Robust Document Clustering by Exploiting Feature Diversity in Cluster Ensembles

Xavier Sevillano, Germán Cobo, Francesc Alías y Joan Claudi Socoró

Departamento de Comunicaciones y Teoría de la Señal

Enginyeria i Arquitectura La Salle. Universidad Ramon Llull

Pg. Bonanova, 8 - 08022 Barcelona

{xavis,gcobo,falias,jclaudi}@salle.url.edu

Resumen: Las prestaciones de los sistemas de clasificación no supervisada de documentos están supeditadas al uso de representaciones textuales óptimas, las cuales no son sólo difíciles de determinar de antemano, sino que pueden variar de un problema de clasificación a otro. Este trabajo propone una metodología basada en diversidad de representaciones y conjuntos de clasificadores no supervisados como primer paso hacia la construcción de sistemas robustos de clasificación no supervisada. Los experimentos realizados sobre tres problemas de categorización binaria de dificultad creciente muestran que el método propuesto es *i)* robusto frente a selecciones no óptimas de la dimensionalidad de las representaciones, y *ii)* capaz de detectar interacciones constructivas entre distintas representaciones textuales, llegando a obtener índices de categorización por consenso superiores a los conseguidos por los clasificadores individuales disponibles.

Palabras clave: Representación de documentos, clasificación no supervisada, conjuntos de clasificadores.

Abstract: The performance of document clustering systems is conditioned by the use of optimal text representations, which are not only difficult to determine beforehand, but also may vary from one clustering problem to another. This work presents an approach based on feature diversity and cluster ensembles as a first step towards building document clustering systems that behave robustly across different clustering problems. Experiments conducted on three binary clustering problems of increasing difficulty show that the proposed method is *i)* robust to near-optimal model order selection, and *ii)* able to detect constructive interactions between different document representations, thus being capable of yielding consensus clusterings superior to any of the individual clusterings available.

Keywords: Document representation, clustering, cluster ensembles.

1 Introduction

In recent years, content-based automatic management of text documents has gained much attention from various research communities, as the need for efficient tools able to filter, classify, index and retrieve documents according to their thematic contents has grown as quickly as the number and size of available digital text document databases.

The text analysis literature covers a wide range of tasks which are instances of text mining, such as document clustering, retrieval and classification. Most techniques posed to solve these problems belong to the machine learning paradigm (Sebastiani, 2002), and their performance heavily relies on finding document representations which reflect the contents of documents to a maximum extent. This issue becomes specially relevant in the context of unsupervised text analysis applications such as document clustering. Moreover, due to the generally low availability of labeled document collections, efficient document clustering techniques are often regarded as a necessary tool to organize unlabeled corpora for subsequent browsing or retrieval (Xu, Liu, and Gong, 2003).

This work deals with robust flat document clustering (i.e. document partitioning), the task of grouping a set of $|\mathcal{D}|$ unlabeled documents in a predefined number of clusters K according to their thematic contents. That is, the only background knowledge available is the number of clusters we want to group the documents in (K), which usually coincides with the expected number of thematic categories contained in the corpus.
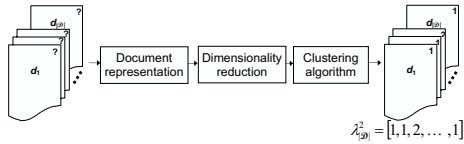
As depicted in figure 1, the required steps

Figure 1: Block diagram of a document clustering system for the particular case of binary clustering ($K = 2$).

for building a document clustering system can be enumerated as follows: *i)* representation of the corpus subject to clustering for computer-based analysis, *ii)* derivation of a dimensionally reduced representation through feature extraction plus model order selection, and *iii)* application of a clustering algorithm. As a result of this process, each of the $|\mathcal{D}|$ documents contained in the corpus is assigned one of K possible labels. In this work, we will refer to the labeling resulting from clustering $|\mathcal{D}|$ documents in K clusters as $\lambda_{|\mathcal{D}|}^{K} \in \mathbb{Z}_{(1...K)}^{|\mathcal{D}|}$, i.e. as a $|\mathcal{D}|$-dimensional vector containing integer labels from 1 to K.

