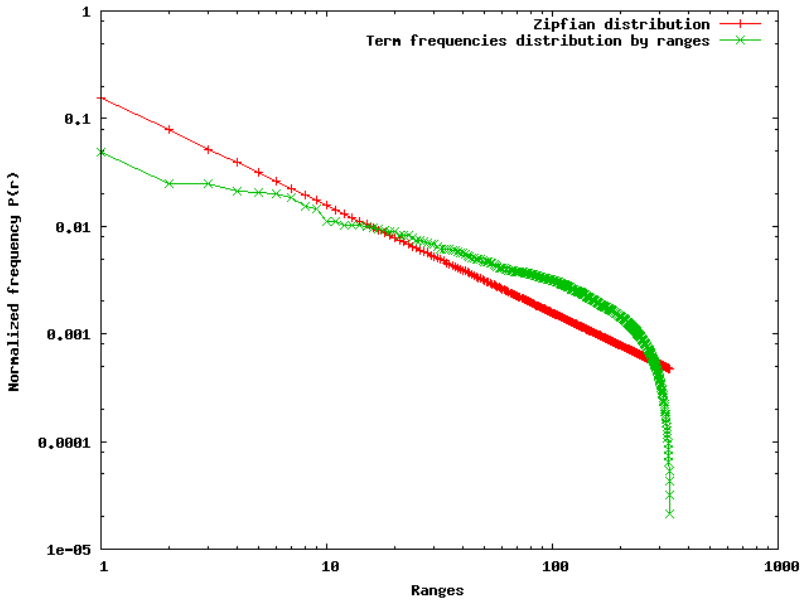


Figure B.24: Stylometry: Range frequency distribution of the *WebKB test* corpus

B.5 The *R8-Reuters train* corpus

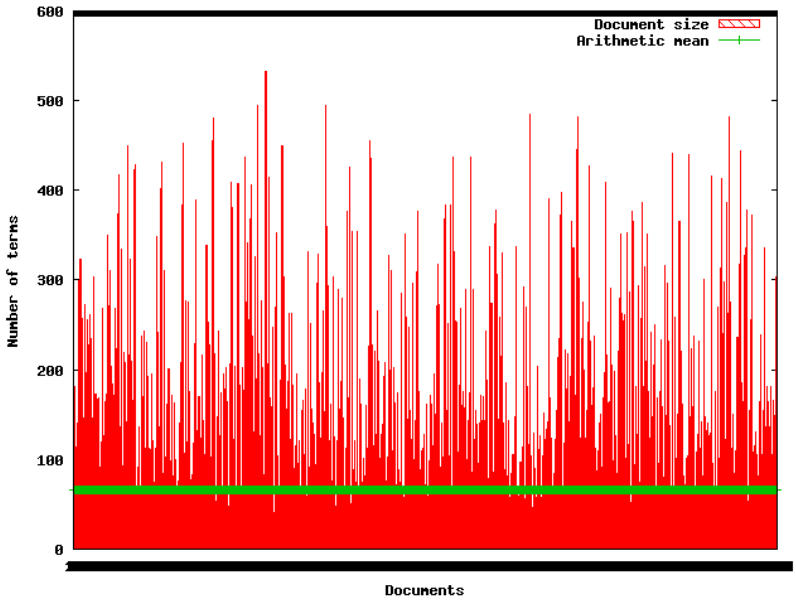


Figure B.25: Document cardinalities of the *R8-Reuters train* corpus

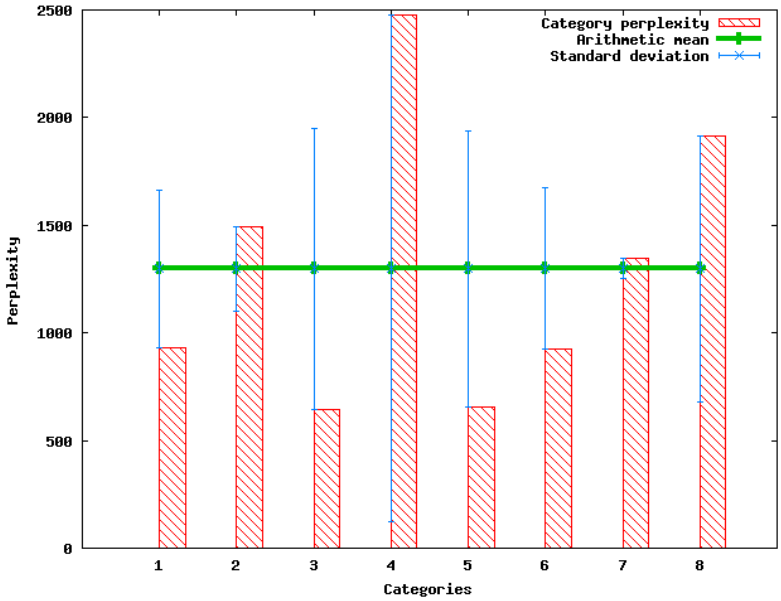


Figure B.26: Perplexity per category of the *R8-Reuters train* corpus

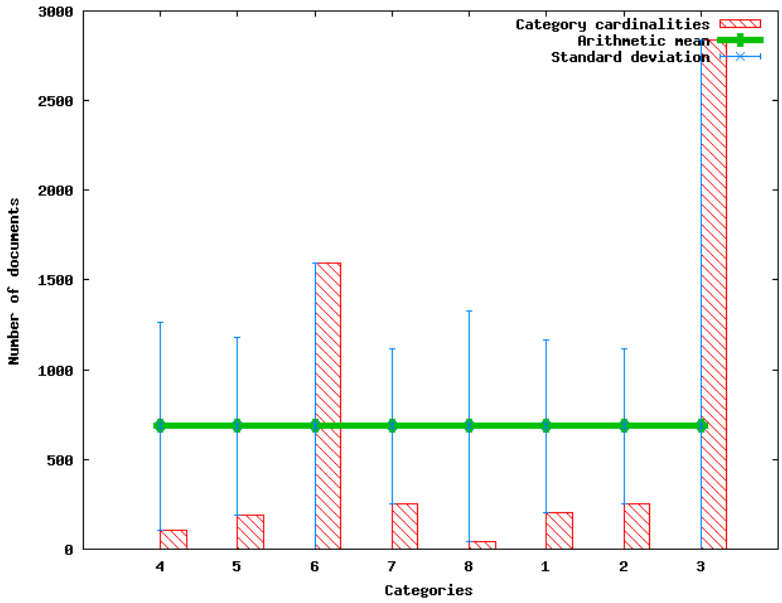


Figure B.27: Imbalance per category of the *R8-Reuters train* corpus

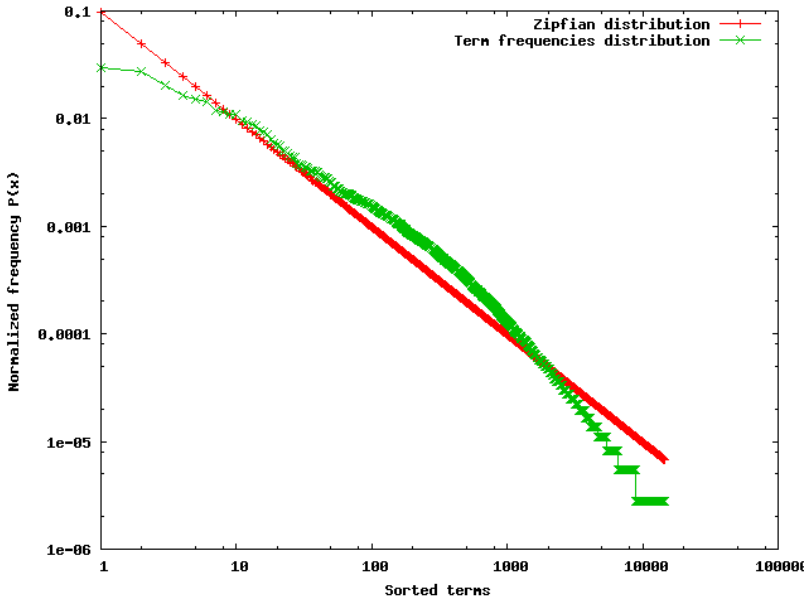


Figure B.28: Stylometry: All term frequency distribution of the *R8-Reuters train* corpus

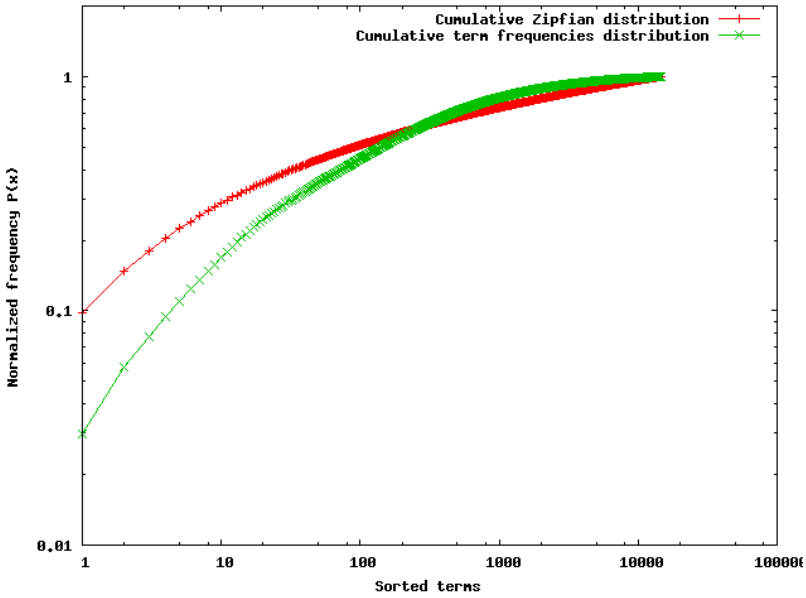


Figure B.29: Stylometry: All term cumulative frequency distribution of the *R8-Reuters train* corpus

