Clustering of Biomedical Scientific Papers

D. Bravo-Alcobendas, C.O.S. Sorzano
Bioengineering Lab.
Univ. San Pablo - CEU
Campus Urb. Montepríncipe, s/n, Boadilla del Monte, Madrid, Spain
coss.eps@ceu.es

Abstract— **In this paper we present a methodology for document clustering based on Non-negative Matrix Factorization (NMF) and ensemble clustering. Thanks to the ensemble clustering the algorithm is less prone to get into a local minimum caused by the initialization of the NMF. Despite the ensemble clustering, the algorithm keeps the semantic interpretability of the NMF and constructs a coocurrence matrix that allows the projection of the documents onto a two-dimensional space suitable for visualization. The algorithm is freely available for the information retrieval community from the Bioengineering Laboratory web page.**

I. Introduction

Document clustering is a useful technique to quickly overview the main groups of a number of documents. It is of particular interest scientific papers, where information is well structured and usually authors keep working on the same topic for several years publishing several papers related to the same theme. Document clustering can help to perform a first classification of the articles into broad topics that can be further explored by the user. The fact of using a clustering approach instead of a classification approach avoids the need to train the algorithm with a number of documents of prespecified classes. Instead, the documents are naturally split into disjoint (although possibly related) classes.

Documents are usually represented as a high-dimensional vector formed by the word counts in each document. These high-dimensional vectors are not well suited for clustering due to their sparseness in high dimensions. Instead, the document vectors are typically projected onto another vector space of lower dimension before clustering avoiding the sparseness in this way. Non-negative Matrix Factorization (NMF) [1] has been successfully applied for this task in several recent works [2], [3], [4]. However, they all perform a single NMF decomposition of the document matrix and derive conclusions from it. This contrasts with the fact that the typical NMF algorithm [1] is an iterative algorithm rather dependent on its initialization. Therefore, the results of a single decomposition are not fully trustable as "the" decomposition of the document matrix and, consequently, the subsequent clustering should take care of this initialization dependence. This has been done in a general clustering framework using NMF in [5]. In the field of information retrieval, [6] tries to avoid the local minima of NMF by alternating between NMF and Probabilistic Latent Semantic Indexing (PLSI). Alternatively, [7] avoids the local minima by performing several NMF decompositions, constructing a hypergraph with all of them, and finally performing a clustering on this hypergraph. A disadvantage of the approach presented in [7] is that the dimension of the lower dimensional space in the NMF decomposition is of the same size as the desired number of clusters, which is implicitly making the assumption that each cluster will lie in its own NMF dimension. In this way, it is assumed that there is no cluster that is a mixture of different topics (e.g., articles whose composition is 50% about a theme (for instance, image registration) and 50% about a different theme (for instance, MRI images)). Moreover, the final clustering loses the semantic topics extracted by each individual NMF, i.e., the documents are correctly clustered but it is not possible to specify the words defining the topics of each one of the clusters.

In this article, we present a document clustering algorithm that takes care of the NMF dependency on the initial decomposition, and that at the end is able of explaining each one of the clusters as a linear combination of semantic topics clearly defined in terms of key words. Additionally, the intermediate data structure used by our algorithm allow a visual inspection of the relative locations of the document clusters. We have tested our algorithm with a short corpus of biomedical articles (journal and conference papers) instead of the traditional article or news abstracts. We find our algorithm quite appropriate as a quick overview tool that clusters biomedical papers into different non-disjoint topics (i.e., a paper may be about Topic 1 and Topic 2 with an energy distribution of 80% and 20%). We also provide a freely accessible software tool that can be obtained from `http://biolab.uspceu.com`.

II. Document preprocessing

The space model vector for information retrieval idea was introduced in [8]. In this model each document is represented as a vector of the frequency of terms contained in it, single words are the simplest and most common representation for documents, but n-grams (groups of n words) can also be used. In this article we use the simple representation based on isolated words.

The words contained in the documents must be filtered to remove words that cannot help to distinguish between themes; in this set we include words such as prepositions, determinants or pronouns. So the first step in our preprocessing is to eliminate these stop-words from the documents to be analyzed.

Our second preprocessing step is to reduce all the words to their stems decreasing so the variability due to different verb conjugations (present, past, gerund), plural or singular appearances of the terms, etc. The set of all stemmed words from the document set to be analyzed constitute the basis on which each document is spanned, i.e., each document is represented by a vector of frequencies of the stemmed words. All those stemmed words with less than three letters are removed from the basis.

Finally, it has been shown [9] that working with frequencies alone is not enough to successfully discriminate between the different classes. For instance, suppose that there is a word that appears in only one document, it cannot be used to cluster similar documents since it does not appear in any other document. On the other extreme, let us assume that in our collection of papers the word "Biomedical" appears in all of them, then this word has no discriminative power to elucidate the theme of any of the articles. To avoid the first effect we remove all words that do not appear in a given minimum number of documents (in our experiments a word must appear in at least 10% of the documents to be considered; however, further exploration is needed to fully determine the effect of this threshold, especially with large datasets). Fig. 1 shows the number of terms that appear in at least n documents for the experiment described in the Results Section.