A fully automatic document clustering system should be able to choose the document representation, the dimensionality of such representation and the clustering technique that maximize some objective classification performance measure. However, although document representation techniques have been compared (Cobo et al., 2006), model order selection approaches have been applied (Kolenda, Hansen, and Sigurdsson, 2000), and clustering methods have been extensively studied (Jain, Murty, and Flynn, 2002), it is still difficult to determine *a priori* the most suitable type of representation and its optimal dimensionality given a particular document clustering problem.

In this context, we propose building a document clustering system able to generate a robust clustering from a bunch of candidate document representations in an unsupervised manner through the use of cluster ensembles. The clustering output by the cluster ensemble should be able to attain at least the same performance as the best individual clustering obtained from the candidate representations. Moreover, we analyze if the cluster ensemble could benefit from constructive interrelations between candidate clusterings in order to improve the performance of the best individual clustering.

2   Cluster ensembles

The use of classifier committees is a well-known approach for boosting performance in the context of supervised text classification (Sebastiani, 2002). However, combining the results of several clustering processes is a fairly more complex task. This is probably the reason why this topic has received little attention until recently (Strehl, 2002; Fred and Jain, 2002; Topchy, Jain, and Punch, 2003). In particular, Strehl's work is one of the first research efforts in this area reporting experiments on document clustering.

The cluster ensembles approach was originally defined for integrating several clusterings by supplying the labelings output by each individual clusterer to a consensus function which yields a global clustering (Strehl, 2002). One of the most appealing capacities of cluster ensembles is their potential to improve the best individual clustering available, provided that sufficient diversity is found among the individual clusterings (Strehl, 2002). In his work, Strehl presents an application of cluster ensembles which tries to enhance clustering performance by using an ensemble of *different* clustering algorithms operating on the *same* data, i.e. a committee of unsupervised classifiers.

The most distinctive feature between our approach and Strehl's is the fact that we do not only want to obtain a consensus clustering that improves individual clusterings, but we also want to construct clustering systems which are robust to variations of the optimal document representation across different clustering problems. Therefore, diversity is provided in our case by the range of features employed to represent documents. In this work, the cluster ensembles consist of C *identical* individual clusterers (in our case, standard K-means -KM- using cosine distance[1]) fed in parallel with *distinct* document representations (Sevillano et al., 2006).

Given the dependence of the cluster ensemble performance on how consensus among individual clusterings is built, we have implemented in this work those consensus functions deemed as top performing in (Strehl, 2002): Cluster-based Similarity Partitioning Algorithm (CSPA) and Meta-Clustering Al-

---

[1]We have chosen KM clustering as it is one of the most popular clustering algorithms. Moreover, cosine distance outperformed Euclidean distance in our experiments (Cobo et al., 2006).

gorithm (MCLA), which are briefly described throughout the following paragraphs.

Both CSPA and MCLA consensus functions require transforming the set of candidate clusterings into a hypergraph, which is built from the binary membership indicator matrices corresponding to each clustering. In fact, the concatenation of those membership matrices constitutes the adjacency matrix of a hypergraph with $|\mathcal{D}|$ vertices and $C \cdot K$ hyperedges (Strehl, 2002).

In CSPA, a simple multiplication of the adjacency matrix of the hypergraph by its transpose yields a similarity-between-documents matrix which, in turn, is used to recluster the documents using a similarity-based clustering algorithm such as METIS (Karypis and Kumar, 1998), giving rise to the consensus clustering.

In contrast, MCLA solves a cluster correspondence problem, as it identifies and consolidates those groups of clusters (or meta-clusters) that share a larger amount of documents (see (Strehl, 2002) for further details).

3 Document representations

In this work, documents are represented using the Vector Space Model (VSM) (Salton, 1989). The initial document representation is term-based (i.e. each dimension of the vector space corresponds to a word appearing in the corpus). Hence, the document collection is represented as a term-by-document matrix $\mathbf{X} \in \mathbb{R}^{|\mathcal{T}| \times |\mathcal{D}|}$, where $|\mathcal{T}|$ is the size of the vocabulary and $|\mathcal{D}|$ is the total number of documents. So as to create feature diversity, three other candidate representations are derived from the term-based representation by means of the following feature extraction techniques[2]:

