

Partitional Clustering Experiments with News Documents

Arantza Casillas¹, Mayte González de Lena², and Raquel Martínez²

¹ Dpt. de Electricidad y Electrónica, Facultad de Ciencias
Universidad del País Vasco
`arantza@we.1c.ehu.es`

² Escuela Superior de CC. Experimentales y Tecnología
Universidad Rey Juan Carlos
`{mt.gonzalez,r.martinez}@escet.urjc.es`

Abstract. We have carried out experiments in clustering a news corpus. In these experiments we have used two partitional methods varying two different parameters of the clustering tool. In addition, we have worked with the whole document (news) and with representative parts of the document. We have obtained good results working with a representative part of the document. The experiments have been carried out with news in Spanish and Basque in order to compare the results in both languages.

1 Introduction

The document clustering deals with the problem of identifying sets of thematically related documents. Document clustering has been investigated for using in a number of different areas: information retrieval, browsing collections of documents, etc; and a number of techniques have been used [3]. We are investigating the use of clustering techniques for addressing the linking of news documents and we are working in two languages: Spanish and Basque. We have employed partitional methods in our experiments. With partitional methods the clusters generated contain objects that agree with a strong pattern. For example, their contents include some shared words or terms; in each cluster there are objects (news) that share a subset of the dimension space. In this paper we present the results of the experiments that we have carried out with two different news corpus, one in Spanish and the other in Basque. In the next Section we briefly describe the documents; Section 3 describe the used clustering tool, the type of parameters and the experiments; in Section 4 we present the results; finally, section 5 summarizes the conclusions drawn from the work carried out.

2 Documents Description

In the project we are involved [4], we are working with a corpus of categorized news. The categories are the Industry Standard IPTC Subject Codes [2]. We have selected for the experiments the sport category in order to test the clustering

of news of the same category. We have selected 37 news of 6 different sports; in Spanish there are: football 16, baseball 2, swimming 2, athletics 6, cycling 7, and skiing 4; in Basque: football 4, cycling 7, pelota 16, swimming 2, athletics 6, and handball 2. The news corpus in Spanish and Basque are not parallel or comparable. The news selection has been random among news of the first days of the 2000 year. The documents have been preprocess in order to work with the lemmas instead of inflected forms. In addition, the words of a stoplist used in Information Retrieval (with articles, determines, ...) have been eliminated of the Spanish documents.

3 Experiment Description

The tool we have selected for experimenting is CLUTO [1]. In addition to the different classes of clustering algorithms, criterion functions and similarity functions, CLUTO can operate on very large datasets with respect to the number of objects (documents) as well as the number of dimensions. In these experiments we have varied 3 different parameters that control how the tool computes the solution: the method, the similarity function, and the clustering criterion function.

- We have used two methods: RB and RBR. In RB method the k clustering solution is computed by performing $k - 1$ repeated bisections. In each step, the cluster that is selected for further partitioning is that one whose bisection will optimize the value of the overall clustering criterion function the most. The RBR method is similar to the previous one, but at the end the overall solution is globally optimized.
- Two similarity function have been used: COS and CORR. These functions determine how the similarity between objects will be calculated. COS represents the cosine function, and CORR the correlation coefficient.
- We have used three clustering criterion functions: I1, I2, H2. The I2 and H2 functions are told to lead generally to very good clustering solutions (see formulas in [1]).

In order to determine if working with the whole document leads to better results than working with a representative part of the document, we have experimented: (1) with the whole document, (2) only with the title and the first paragraph, and finally (3) with the title and the first paragraph but increasing the weight of the title words. This aspect can be very important in reducing the computational cost of the clustering when a large corpus of news must be clustered.

4 Results

We carried out a manual clustering in order to test the experiments results. The manual clustering consisted of grouping the documents by sport category (football, cycling, ...). This manual clustering is used by the clustering tool in

order to compute the quality of the clustering solution using external quality measures. The tool computes this quality in terms of *entropy* and *purity* (see formulas in [6]). Small entropy values and large purity values indicate good clustering solution.

The results of the experiments can be seen in Table 1 and Table 2. Each table reflects the three best results in connection with *entropy* and *purity* showing the parameters that have been used. In addition, we propose the *coherence* metric in order to show other quality metric of the clustering solution. We consider that a cluster is coherent if it has at least two clearly related objects (news). The percentage of coherence is the percentage of coherent clusters in the solution.