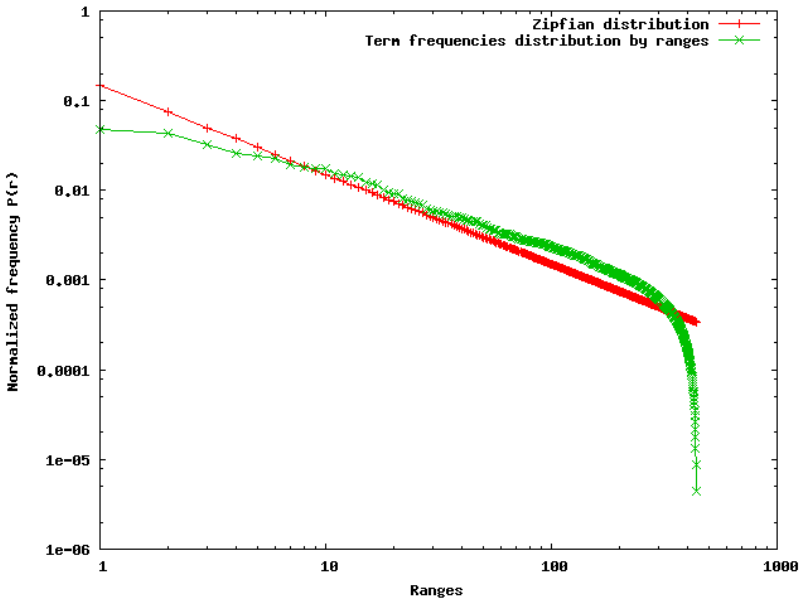


Figure B.30: Stylometry: Range frequency distribution of the *R8-Reuters train* corpus

B.6 The *R8-Reuters test* corpus

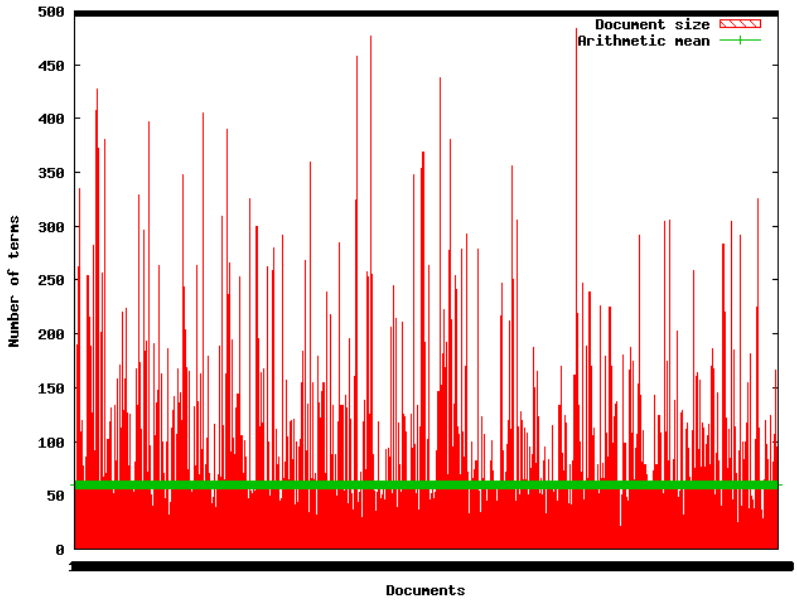


Figure B.31: Document cardinalities of the *R8-Reuters test* corpus

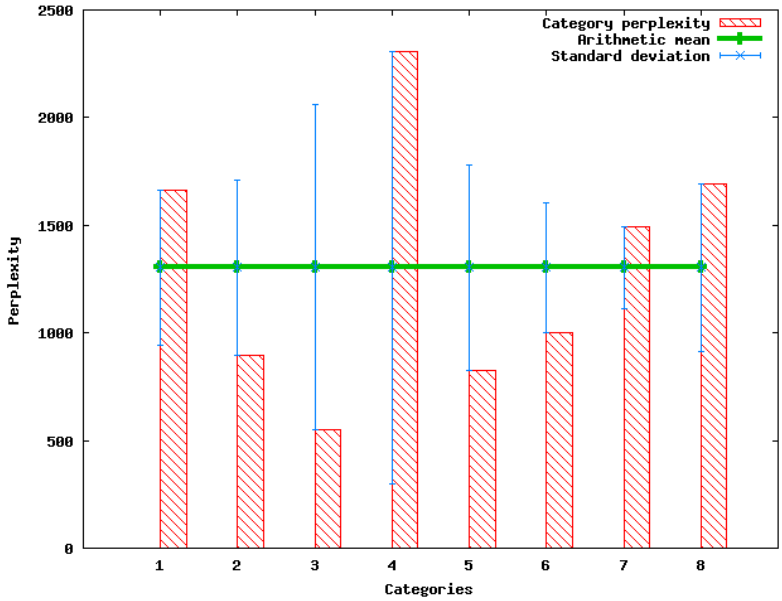


Figure B.32: Perplexity per category of the *R8-Reuters test* corpus

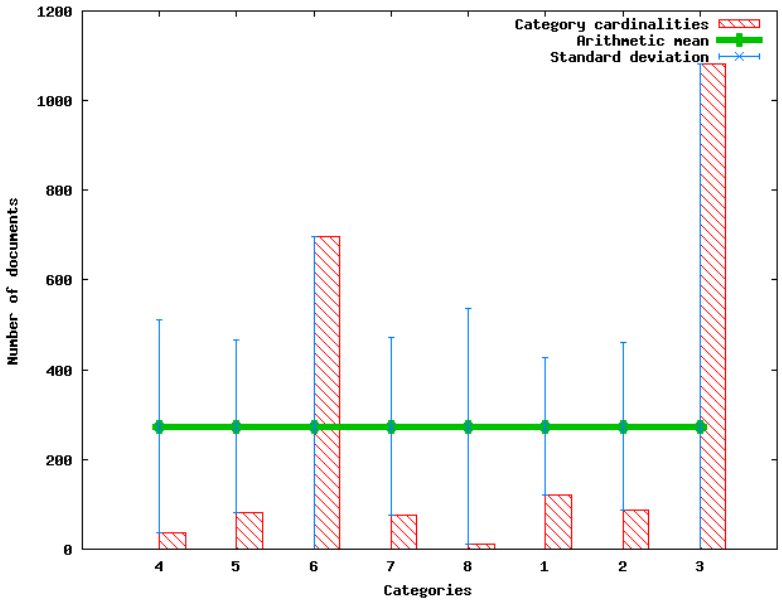


Figure B.33: Imbalance per category of the *R8-Reuters test* corpus

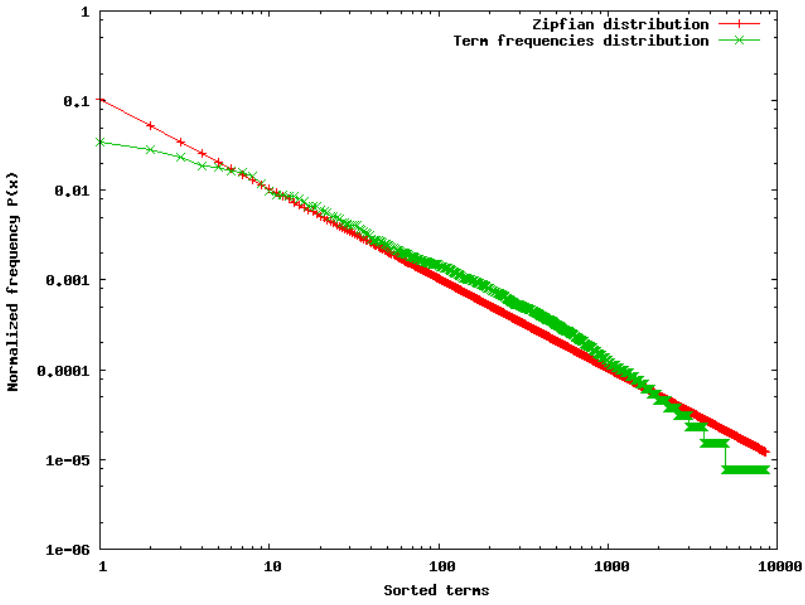


Figure B.34: Stylometry: All term frequency distribution of the *R8-Reuters test* corpus

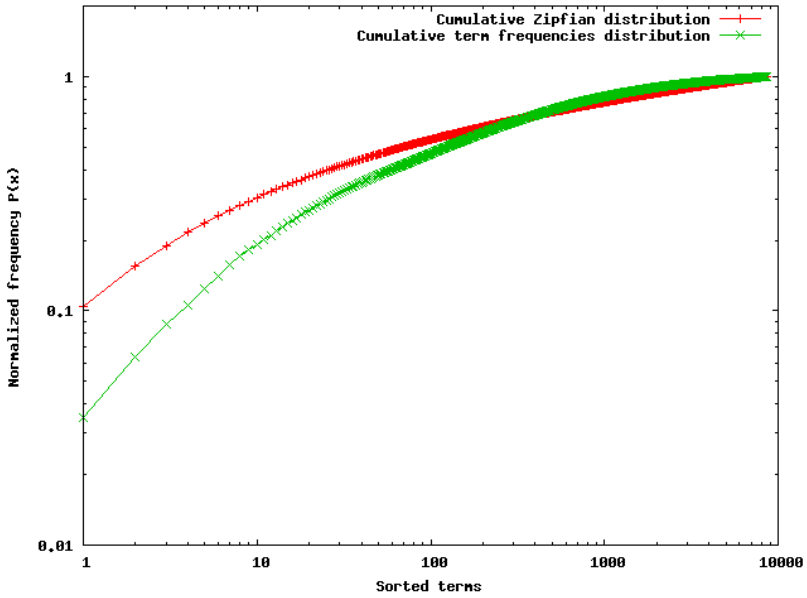


Figure B.35: Stylometry: All term cumulative frequency distribution of the *R8-Reuters test* corpus