- *Latent Semantic Indexing* (LSI) (Deerwester et al., 1990) allows to perform dimensionality reduction by retaining the singular vectors associated with the largest singular values resulting from the Singular Value Decomposition (SVD) of the term-by-document matrix $\mathbf{X}$:

$$\mathbf{X} = \mathbf{U} \cdot \mathbf{\Sigma} \cdot \mathbf{V}^T \qquad (1)$$

where matrix $\mathbf{\Sigma}$ contains the singular values ordered in decreasing order and matrices $\mathbf{U}$ and $\mathbf{V}^T$ contain the left and right singular vectors, respectively. Dimensionality reduction is conducted by retaining the first M rows of matrix $\mathbf{V}^T$, that contain the location of the $|\mathcal{D}|$ documents in a M-dimensional orthogonal space, in which clustering is conducted.

- *Independent Component Analysis* (ICA) (Kolenda, Hansen, and Sigurdsson, 2000) is based on the assumption that the document collection (matrix $\mathbf{X}$) is generated by an unknown linear combination of M statistically independent hidden topics.

The use of ICA in text analysis is usually preceded by LSI, as this procedure is equivalent to the usual whitening step that simplifies ICA algorithms (Hyvarinen, Karhunen, and Oja, 2001). Applying ICA on the LSI data yields an estimation $\tilde{\mathbf{S}}$ of the M independent topics which generated the documents:

$$\tilde{\mathbf{S}} = \mathbf{W} \cdot \mathbf{X} \qquad (2)$$

where $\mathbf{W}$ is known as the separating matrix. Subsequently, the clustering algorithm is fed with the $M \times |\mathcal{D}|$ matrix $\tilde{\mathbf{S}}$.

- *Non-Negative Matrix Factorization* (NMF) (Lee and Seung, 1999) is a technique that factorizes the non-negative term-by-document matrix $\mathbf{X}$ into the product of two non-negative matrices $\mathbf{W}$ and $\mathbf{H}$:

$$\mathbf{X} = \mathbf{W} \cdot \mathbf{H} \qquad (3)$$

NMF assumes that the document collection ($\mathbf{X}$) is generated by the sum of a set of M hidden non-negative variables (i.e. topics), contained in the $M \times |\mathcal{D}|$ matrix $\mathbf{H}$. Hence, the clustering process is conducted on the corresponding latent topic $M \times |\mathcal{D}|$ matrix $\mathbf{H}$.

A key issue concerning the VSM is automatically choosing its optimal dimensionality M (model order selection). However, in this work we conduct supervised dimensionality selection in order to focus on the performance of the cluster ensemble solely.

---

[2]Other representations such as term selection plus change of basis (Srinivasan, 2002) were tested but finally discarded as they resulted in poorer clustering performance.
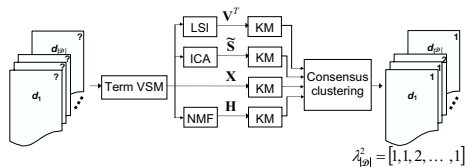
Figure 2: Block diagram of the proposed document clustering system based on feature diversity and cluster ensembles for the particular case of binary clustering ($K = 2$).

4 Experiments

Experiments have been conducted on the miniNewsgroups[3] corpus, a subset of the 20 Newsgroups document collection that contains 100 documents from each newsgroup. In this work, we have focused our attention in the three binary clustering problems described in (Srinivasan, 2002): *i)* the `NG1&NG2` problem: two well-separated categories (alt.atheism -`NG1`- and comp.graphics -`NG2`-), *ii)* the `NG10&NG11` problem: two categories with some overlap (rec.sport.baseball -`NG10`- and rec.sport.hockey -`NG11`-), and *iii)* the `NG18&NG19` problem: two highly overlapped categories (talk.politics.mideast -`NG18`- and talk.politics.misc -`NG19`-). Hence, in our experiments, $|\mathcal{D}| = 200$ and $K = 2$.

Firstly, the documents are represented in the term-based VSM using the normalized *tfidf* weighting scheme (Sebastiani, 2002) and subsequently transformed into the LSI, ICA and NMF representations. Then, four KM clusterers are fed in parallel with these document representations, and a consensus clustering is built upon the labelings generated by these clusterers (see figure 2). Each clustering process is conducted 10 times in order to reduce the influence of the random initialization of the KM clusterers and attain statistically reliable results (Cobo et al., 2006). Hence, all the results presented throughout the following sections correspond to the average of 10 trials.