Num. clusters & Part of docu.	Method	Similarity Function	Criterion Function	Entropy	Purity	% Coherence
10 cl. & The whole document	RBR	COS	I2	0.256	0.784	100
	RB	COS	I2	0.320	0.730	100
	RB	COS	H2	0.335	0.703	90
10 cl. & Title, First paragraph	RB	COS	I2	0.292	0.703	80
	RB	CORR	H2	0.298	0.703	80
	RBR	CORR	H2	0.298	0.703	80
10 cl. & First parag. weighted title	RB	COS	H2	0.292	0.676	70
	RBR	COS	I2	0.342	0.676	90
	RB	COS	I1	0.347	0.703	90
6 cl. & The whole document	RB	COS	I2	0.456	0.622	100
	RB	COS	H2	0.460	0.649	100
	RBR	COS	I2	0.461	0.622	100
6 cl. & Title, First paragraph	RB	COS	H2	0.401	0.676	100
	RB	COS	I2	0.463	0.595	100
	RBR	COS	I2	0.466	0.622	100
6 cl. & First parag. weighted title	RB	COS	I1	0.445	0.649	100
	RBR	COS	I1	0.445	0.649	100
	RB	COS	I2	0.476	0.595	100

Table 1. Results of the three best combinations of Spanish document clustering

Working with a number of clusters equal than the number of different sports the news belong to, that is 6, the best results are obtained taken into account only the title and the first paragraph of each news. However, if the number of cluster increases, the best results correspond to the whole document. The best clustering solutions have been obtained in most of the tests with the I2 or H2 clustering criterion functions. With regard to the others parameters, there are appreciable differences among both groups of news. Whereas in Spanish the RB method is the best in most of the cases, in Basque the best is the RBR method. With regard to the similarity function, the cosine function (COS) leads to better results with the Spanish news, whereas the correlation coefficient (CORR) works better in half of the Basque ones.

5 Conclusions

The best clustering solutions have been obtained with different parameters (method and similarity function) in both groups of news. Each type of document and language will require experimentation in order to determine the best combination

Num. clusters & Part of docu.	Method	Similarity Function	Criterion Function	Entropy	Purity	% Coherence
10 cl. & The whole document	RBR	CORR	I2	0.293	0.730	80
	RBR	CORR	I1	0.303	0.730	100
	RB	CORR	I2	0.323	0.676	70
10 cl. & Title, First paragraph	RBR	CORR	H2	0.306	0.730	100
	RBR	COS	H2	0.340	0.703	90
	RB	COS	I2	0.347	0.676	100
10 cl. & First parag. double title	RB	COS	H2	0.368	0.676	100
	RBR	COS	I1	0.376	0.649	70
	RB	COS	I1	0.376	0.649	70
6 cl. & The whole document	RBR	COS	I2	0.513	0.595	100
	RBR	CORR	I2	0.514	0.595	100
	RB	CORR	H2	0.529	0.541	100
6 cl. & Title, First paragraph	RBR	COS	H2	0.446	0.622	100
	RB	COS	H2	0.448	0.649	100
	RBR	COS	I2	0.479	0.541	100
6 cl. & First parag. double title	RBR	CORR	I2	0.495	0.595	100
	RBR	CORR	H2	0.525	0.568	100
	RB	CORR	I2	0.528	0.595	100

Table 2. Results of the three best combinations of Basque document clustering

of parameters. When reducing the computational cost is a critical criteria in a particular clustering task, our experiments show that working with the title and the first paragraph of the news leads to good enough results in entropy in some cases. However, in other domain this conclusion could be uncertain. With regard to the number of clusters, the more clusters there are the entropy metric improves, but the coherence decreases in some cases, so working with the whole document is required in order to obtain better results.

6 Acknowledgments

This research is being supported by the Spanish Research Agency, project HERMES (TIC2000-0335-C03-03).

References

1. “CLUTO. A Clustering Toolkit. Release 2.1”. <http://www-users.cs.umn.edu/~karypis/cluto/>.
2. Industry Standard IPTC Subject Codes. <http://www.sipausa.com/ipctsubject-codes.htm>.
3. A. Gelbukh, G. Sidorov, A. Guzman-Arenas. “Use of a weighted topic hierarchy for text retrieval and classification.” *Text, Speech and Dialogue. Proc. TSD-99. Lecture Notes in Artificial Intelligence, No. 1692, Springer*, 130-135, 1999.
4. “Project HERMES (Hemerotecas Electrónicas: Recuperación Multilingue y Extracción Semántica)” of the Spanish Research Agency, (TIC2000-0335-C03-03). <http://terral.ieec.uned.es/hermes/>.
5. Y. Zhao and G. Karypis. “Evaluation of hierarchical clustering algorithms for document data sets”. *CIKM*, 2002.
6. Y. Zhao and G. Karypis. “Criterion functions for document clustering: Experiments and analysis”. <http://cs.umn.edu/~karypis/publications>.