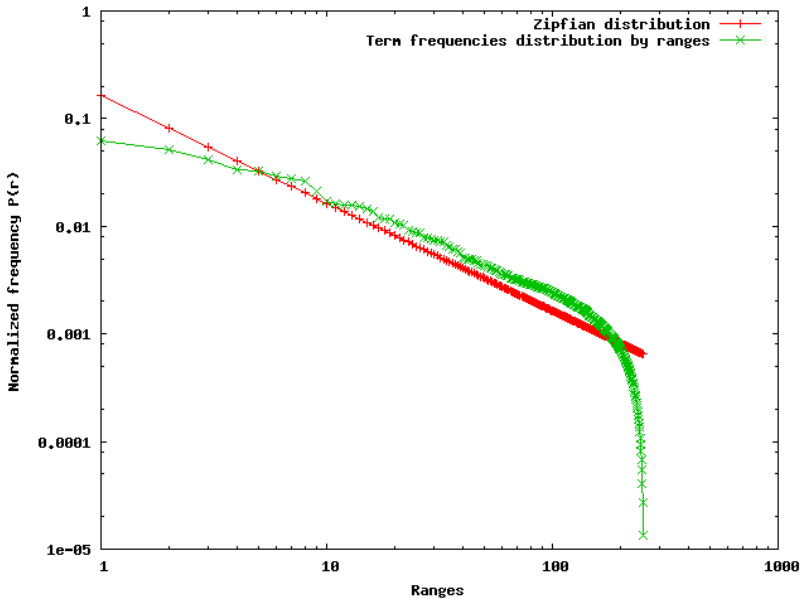


Figure B.36: Stylometry: Range frequency distribution of the *R8-Reuters test* corpus

B.7 The *R52-Reuters train* corpus

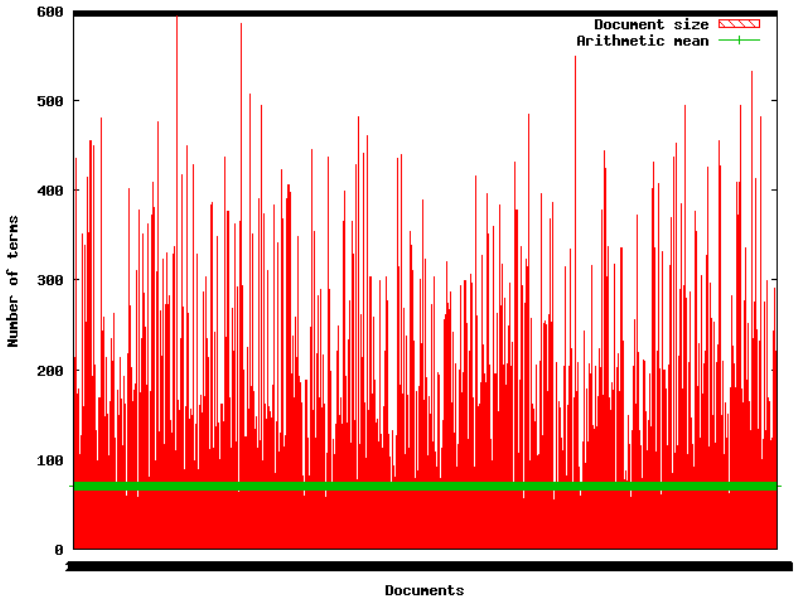


Figure B.37: Document cardinalities of the *R52-Reuters train* corpus

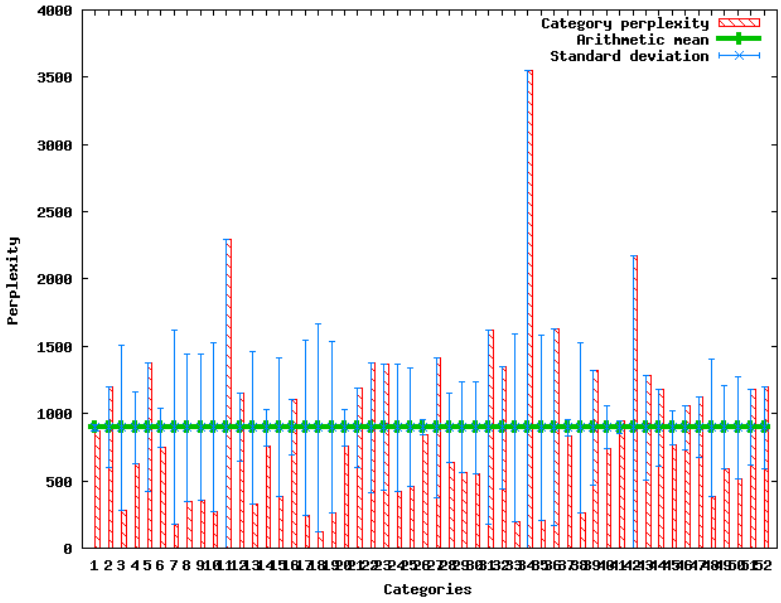


Figure B.38: Perplexity per category of the *R52-Reuters train* corpus

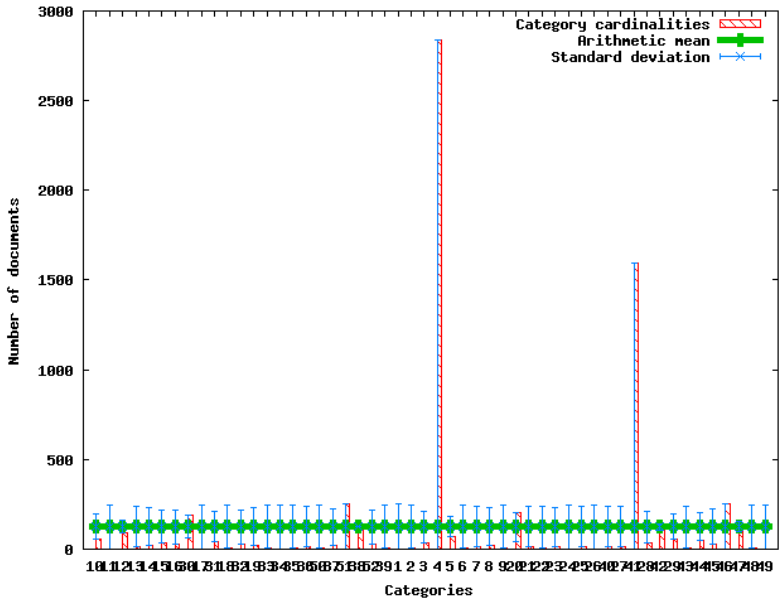


Figure B.39: Imbalance per category of the *R52-Reuters train* corpus

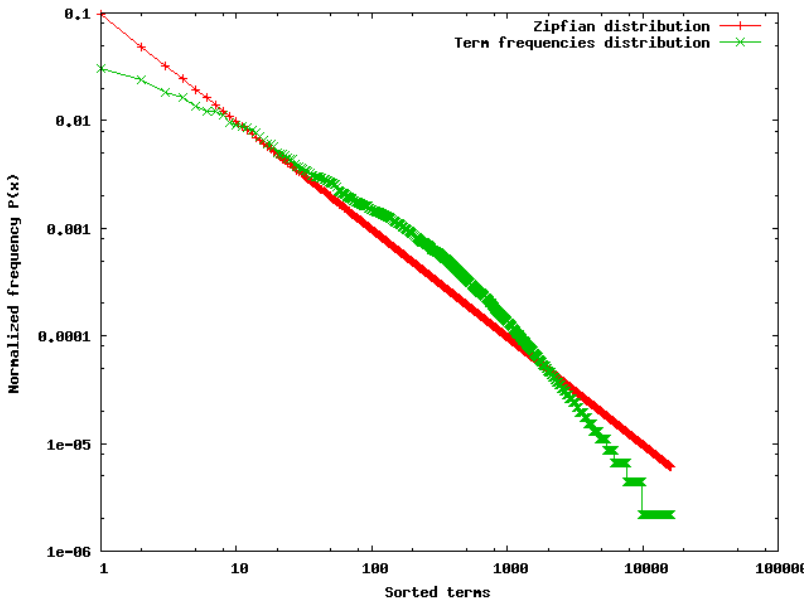


Figure B.40: Stylometry: All term frequency distribution of the *R52-Reuters train* corpus

