A Hierarchical Consensus Architecture 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. Barcelona (Spain)
{xavis,gcobo,falias,jclaudi}@salle.url.edu

Abstract. A major problem encountered by text clustering practitioners is the difficulty of determining *a priori* which is the optimal text representation and clustering technique for a given clustering problem. As a step towards building robust document partitioning systems, we present a strategy based on a hierarchical consensus clustering architecture that operates on a wide diversity of document representations and partitions. The conducted experiments show that the proposed method is capable of yielding a consensus clustering that is comparable to the best individual clustering available even in the presence of a large number of poor individual labelings, thus largely avoiding the problems derived from selecting non-optimal clustering configurations.

1 Introduction

The low availability of labeled document collections has made document clustering techniques become a necessary tool to organize unlabeled corpora according to their thematic contents. However, finding a thematically meaningful partition of an unlabeled document collection is not a straightforward task. This is mainly due to the difficulty of blindly choosing the optimal document representation and clustering method that, for a given a clustering problem, ensure the best match between the true labeling of the documents and the partitioning results.

That is, the performance of document clustering systems requires finding representations of documents that reflect their contents to a maximum extent. Unfortunately, it is difficult to determine *a priori* the optimal type of representation and its dimensionality given a particular document clustering problem. We call this situation the *data representation dependence* effect. Moreover, the application of different clustering methods on the same data often yields different partitions. This gives rise to what we call the *algorithm dependence* effect.

Due to both effects, obtaining the optimal partition of an unlabeled corpus on a single run of a clustering algorithm fed by a specific text representation is a rather challenging aim. Allowing for these circumstances, in this work we define a generic framework for deriving a robust partition of a document collection by building a cluster ensemble upon a wide range of text representations and partitions, following a hierarchical consensus strategy. This proposal is a step towards setting the text clustering practitioner free from the obligation of blindly choosing a single document representation and clustering technique.

2 Hierarchical Consensus Clustering Architecture

The hierarchical consensus clustering architecture is a modular and flexible proposal that allows to obtain a robust partition of the document collection subject to clustering. In this paper, it is assumed that the only knowledge available is the number of clusters we want to group the documents in (i.e. the expected number of thematic categories). Figure 1 depicts the specific implementation of the hierarchical consensus clustering architecture employed in this work.

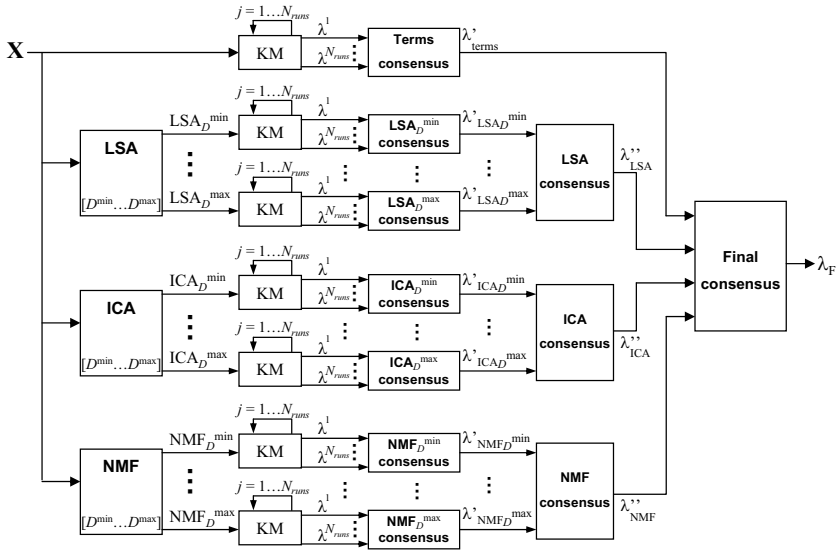

Fig. 1. Hierarchical consensus clustering architecture using k-means (KM) clustering and terms, LSA, ICA and NMF document representations.

The document corpus is initially represented in a term-based vector space (term-by-document matrix $\mathbf{X}$), using the *tfidf* weighting scheme [4]. However, it is a commonplace that the efficiency of clustering systems can be improved by the use of document representations based on feature extraction techniques [5]. Therefore, alternative document representations are derived by applying Latent Semantic Analysis (LSA) [1], Independent Component Analysis (ICA) [2] and Non-negative Matrix Factorization (NMF) [3], with dimensionalities D ranging from $D^{\min}$ to $D^{\max}$. The interval $[D^{\min}, D^{\max}]$ should be wide enough so that the optimal dimensionality of each representation is included in it, but need not be equal for all the representations.

Subsequently, diverse partitions are created by running the k-means (KM) algorithm N_{runs} times with random centroids initialization on each distinct representation[1], yielding what we call *individual clusterings*, denoted as $\lambda^1, \ldots, \lambda^{N_{runs}}$

[1] A more generic strategy for introducing algorithm diversity at this stage would consist in applying distinct clustering algorithms on the same data.

in figure 1. These clusterings form a cluster ensemble which is fed into a first consensus stage, which allows us to overcome the algorithm dependence effect. As a result, a consensus labeling for each distinct document representation is obtained (e.g. for LSA, $\{\lambda'_{\mathrm{LSA}_{D^{\min}}}, \ldots, \lambda'_{\mathrm{LSA}_{D^{\max}}}\}$, or $\lambda'_{\mathrm{terms}}$ for terms).