Throughout the following sections several experiments are presented. Firstly, we seek the optimal dimensionality of each document representation. Secondly, the performance and robustness of cluster ensembles and consensus clusterings are analyzed. And finally,

---

[3] The miniNewsgroups corpus is available online at `http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html`.

the effect of evaluating the best individual and the best consensus clusterings in terms of different classification efficiency measures is studied.

4.1 Finding the optimal document representations

The first experiment consists in conducting supervised model order selection for each representation technique in each clustering problem. To that effect, the clustering $\lambda_{|\mathcal{D}|}^{K}$ obtained from each document representation and the documents' original labeling $\kappa_{|\mathcal{D}|}^{K}$ are compared in terms of their normalized mutual information (NMI):

$$\text{NMI}(\lambda_{|\mathcal{D}|}^{K}, \kappa_{|\mathcal{D}|}^{K}) = \frac{2}{|\mathcal{D}|} \sum_{i=1}^{K} \sum_{j=1}^{K} n_i^j \log_{K \cdot K} \left(\frac{n_i^j \, |\mathcal{D}|}{n^j \, n_i} \right) \tag{4}$$

where n_i^j denotes the number of documents assigned to cluster i by $\lambda_{|\mathcal{D}|}^{K}$ and in class j according to $\kappa_{|\mathcal{D}|}^{K}$, n^j is the number of documents belonging to class j as designed by $\kappa_{|\mathcal{D}|}^{K}$ and n_i the number of documents in cluster i according to $\lambda_{|\mathcal{D}|}^{K}$ (Strehl, 2002).

Regarding feature extraction based document representations (i.e. LSI, ICA and NMF), the optimal dimensionality is found by performing a sweep from 2 to 100 dimensions. However, only results up to 20 dimensions are shown in figure 3, as the maxima of NMI are always found within this range. In the case of the term-based representation, we seek the optimal dimensionality by simple term selection based on ranking each term according to its *tfidf* weight. Details about the optimal dimensionality of each document representation technique are described throughout the following paragraphs.

4.1.1 `NG1&NG2` problem

As shown in figure 3a, the best individual clustering results are achieved when the KM clusterer operates on a 2-dimensional LSI, a 4-d(imensional) ICA and a 9-d NMF space. In the case of the term-based document representation, we observed that NMI experienced a monotonic increase, yielding the best performance when all terms were considered (10184 in total). Therefore, the maximum NMI corresponding to the term-based clustering is depicted in figure 3a as a constant baseline. To sum up, inspection of this dia-

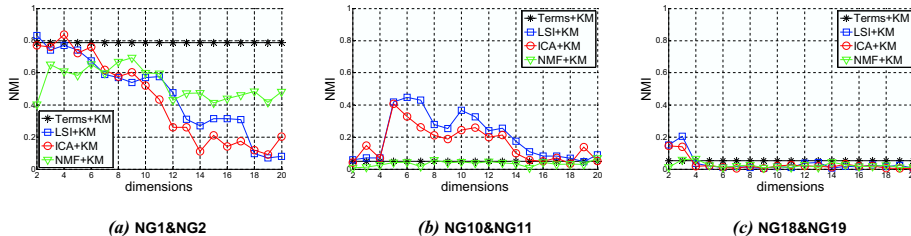*(a) NG1&NG2*   *(b) NG10&NG11*   *(c) NG18&NG19*

Figure 3: NMI between the original labeling and the clusterings obtained by KM fed with terms, LSI, ICA and NMF document representations as a function of the dimensionality for the three clustering problems.

gram reveals that the top performing document representation in this clustering problem is 4-d ICA.

4.1.2   `NG10&NG11` **problem**

In this case, the best individual feature extraction based clustering results are achieved on 6-d LSI, 5-d ICA and 8-d NMF representations (see figure 3b). Term-based clustering also achieves its maximum NMI when all terms were considered (7094 in this case), so its corresponding NMI is also depicted in figure 3b as a constant baseline. In conclusion, the optimal document representation in this clustering problem is 6-d LSI.

4.1.3   `NG18&NG19` **problem**

Results regarding this third clustering problem are presented in figure 3c. The best clustering results are achieved on 3-d LSI, 2-d ICA, 4-d NMF and (all) 9181 terms representations. Just like in the previous experiment, the LSI document representation achieves the best clustering results in terms of NMI.