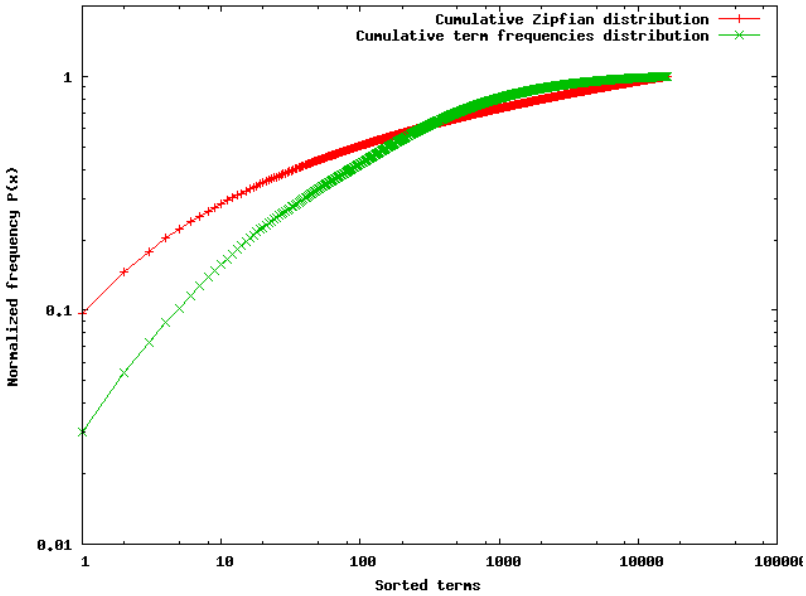


Figure B.41: Stylometry: All term cumulative frequency distribution of the *R52-Reuters train* corpus

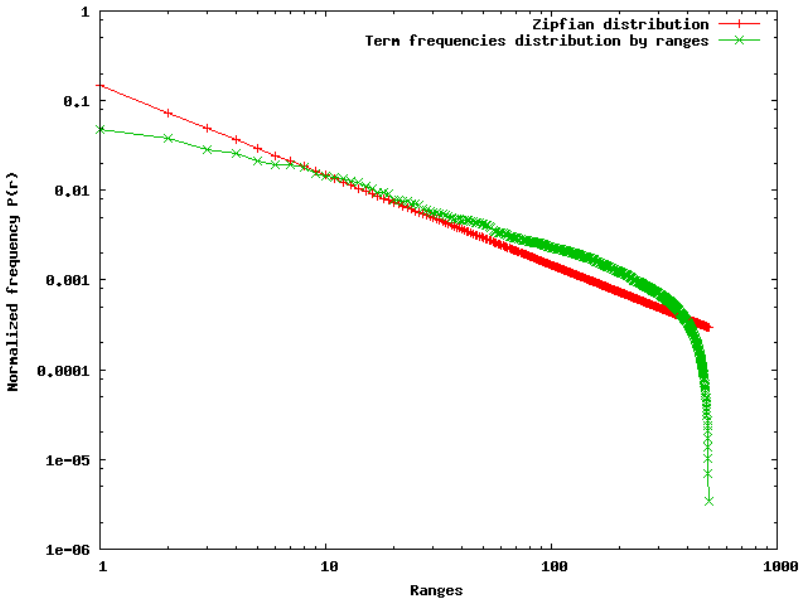


Figure B.42: Stylometry: Range frequency distribution of the *R52-Reuters train* corpus

B.8 The *R52-Reuters test* corpus

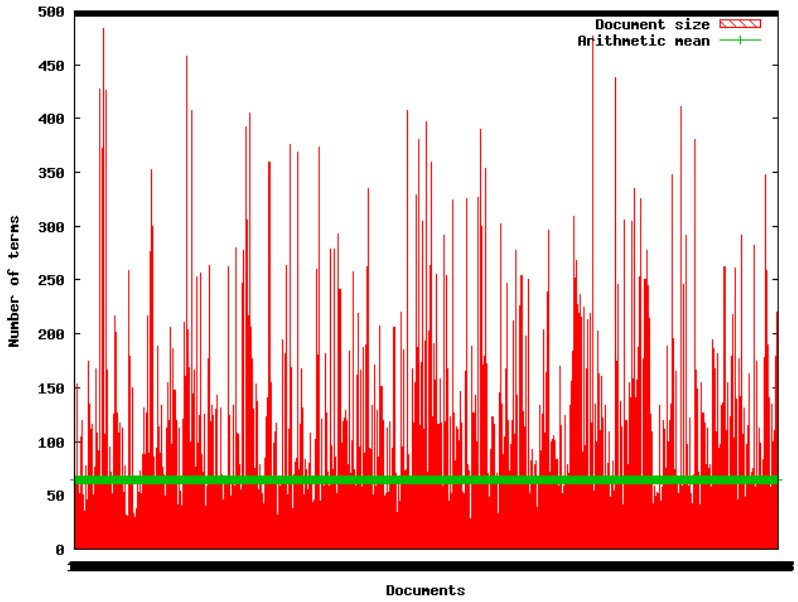


Figure B.43: Document cardinalities of the *R52-Reuters test* corpus

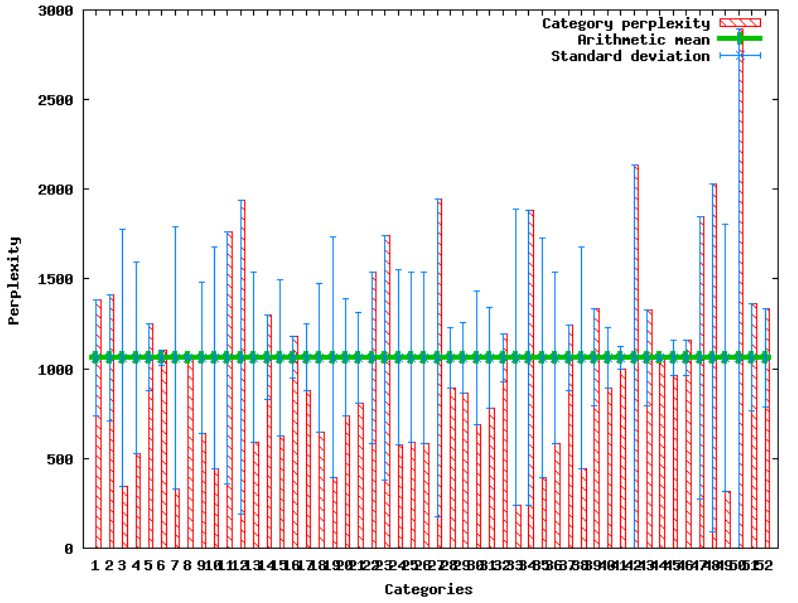


Figure B.44: Perplexity per category of the *R52-Reuters test* corpus

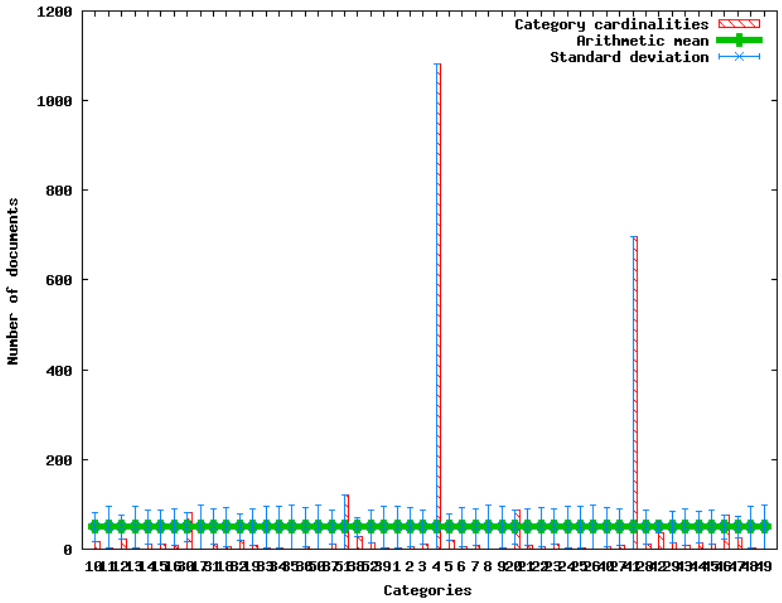


Figure B.45: Imbalance per category of the *R52-Reuters test* corpus

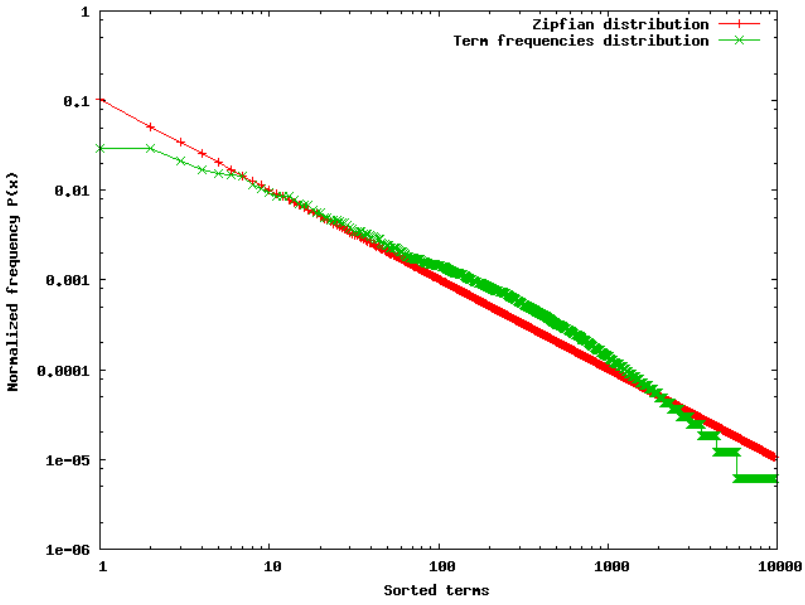


Figure B.46: Stylometry: All term frequency distribution of the *R52-Reuters test* corpus