Next, the labelings corresponding to each feature extraction based document representation method are fed into a second consensus stage, which yields a single labeling per representation scheme (i.e. λ''_{LSA}, λ''_{ICA} and λ''_{NMF} in figure 1). Finally, these labelings (together with $\lambda'_{\mathrm{terms}}$) are used for building the final consensus clustering λ_{F}. Hopefully, λ_{F} will be very similar to the best individual clustering available, thus overcoming the data representation dependence effect.

3 Experiments

Experiments have been conducted on the miniNewsgroups corpus, a subset of the 20 Newsgroups document collection that contains 100 documents from each newsgroup. More specifically, six categories of the miniNewsgroups corpus have been used in the following experiments (`comp.graphics`, `rec.autos`, `sci.crypt`, `misc.forsale`, `talk.politics.misc` and `talk.religion.misc`).

In order to build the consensus clusterings, we have compared the following consensus functions: Cluster-Similarity Partitioning Algorithm (CSPA), Hyper-Graph Partitioning Algorithm (HGPA) and Meta-Clustering Algorithm (MCLA) [6]. The tunable parameters of the hierarchical architecture depicted in figure 1 have been set to $D^{\min} = 2$, $D^{\max} = 50$ and $N_{runs} = 30$, hence giving rise to a total of 4440 individual clusterings. Clustering results are evaluated in terms of the F1 measure [4] with respect the true category membership of each document.

Figure 2a presents the F1 measure of the individual and consensus clusterings at the first stage of the hierarchy on a particular case (LSA with $D = 20$). It can be observed that fairly diverse F1 measure values are achieved across the 30 KM algorithm runs, which somehow illustrates the algorithm dependence effect. However, the MCLA and CSPA consensus functions are capable of keeping close to the best individual clusterings, while HGPA does not.

In figure 2b, the results of building a LSA consensus clustering on the labelings output by the previous consensus stage are presented, illustrating the influence of the document representation dimensionality. Again, the MCLA and CSPA consensus functions yield λ''_{LSA} clusterings which are very close or even slightly better than the best clustering input to the consensus function.

Finally, the consensus λ_{F} is evaluated by means of the F1 measure histogram depicted in figure 2c, thus comparing the final labeling output by each consensus function to each one of the 4440 individual clusterings. Once more, the MCLA and CSPA are the best performing consensus functions. But more important, their ability to keep track of the best input clusterings is notable, as there are only 11 (out of 4440) individual clusterings superior to the λ_{F} consensus clustering derived through MCLA (19 for CSPA). In other words, the proposed hierarchical consensus clustering architecture has yielded (using the MCLA or CSPA consensus functions) a final clustering that is better than the 99.5% of

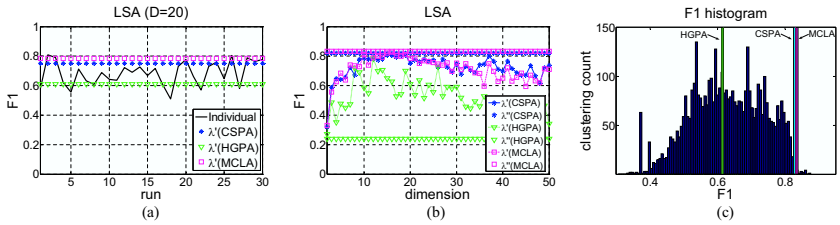

Fig. 2. F1 measure of consensus clusterings: (a) across 30 runs of the KM algorithm, (b) across the $[D^{\min} = 2, D^{\max} = 50]$ dimensionality range, and (c) F1 histogram comparison between the final consensus and individual clusterings.

the individual clusterings, achieving a F1 measure that, in relative terms, is only a 5.5% worse than the best individual clustering. These results reflect high robustness against the data representation and algorithm dependence effects.

4 Conclusions

This paper has presented a novel strategy for robust document clustering by means of an open hierarchical consensus clustering architecture that operates on highly diverse document representations and partitions. As a result, a final consensus labeling comparable to the best input clustering can be obtained, even in the presence of many poor clusterings. Moreover, in comparison to the classic flat consensus approach, the proposed hierarchical architecture delivers better clustering results at a lower computational cost (e.g. execution of the three hierarchical consensus is 7 times faster than a single flat consensus). In fact, the execution of flat consensus via the top performing MCLA function failed due to extremely high memory requirements given the large size of the cluster ensemble.

References

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

[2] Kolenda, T., Hansen, L.K. and Sigurdsson, S.: Independent Components in Text. In: Girolami, M. (ed.): Advances in Independent Component Analysis. Springer-Verlag, Berlin Heidelberg New York (2000) 241–262

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

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

[5] Shafiei, M., Wang, S., Zhang, R., Milios, E., Tang, B., Tougas, J. and Spiteri, R.: A Systematic Study of Document Representation and Dimension Reduction for Text Clustering. Technical Report CS-2006-05. Dalhousie University (2006)

[6] Strehl, A. and Ghosh, J.: Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions. JMLR, Vol. 3, (2002) 583–617