In summary, the results of these experiments show how optimal document representations vary across different clustering problems, not only in terms of their dimensionality, but also in terms of the feature employed. Therefore, the need for designing clustering systems robust to these variations seems to be reasonable.

4.2   Cluster ensembles for robust clustering

The second experiment consists in building consensus clusterings from the four parallel KM clusterers by means of the CSPA and MCLA consensus functions. In this experiment we perform a twofold analysis. Firstly, we analyze the behaviour of such consensus functions by comparing their performance

when the cluster ensemble is fed with clusterings obtained using optimal *vs.* suboptimal dimensionalities for document representation. Hence, as a first step to simulate suboptimal model order selection, we created feature extraction based near-optimal representations by choosing those attaining the second highest NMI in the dimensionality sweep presented in section 4.1. As regards term based representations, suboptimal representations were created by applying term selection, in this case, discarding the 20% of the terms with lowest *tfidf* weight. The dimensionalities of the optimal and suboptimal document representations employed hereafter are presented in table 1.

	Feature	Optimal	Suboptimal
NG1& NG2	Terms	10184	8147
	LSI	2	4
	ICA	4	2
	NMF	9	8
NG10& NG11	Terms	7094	5675
	LSI	6	7
	ICA	5	6
	NMF	8	6
NG18& NG19	Terms	9181	7344
	LSI	3	2
	ICA	2	3
	NMF	4	3

Table 1: Dimensionality of optimal and the selected suboptimal document representations.

Secondly, as equal or better clustering results are obtainable using very low-dimensional extracted features in comparison to the high-dimensional term representation (see figure 3), we analyze the relevance of the term-based document representation in this
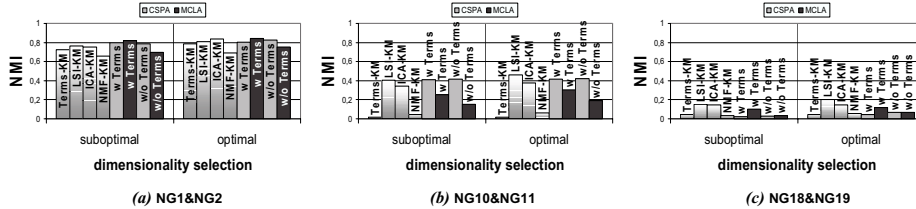
Figure 4: NMI of the four individual clusterings and the consensus functions with or without terms using suboptimal (left bar plot) and optimal (right bar plot) dimensionality selection.

context. Hence, we create consensus labelings both considering (w Terms) and ignoring (w/o Terms) the term-based clustering. The quality of the individual and consensus clusterings is evaluated in terms of the NMI with respect to the documents' original labeling. The results of this experiment are presented and discussed in the following subsections.

4.2.1 NG1&NG2 problem

As shown in figure 4a, both CSPA and MCLA consensus functions perform quite similarly when terms are considered. However, MCLA is negatively affected by the absence of terms, in contrast to CSPA. With regard to the optimal dimensionality experiment (right bar plot in figure 4a), the winning consensus function (MCLA w Terms) is, in terms of NMI, slightly better than the best individual clustering (4-d ICA-KM). Moreover, when suboptimal clusterings are fed into MCLA, the best resulting consensus clustering (left bar plot in figure 4a) is better than the best individual suboptimal clustering (7% average relative improvement). And more important, it is almost equivalent to the MCLA w Terms consensus clustering obtained in the optimal dimensionality selection case, which suggests that this consensus function is able to cope with near-optimal model order selection in this clustering problem.

4.2.2 NG10&NG11 problem

The results of this experiment are depicted in figure 4b, which shows that CSPA outperforms MCLA in this case. Again, we observe that MCLA is clearly spoiled when terms are ignored, whereas CSPA even experiences a slight performance improvement. When consensus is built upon optimal dimensionality document representations (right bar plot in figure 4b), the best consensus function (CSPA w/o Terms) achieves lower NMI than the optimal individual clustering (6-d

LSI-KM) (8% average relative decrease). In contrast, when suboptimal clusterings are fed into CSPA, the resulting best consensus clustering (left bar plot in figure 4b) achieves better results than the best individual suboptimal clustering (3% average relative improvement). And again, it is nearly equivalent to the CSPA consensus clustering obtained in the optimal dimensionality selection case, which suggests that consensus functions are robust in front of near-optimal model order selection.