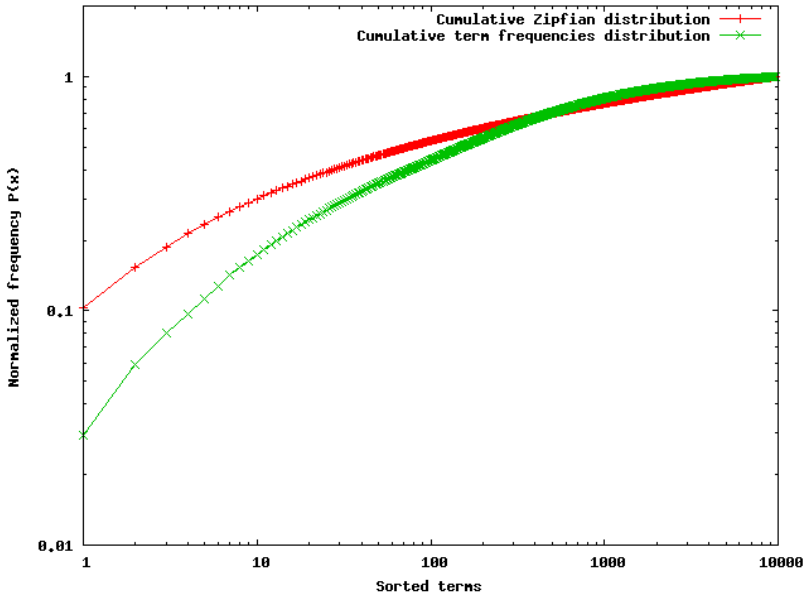


Figure B.47: Stylometry: All term cumulative frequency distribution of the *R52-Reuters test* corpus

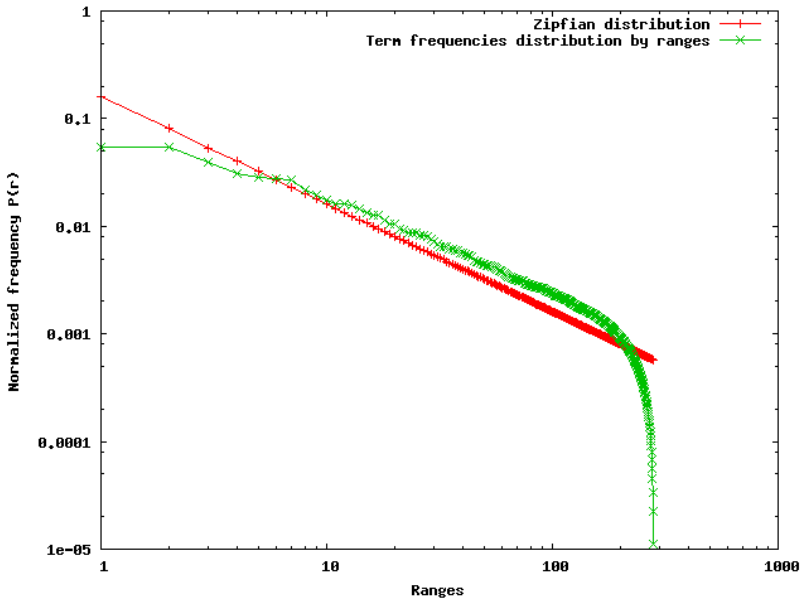


Figure B.48: Stylometry: Range frequency distribution of the *R52-Reuters test* corpus

B.9 The *20 Newsgroups train* corpus

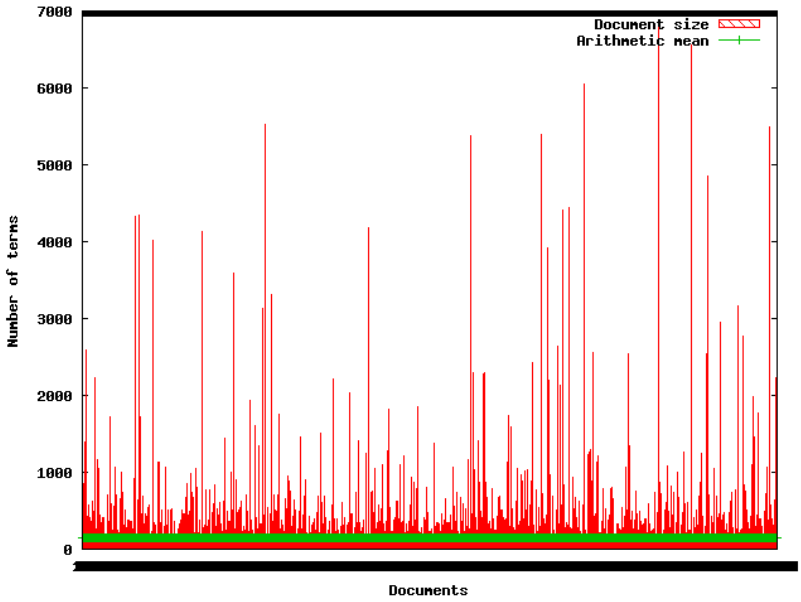


Figure B.49: Document cardinalities of the *20 Newsgroups train* corpus

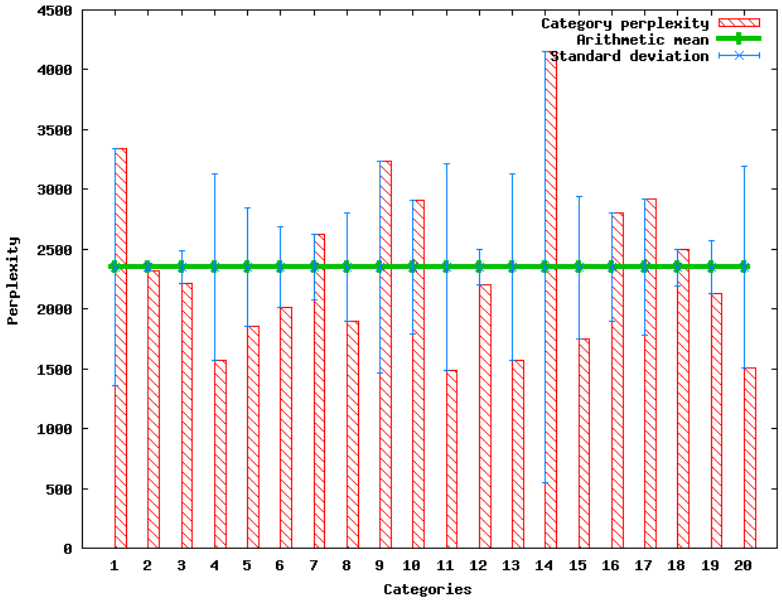


Figure B.50: Perplexity per category of the *20 Newsgroups train* corpus

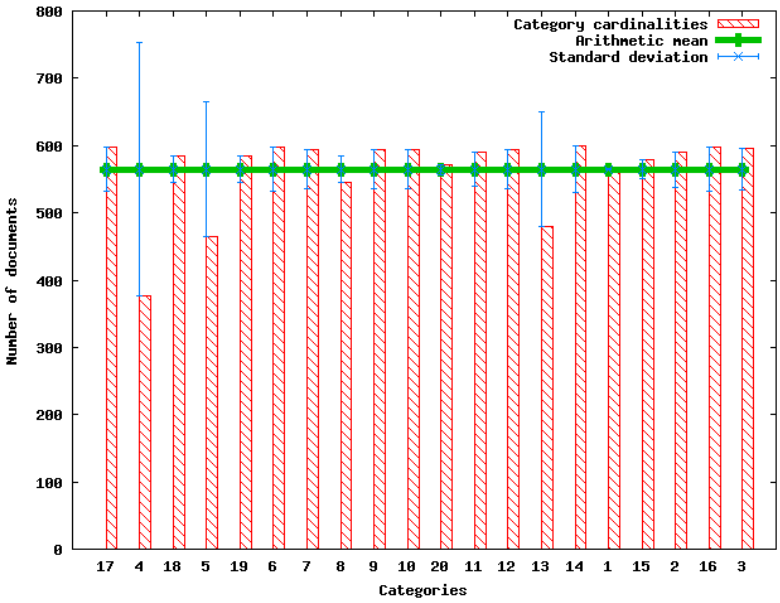


Figure B.51: Imbalance per category of the *20 Newsgroups train* corpus

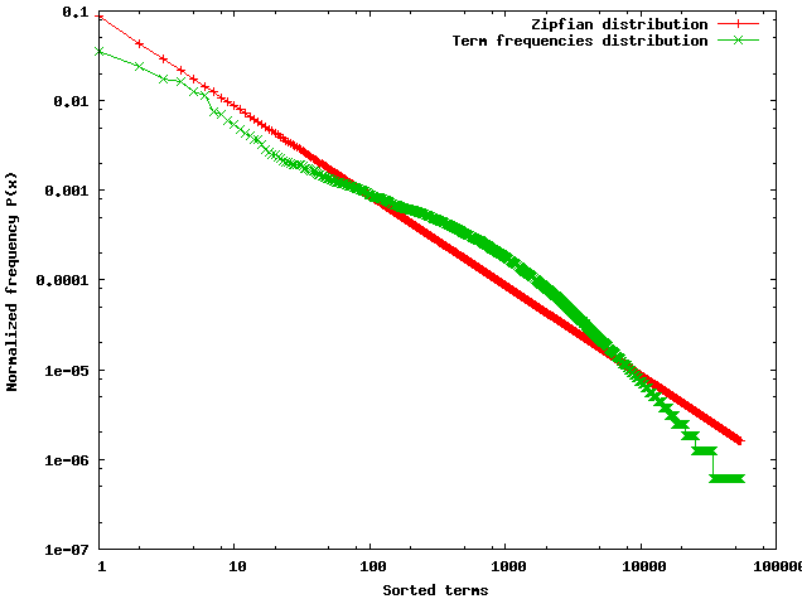


Figure B.52: Stylometry: All term frequency distribution of the *20 Newsgroups train* corpus

