Feature Diversity in Cluster Ensembles for Robust Document Clustering

Xavier Sevillano, Germán Cobo, Francesc Alías and Joan Claudi Socoró
Department of Communications and Signal Theory
Enginyeria i Arquitectura La Salle. Ramon Llull University
Pg. Bonanova 8, 08022 - Barcelona, Spain
{xavis,gcobo,falias,jclaudi}@salle.url.edu

ABSTRACT

The performance of document clustering systems depends on employing optimal text representations, which are not only difficult to determine beforehand, but also may vary from one clustering problem to another. As a first step towards building robust document clusterers, a strategy based on feature diversity and cluster ensembles is presented in this work. Experiments conducted on a binary clustering problem show that our method is robust to near-optimal model order selection and able to detect constructive interactions between different document representations in the test bed.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text Analysis*; I.5.3 [**Pattern Recognition**]: Clustering—*Algorithms*

General Terms

Algorithms, Design, Experimentation, Performance

Keywords

Document clustering, feature extraction, cluster ensembles

1. INTRODUCTION

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 performance measure. Though clustering techniques have been extensively studied and compared (see [3] for a survey), it is still difficult to determine *a priori* the optimal type of representation and its dimensionality given a particular document clustering problem.

In this context, it would be interesting to develop a system able to generate a global clustering from a bunch of candidate document representations in an unsupervised manner, attaining at least the same performance as the best individual clustering (BIC) obtained from those representations. Moreover, it would be interesting to analyze if this approach could benefit from constructive interrelations between candidate clusterings in order to improve the BIC performance. In this work, we make an initial approach to this issue by creating cluster ensembles fed with different representations of the document collection subject to clustering.

SIGIR'06, August 6–11, 2006, Seattle, Washington, USA.
ACM 1-59593-369-7/06/0008.

2. CLUSTER ENSEMBLES

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 [7]. One of the most appealing capacities of cluster ensembles is their potential to improve the BIC available, provided that sufficient diversity is found among the individual clusterings [7].

In our case, the cluster ensembles consist of several identical individual clusterers (in our case, standard k-means -KM- using cosine distance) fed in parallel with distinct document representations. Hence, diversity is provided by the range of features employed to represent documents.

The performance of the cluster ensemble is totally dependent on the consensus function employed. In this work, we have implemented those consensus functions deemed as top performing in [7]: Cluster-based Similarity Partitioning Algorithm (CSPA) and Meta-Clustering Algorithm (MCLA) (see [7] for further details).

3. DOCUMENT REPRESENTATIONS

In this work, documents are represented in the Vector Space Model (VSM) [5]. The initial document representation is term-based (i.e. each dimension of the vector space corresponds to a word appearing in the corpus). So as to create feature diversity, three other candidate representations are derived from the term-based representation by means of feature extraction techniques, namely Latent Semantic Indexing (LSI) [1], Independent Component Analysis (ICA) [2] and Non-negative Matrix Factorization (NMF) [4]. Other representations such as term selection plus change of basis [6] were discarded due to their inferior performance.

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

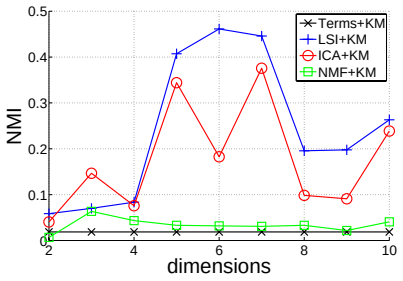

Figure 1: Averaged NMI between the original labeling and the clusterings obtained by k-means fed by terms, LSI, ICA and NMF document representations as a function of the dimensionality.

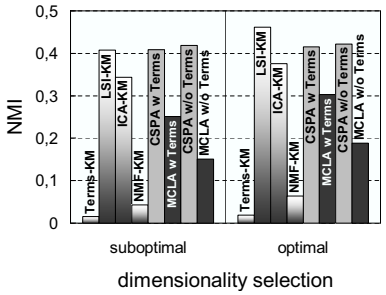

Figure 2: NMI of the four individual clusterings and the consensus functions with or without terms using suboptimal (left) and optimal (right) dimensionality selection.

4. EXPERIMENTS AND DISCUSSION

Experiments have been conducted on the miniNewsgroups[1] corpus, a subset of the 20 Newsgroups document collection that contains 100 documents from each newsgroup. As a first step towards robust clustering, we have focused our attention in a medium difficulty binary clustering problem where the categories present some overlap (`rec.sport.hockey` and `rec.sport.baseball`)[6].

Firstly, the documents are represented in the VSM using the normalized *tfidf* weighting scheme [5] and subsequently transformed to 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. 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.

The first experiment consists in conducting supervised model order selection for each representation technique by computing the normalized mutual information (NMI) between each clustering and the documents' original labeling. Regarding feature extraction-based document representations (i.e. LSI, ICA and NMF), the optimal dimensionality in terms of NMI is found by performing a sweep from 2 to 100 dimensions. However, only results up to 10 dimensions are shown in figure 1, as the maxima of NMI are always found within this range. We conclude that the best individual clustering results are achieved when the KM clusterer operates on a 6-dimensional LSI, a 7-d(imensional) ICA and a 3-d NMF space. In the case of term-based representations, we seek the optimal dimensionality by simple term selection based on ranking each term according to its *tfidf* weight. We observed that NMI experienced a monotonic increase, yielding the best performance when all terms were considered (7094 in total). Therefore, its corresponding NMI is depicted in figure 1 as a constant baseline.

The second experiment consists in building a consensus clustering 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 performance and robustness of such consensus functions by feeding them with clusterings obtained from document representations of optimal and suboptimal dimensionality. To that effect, we create KM clusterings from 5-d LSI, 5-d ICA, 4-d NMF and 5675-term (i.e. 80% of the terms with highest *tfidf*) document representations in order to simulate near-optimal model order selection. Secondly, as clustering based on terms (constant baseline performance in figure 1) is inferior to the individual clusterings based on extracted features, we analyze the relevance of the term-based representation in this context. Hence, we create consensus labelings both considering and ignoring the term-based clustering.

It can be observed from figure 2 that CSPA clearly outperforms MCLA. Moreover, we observe that MCLA is dramatically spoiled when terms are ignored, whereas CSPA even experiences a slight performance improvement. With regard to the optimal dimensionality experiment (right bar plot in figure 2), we can see that the winning consensus function (CSPA w/o terms) achieves lower NMI (8% average relative decrease) than the BIC (6-d LSI-KM). In contrast, when suboptimal clusterings are fed into CSPA, the resulting consensus clustering (left bar plot in figure 2) achieves slightly better results than the BIC (3% average relative increase). More important, it is equivalent to the CSPA consensus clustering obtained in the optimal dimensionality selection case, which indicates that this consensus function is able to cope with near-optimal model order selection.

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

5. REFERENCES

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

[2] C.-L. Isbell and P. Viola. Restructuring Sparse High Dimensional Data for Effective Retrieval. *Adv. in Neural Information Proc. Systems*, 11:480–486, 1999.

[3] A. Jain, M. Murty, and P. Flynn. Data Clustering: a Survey. *ACM Computing Surveys*, 31(3):264–323, 1999.

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

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

[6] S.-H. Srinivasan. Features for Unsupervised Document Classification. In *Proc. of CoNLL-2002*, pages 36–42.

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