4.2.3 NG18&NG19 problem

Figure 4c shows the quality of individual and consensus clusterings in terms of NMI in the most difficult clustering problem. We observe that none of the consensus functions is able to improve the best individual clustering, neither in the optimal nor in the suboptimal dimensionality selection case (in fact, dramatic average relative decreases around 35% in NMI are observed). With respect to the presence or absence of terms, MCLA and CSPA show a behaviour similar to that reported in the previous experiments.
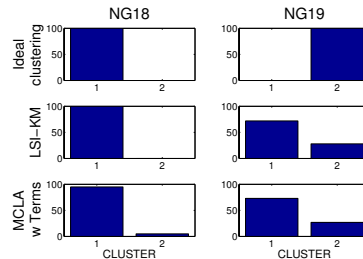

Figure 5: Average document distribution histograms of classes NG18 and NG19 under an ideal clustering, the best individual clustering (LSI-KM) and the best consensus clustering (MCLA w Terms) in terms of NMI.
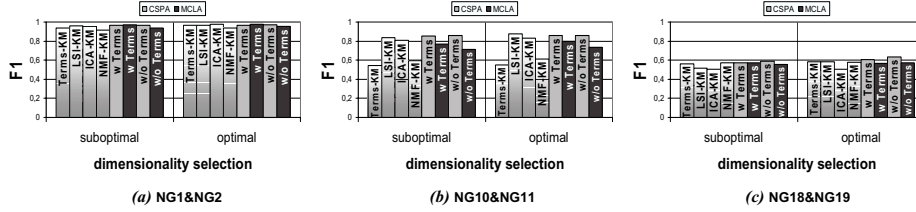
Figure 6: F_1 of the four individual clusterings and the consensus functions with or without terms using suboptimal (left bar plot) and optimal (right bar plot) dimensionality selection.

4.3 Clustering evaluation by F_1

Driven by the poor results obtained in this last experiment, we analyzed the distributions of documents resulting from the individual and consensus clusterings that achieved the highest NMI. In particular, we studied how many documents from each class (NG18 and NG19) were assigned in average to each cluster (see figure 5).

The upper row of figure 5 depicts the ideal clustering of the documents belonging to NG18 and NG19, and it is provided as a reference. However, as shown in its second row, the best individual clustering in terms of NMI (LSI-KM) fails to separate both classes. In particular, all the documents of class NG18 and 72% of the documents belonging to NG19 are assigned to the same cluster. By inspecting the third row of figure 5, we observe that the best consensus clustering in terms of NMI (MCLA w Terms) yields a similar document distribution. Such a clustering results in a very high recall and a very low precision with respect NG18 (and obviously, the opposite behaviour with respect NG19). According to these results, it seems that evaluating clusterings in terms of NMI can give rise to poor class separations, specially in difficult clustering scenarios.

So as to obtain better consensus clusterings, we evaluated employing a performance measure which, in contrast to NMI, takes into account the soundness and completeness of the resulting clusters, i.e. the well-known F_1 measure (the harmonic mean of precision and recall) (Sebastiani, 2002). As in the case of NMI, the best clusterings will be those attaining the highest average F_1. The results of these experiments are shown in figure 6.

We observe that in the NG1&NG2 and NG10&NG11 clustering problems there exists a correspondence between NMI and F_1 evaluations, as those clusterings deemed as the best are the same in both cases. Moreover, the performance between the best individual and the best consensus clustering is comparable irrespective of whether NMI or F_1 is employed (compare figure 4a to 6a and figure 4b to 6b).

However, in the third clustering problem (NG18&NG19), the best consensus clustering in terms of F_1 (CSPA w/o Terms) differs from the best consensus clustering in terms of NMI (MCLA w Terms). The same applies to the best individual clustering in the suboptimal dimensionality selection case (LSI-KM yields the highest NMI whereas NMF-KM attains maximum F_1 measure). And more important, the best consensus clusterings attain a higher F_1 than the best individual clusterings (average relative improvements of 8% in the optimal dimensionality case and 3% in the suboptimal case), which is in sharp contrast with the situation depicted in figure 4c, where consensus clusterings were much poorer than the best individual clustering available.