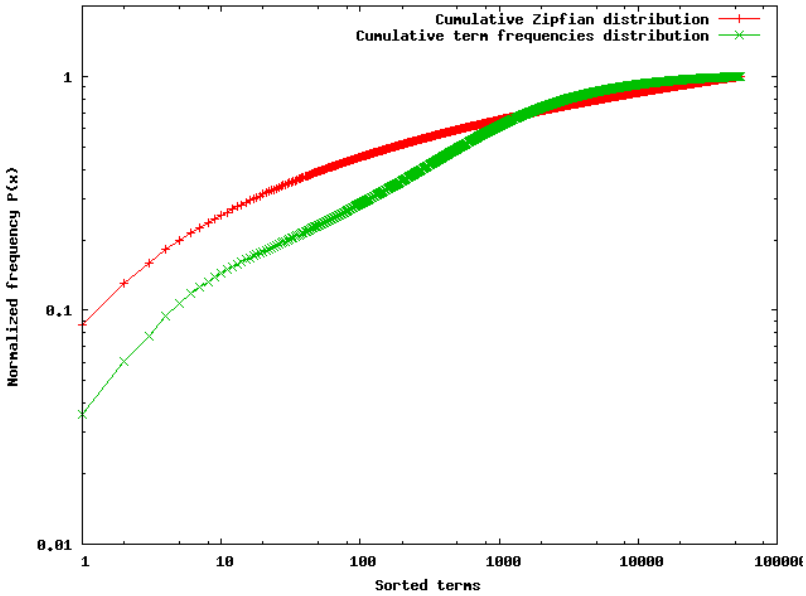


Figure B.53: Stylometry: All term cumulative frequency distribution of the *20 Newsgroups train* corpus

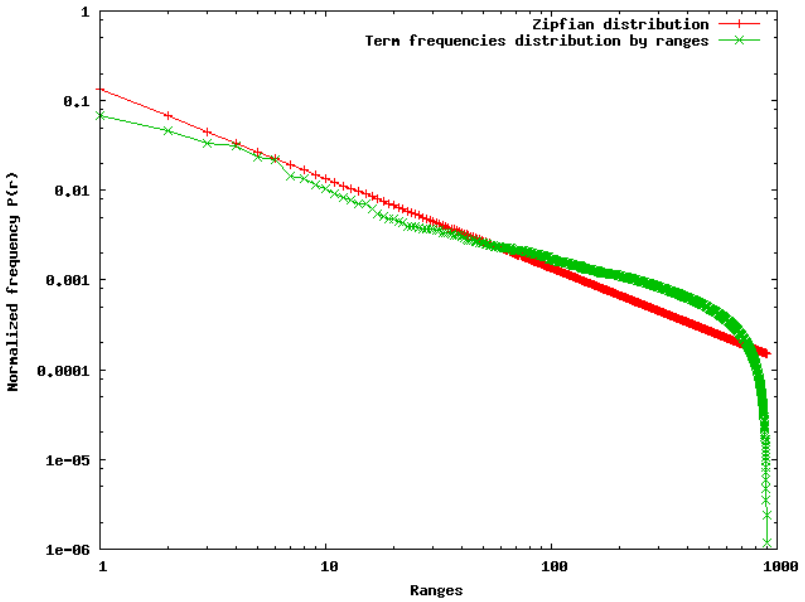


Figure B.54: Stylometry: Range frequency distribution of the *20 Newsgroups train* corpus

B.10 The *20 Newsgroups test* corpus

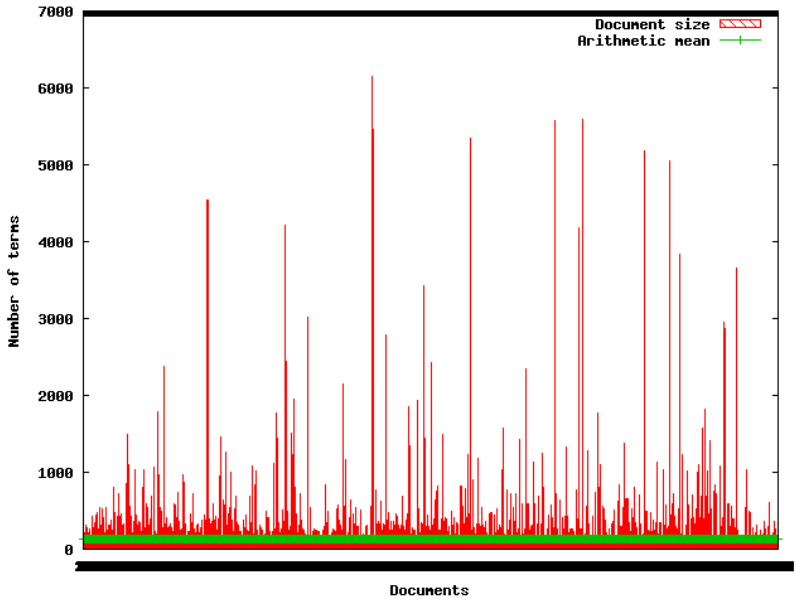


Figure B.55: Document cardinalities of the *20 Newsgroups test* corpus

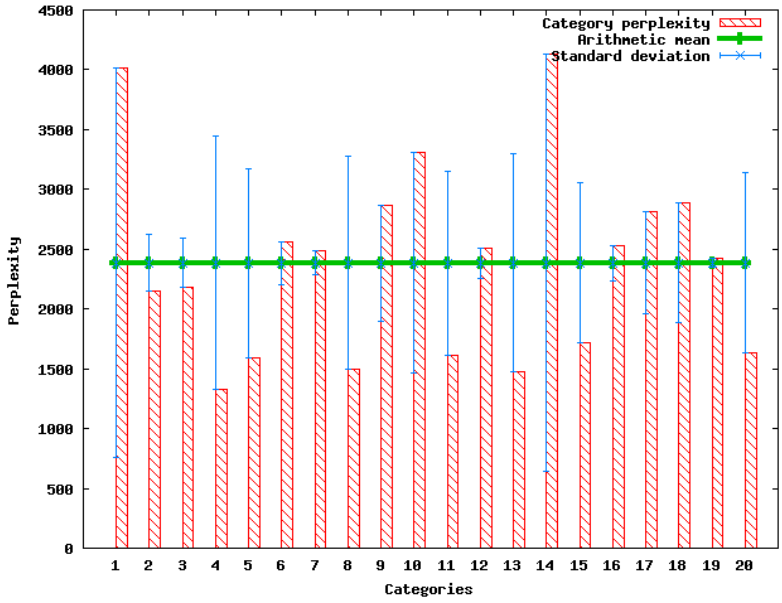


Figure B.56: Perplexity per category of the *20 Newsgroups test* corpus

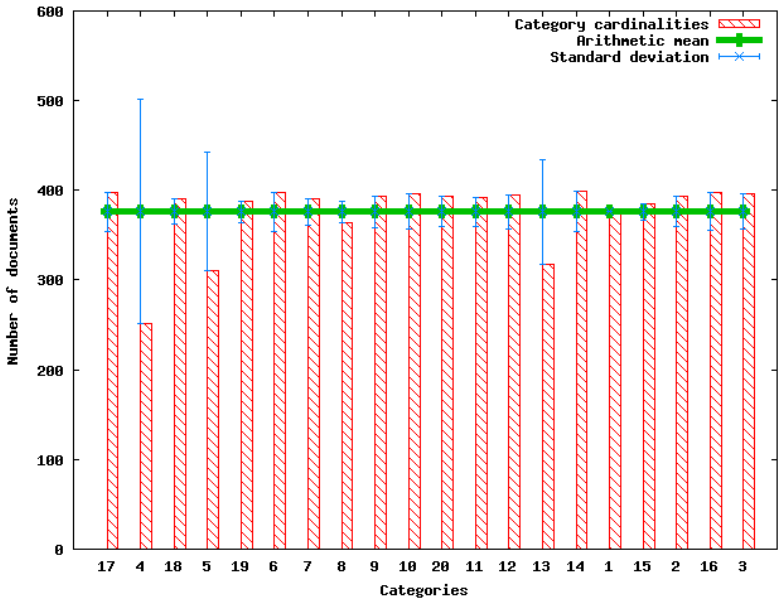


Figure B.57: Imbalance per category of the *20 Newsgroups test* corpus

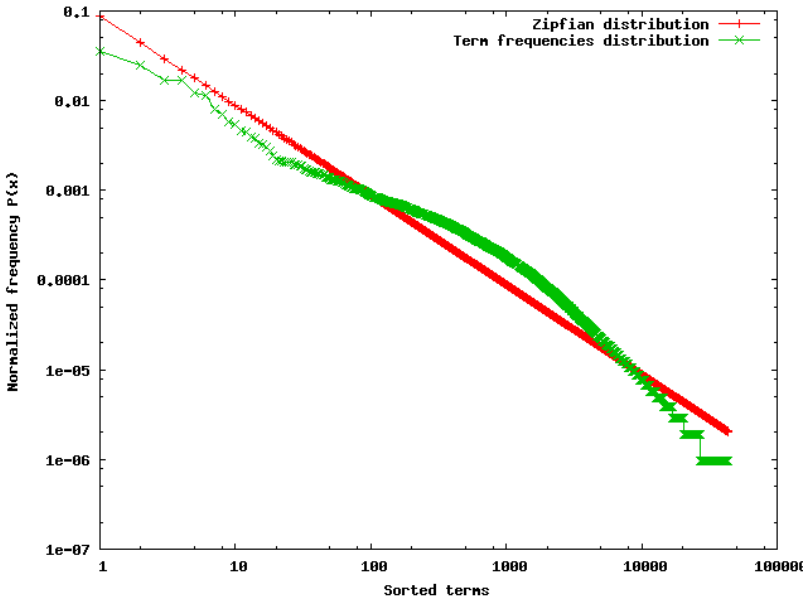


Figure B.58: Stylometry: All term frequency distribution of the *20 Newsgroups test* corpus

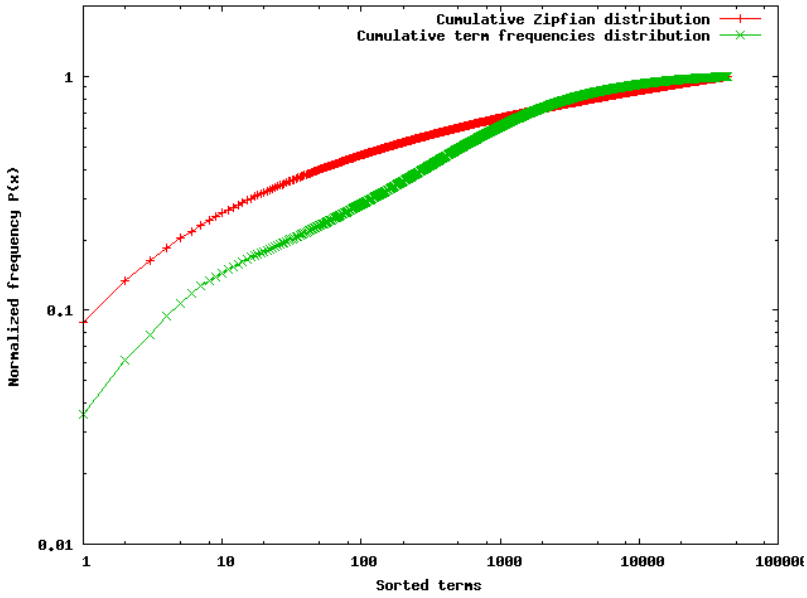


Figure B.59: Stylometry: All term cumulative frequency distribution of the *20 Newsgroups test* corpus

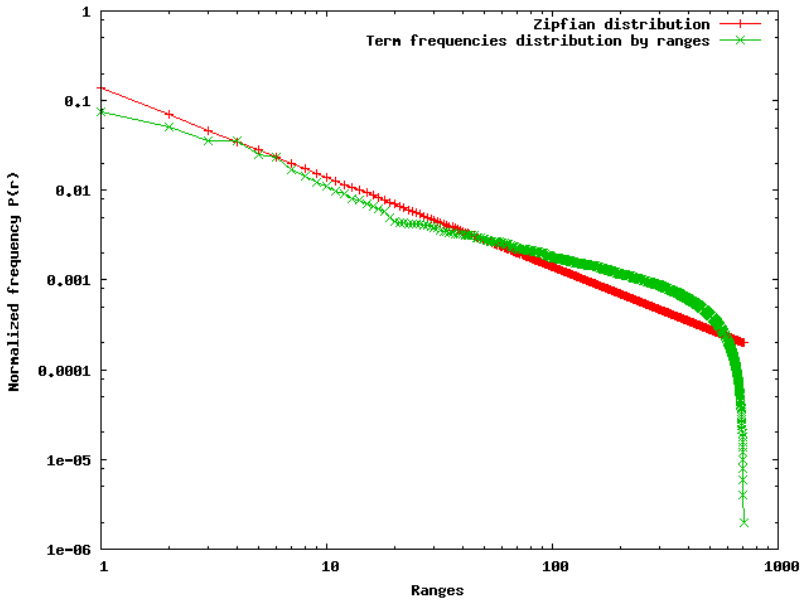


Figure B.60: Stylometry: Range frequency distribution of the *20 Newsgroups test* corpus

Appendix C

Word by word analysis in the WSI-SemEval data collection

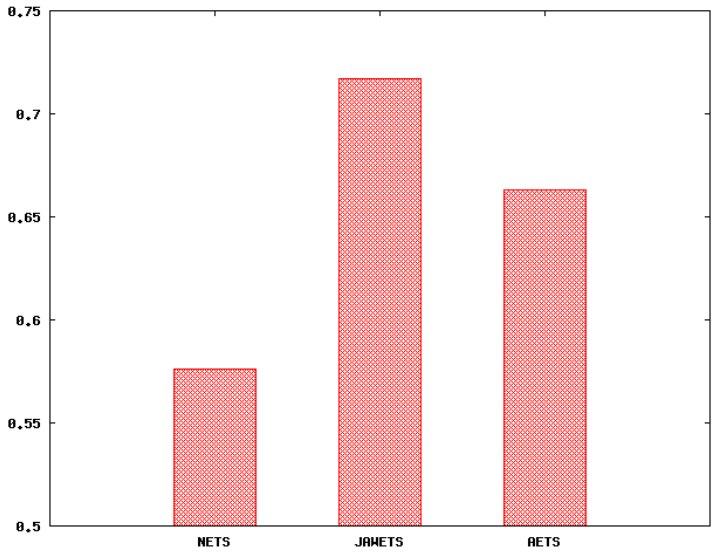


Figure C.1: Effect of the Self-term expansion technique on the *WSI-SemEval* data collection

Figure C.1 presents the arithmetic mean for the three approaches that evaluated the WSI-Semeval data collection, namely NETS, JAWETS & AETS. The average was calculated over 100 corpora. Each dataset refers to one ambiguous word and the word by word obtained results is shown in Tables C.1 and C.2.

Table C.1: Word by word results with the WSI-SemEval data collection (first 50 ambiguous words)

Ambiguous word	**Gold class**	**AETS clusters**	**JAWETS clusters**	**NETS clusters**	**AETS FScore**	**JAWETS FScore**	**NETS FScore**
affect.v	1	3	3	5	0.882	0.914	0.538
allow.v	2	6	3	6	0.530	0.882	0.559
announce.v	1	3	2	5	0.919	0.974	0.710
approve.v	2	4	2	3	0.689	1.000	0.675
area.n	3	9	2	8	0.584	0.662	0.572
ask.v	5	9	2	8	0.610	0.587	0.543
attempt.v	1	1	4	3	1.000	0.571	0.824
authority.n	3	3	2	4	0.581	0.646	0.588
avoid.v	1	3	2	4	0.857	0.815	0.720
base.n	4	3	1	4	0.771	0.588	0.756
begin.v	2	6	3	6	0.561	0.595	0.534
believe.v	2	10	2	5	0.528	0.756	0.536
bill.n	3	7	8	8	0.925	0.640	0.799
build.v	3	4	4	6	0.684	0.677	0.528
buy.v	5	6	2	6	0.644	0.740	0.469
capital.n	2	6	3	8	0.733	0.820	0.499
care.v	2	2	1	4	0.662	0.722	0.694
carrier.n	2	3	1	2	0.974	0.722	0.819
cause.v	1	7	4	5	0.747	0.779	0.676
chance.n	3	2	1	2	0.544	0.557	0.604
claim.v	3	2	2	3	0.738	0.719	0.509
come.v	9	6	2	7	0.359	0.330	0.369
complain.v	2	3	2	4	0.582	0.671	0.582
complete.v	2	4	2	5	0.774	0.902	0.673
condition.n	2	5	1	7	0.766	0.752	0.755
contribute.v	2	6	3	4	0.610	0.526	0.500
defense.n	6	3	3	4	0.370	0.474	0.535
describe.v	1	3	1	5	0.882	1.000	0.774
development.n	2	5	4	5	0.625	0.571	0.500
disclose.v	2	4	3	5	0.690	0.635	0.634
do.v	2	9	5	10	0.659	0.673	0.612
drug.n	2	4	4	6	0.873	0.836	0.786
effect.n	2	5	3	5	0.807	0.712	0.608
end.v	3	5	3	8	0.550	0.419	0.491
enjoy.v	2	5	1	6	0.469	0.673	0.694
estimate.v	1	6	2	6	0.476	0.897	0.667
examine.v	1	3	2	2	0.500	0.800	0.800
exchange.n	3	10	2	9	0.692	0.668	0.655
exist.v	1	4	2	6	0.842	0.900	0.483
explain.v	2	4	3	4	0.743	0.643	0.690
express.v	1	2	2	2	0.947	0.824	0.824
feel.v	2	7	3	10	0.566	0.539	0.534
find.v	3	6	5	6	0.701	0.584	0.585
fix.v	2	2	2	2	1.000	1.000	1.000
future.n	3	8	5	12	0.822	0.847	0.618
go.v	9	9	3	10	0.405	0.384	0.394
grant.v	2	3	2	3	0.667	0.680	0.667
hold.v	8	6	4	7	0.543	0.420	0.537
hope.v	1	5	1	5	0.900	1.000	0.731
hour.n	2	6	8	7	0.636	0.688	0.485

Table C.2: Word by word results with the WSI-SemEval data collection (last 50 ambiguous words)