Furthermore, figure 7 shows the documents distribution corresponding to the consensus clustering that attained maximum F_1.
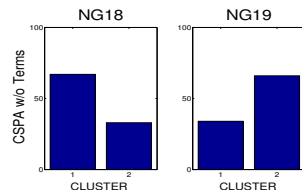

Figure 7: Average document distribution histograms of classes NG18 and NG19 under the best consensus clustering in terms of F_1 (CSPA w/o Terms).

In contrast to the situation depicted in figure 5, the best consensus clustering in terms of F_1 (CSPA w/o Terms) tends to separate both classes, although quite poorly. However,

this trend is captured by the F_1 measure, but not by NMI. Therefore, we conclude that it seems more appropriate to use F_1 than NMI for evaluating consensus clusterings, specially in the context of difficult clustering problems.

5 Conclusions

In this work, a study on the application of cluster ensembles to exploit feature diversity for building robust document clustering systems has been presented. This strategy has proven to be *i)* robust to small errors in model order selection across different clustering problems, and *ii)* able to generate consensus clusterings that, in terms of the F_1 measure, mostly improve any of the individual clusterings in the ensemble.

This paper constitutes a first step towards building robust document clustering systems, and therefore, several challenges still lie ahead: firstly, it is necessary to implement a supra-consensus function which allows to choose the most appropriate consensus technique (MCLA or CSPA) for each scenario in an unsupervised manner. Secondly, the robustness of the proposal must be checked under more severe experimental conditions, e.g. addressing clustering problems of more than two categories and employing data representations far-off the optimal ones. Thirdly, unsupervised performance evaluation measures such as Averaged NMI (Strehl, 2002) should be employed, as the original labeling of the documents will not be available in real clustering applications. And finally, widening the range of document representations, clustering algorithms and cluster ensemble methodologies is necessary for finding the limitations of our proposal.

References

Cobo, G., X. Sevillano, F. Alías, and J.C. Socoró. 2006. Técnicas de Representación de Textos para Clasificación no Supervisada de Documentos. *Procesamiento de Lenguaje Natural*, 37.

Deerwester, S., S.-T. Dumais, G.-W. Furnas, T.-K. Landauer, and R. Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal American Society Information Science*, 6(41):391–407.

Fred, A.L.N. and A.K. Jain. 2002. Data Clustering Using Evidence Accumulation. In *Proceedings of the 16th International Conference on Pattern Recognition*, pages 276–280, Quebec City, Canada.

Hyvarinen, A., J. Karhunen, and E. Oja. 2001. *Independent Component Analysis*. John Wiley and Sons.

Jain, A., M. Murty, and P. Flynn. 2002. Data Clustering: a Survey. *ACM Computing Surveys*, 31(3):264–323.

Karypis, G. and V. Kumar. 1998. A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs. *SIAM Journal on Scientific Computing*, 20(1):359–392.

Kolenda, T., L.K. Hansen, and S. Sigurdsson. 2000. Independent components in text. In M.Girolami, editor, *Advances in Independent Component Analysis*. Springer-Verlag, pages 241–262.

Lee, D.D. and H.S. Seung. 1999. Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature*, 401:788–791.

Salton, G. 1989. *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley.

Sebastiani, F. 2002. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1):1–47.

Sevillano, X., G. Cobo, F. Alías, and J.C. Socoró. 2006. Feature Diversity in Cluster Ensembles for Robust Document Clustering. In *Proceedings of the 29th ACM SIGIR Conference*, Seattle, WA, USA.

Srinivasan, S.H. 2002. Features for Unsupervised Document Classification. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 36–42, Taipei, Taiwan.

Strehl, A. 2002. Relationship-based Clustering and Cluster Ensembles for High-Dimensional Data Mining. The University of Texas at Austin, May. PhD thesis.

Topchy, A., A.K. Jain, and W. Punch. 2003. Combining Multiple Weak Clusterings. In *Proceedings of the 3rd IEEE International Conference on Data Mining*, pages 331–338, Melbourne, FLA, USA.

Xu, W., X. Liu, and Y. Gong. 2003. Document Clustering Based on Non-Negative Matrix Factorization. In *Proceedings of the 26th ACM SIGIR Conference*, pages 267–273, Toronto, Canada.