Ambiguous word	**Gold class**	**AETS clusters**	**JAWETS clusters**	**NETS clusters**	**AETS FScore**	**JAWETS FScore**	**NETS FScore**
improve.v	1	2	4	3	0.933	0.720	0.857
job.n	3	5	2	5	0.529	0.697	0.540
join.v	4	6	4	4	0.543	0.469	0.440
keep.v	7	6	2	9	0.514	0.466	0.419
kill.v	2	6	2	5	0.578	0.709	0.578
lead.v	6	4	5	6	0.467	0.401	0.399
maintain.v	2	2	2	4	0.820	0.820	0.743
management.n	2	6	2	5	0.681	0.663	0.735
move.n	2	7	4	6	0.876	0.850	0.699
need.v	2	8	5	8	0.567	0.571	0.532
negotiate.v	1	2	4	4	0.941	0.800	0.800
network.n	3	7	10	6	0.819	0.728	0.763
occur.v	3	4	4	3	0.696	0.635	0.558
order.n	4	6	2	7	0.860	0.652	0.641
part.n	3	9	2	9	0.631	0.629	0.448
people.n	2	13	9	10	0.674	0.529	0.635
plant.n	2	6	4	9	0.910	0.936	0.867
point.n	5	11	2	13	0.517	0.774	0.592
policy.n	2	7	4	6	0.662	0.725	0.485
position.n	5	7	2	8	0.559	0.464	0.568
power.n	3	7	3	6	0.538	0.586	0.617
prepare.v	2	2	3	4	0.714	0.794	0.672
president.n	3	14	4	14	0.531	0.696	0.756
produce.v	3	8	4	9	0.500	0.646	0.457
promise.v	2	2	1	3	0.688	0.743	0.550
propose.v	2	3	2	4	0.881	0.792	0.599
prove.v	2	4	2	5	0.606	0.622	0.575
purchase.v	1	2	3	5	0.846	0.800	0.800
raise.v	7	5	4	4	0.365	0.403	0.428
rate.n	2	9	13	13	0.613	0.679	0.547
recall.v	2	5	3	3	0.493	0.549	0.737
receive.v	2	8	3	7	0.693	0.785	0.465
regard.v	2	2	2	3	0.692	0.653	0.717
remember.v	1	3	2	4	0.818	0.870	0.762
remove.v	1	1	2	4	1.000	0.970	0.741
replace.v	1	2	2	2	0.929	0.800	0.800
report.v	2	8	3	7	0.629	0.850	0.551
rush.v	1	1	2	2	1.000	0.923	0.727
say.v	4	10	11	15	0.671	0.939	0.492
see.v	5	7	2	5	0.470	0.486	0.437
set.v	6	5	4	7	0.375	0.313	0.345
share.n	2	20	11	26	0.800	0.909	0.602
source.n	5	6	3	10	0.576	0.444	0.474
space.n	2	3	4	4	0.608	0.714	0.833
start.v	6	6	3	7	0.465	0.507	0.328
state.n	3	10	7	10	0.580	0.635	0.478
system.n	4	12	8	8	0.463	0.480	0.423
turn.v	11	8	5	8	0.374	0.359	0.338
value.n	2	8	5	6	0.651	0.577	0.496
work.v	6	5	2	4	0.461	0.542	0.544
Average over 100 words	2.87	5.57	3.35	6.16	0.663	0.717	0.576

In Figure C.2 we may see all the ambiguous words for which both, the AETS and JAWETS approach obtained better performance than the NETS approach. Figure C.3 shows those ambiguous words where at least one of the expanded versions ouperformed the unexpanded one. Finally, in Figure C.4 we may observe the ambiguous words to which none of the self-term expansion approaches outperformed the NETS approach.

In conclusion, the self-term expansion outperformed the NETS approach for the 87% of corpora provided in the word sense induction task of the SemEval 2007.

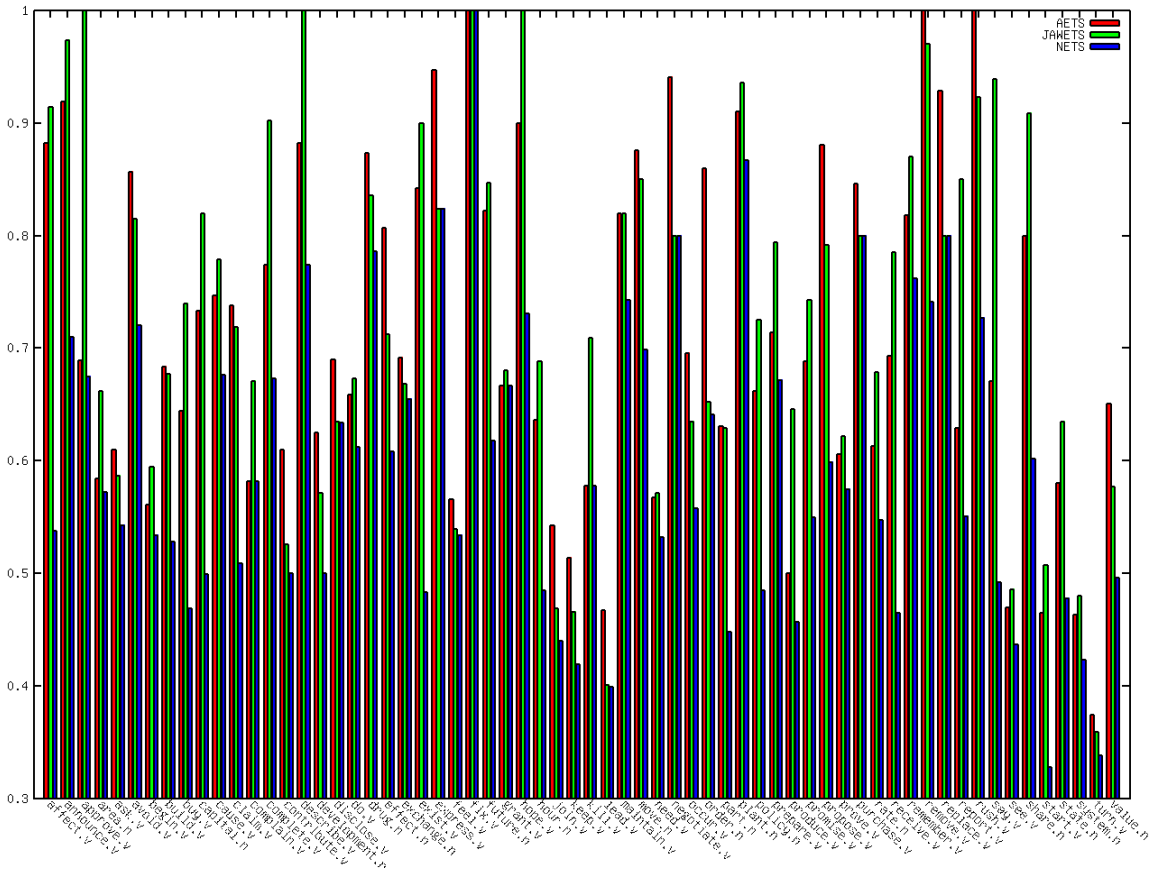


Figure C.2: Ambiguous words for which both AETS and JAWETS obtained better performance than the NETS approach

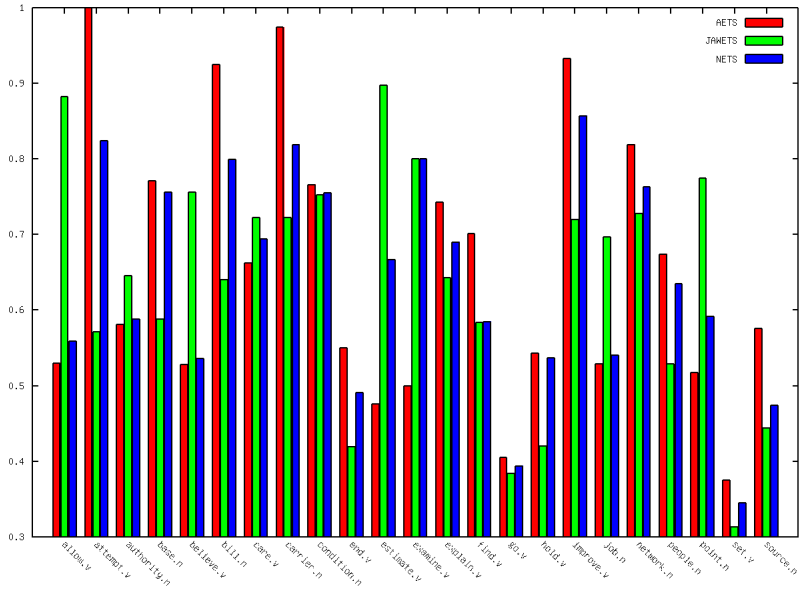


Figure C.3: Ambiguous words for which either AETS or JAWETS obtained better performance than the NETS approach

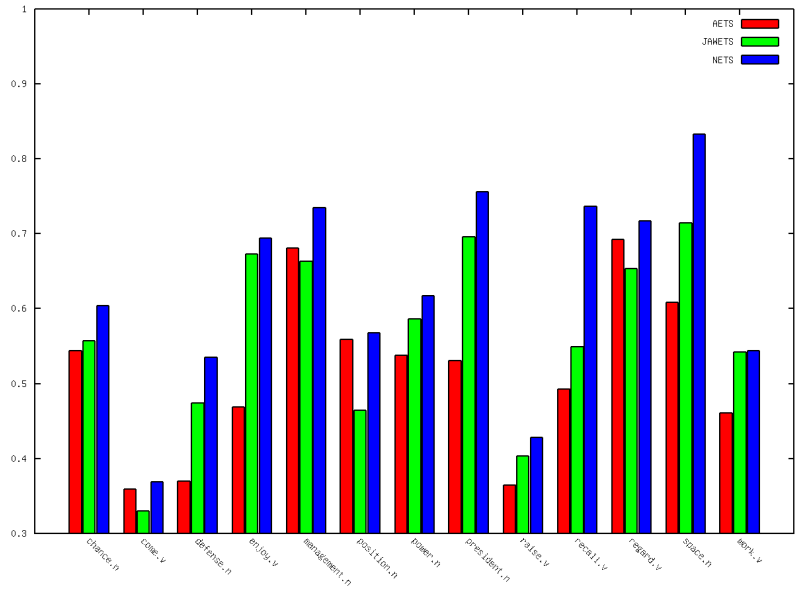


Figure C.4: Ambiguous words for which none of AETS and JAWETS obtained better performance than the NETS approach