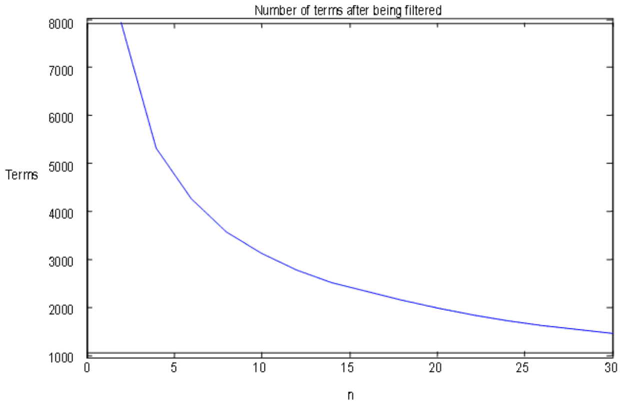


Fig. 1. Number of terms appearing in at least n documents.

To avoid the second effect, the term frequency is translated into a term relevance as follows:

$$tr = tf \log\left(\frac{N}{df}\right) \tag{1}$$

where N is total number of documents in the corpus being analyzed, df is number of documents containing the term, and tf the term frequency. This conversion is also known as the inverse document frequency or IDF [9].

III. Dimensional reduction

The vector dimension of the documents after preprocessing is in the order of thousands (2000-3000 terms typically). Moreover, this dimension increases considerably with the number of documents to evaluate. Documents in these high dimensional spaces are very sparsely scattered and clustering them is extremely difficult (this is known as the curse of dimensionality [10]). For this reason it is essential to perform a treatment to reduce the dimensionality of the dataset while maintaining as much information as possible, this step is known as dimensionality reduction.

Non-negative Matrix Factorization (NMF) is a linear and non-negative representation of a dataset that has been successfully applied in the field of information retrieval for the purpose of identifying semantic topics [2], [3], [4], [6], [7] as follows. Let us call D the matrix formed by the term relevancies computed in the previous section. Each document vector is one of the columns of this matrix. NMF decomposes this matrix as the product of two other matrices with non-negative elements (see Fig. 2)

$$D \approx TD' \tag{2}$$

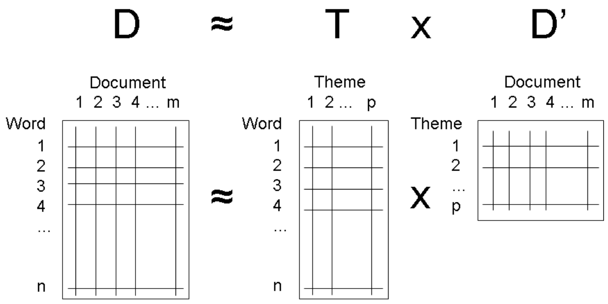


Fig. 2. Schematic representation of the matrix decomposition performed by NMF.

The documents in matrix D are n-dimensional while in D' are p-dimensional, with $p \ll n$. Interestingly, this decomposition provides a basis for the documents that can be interpreted as the different themes of the documents, and the document representation in D' is simply a non-negative linear combination of the different themes. The NMF problem addresses the minimization of the reconstruction error (squared Frobenius norm of the difference between the original document matrix and the reconstructed one):

$$T^*, D'^* = \arg\min_{T,D'} \|D - TD'\|_F^2 \tag{3}$$

Exact recovery is not possible unless a high number of themes is employed, but this latter condition makes the document clustering more difficult due to the sparseness of high-dimensional spaces. Fig. 3 shows the reconstruction error (measured as given in Eq. (3)) as a function of the number of themes. It can be seen that for a medium number of themes (between 10 and 50), the reconstruction error decreases nearly linearly with the number of themes.

Unfortunately, the NMF decomposition is not unique [7] and, moreover, the iterative algorithm employed has a strong dependency with the initial decomposition (which is usually initialized at random). To avoid this negative effect, we propose to perform several NMF decompositions, each one with a different random initialization. For each NMF

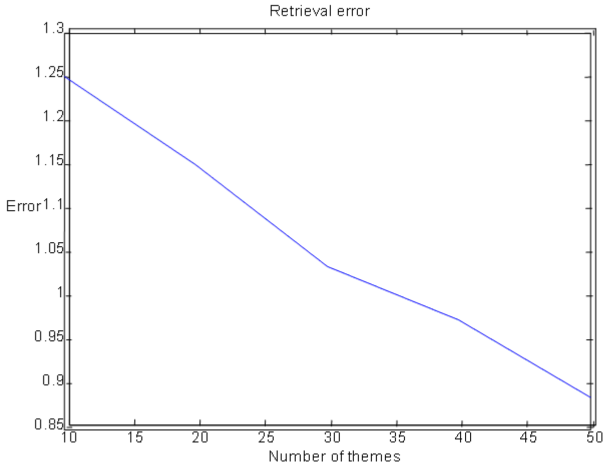


Fig. 3. Dependency of the reconstruction error with the number of themes in the NMF decomposition.

decomposition, we propose to cluster the documents in their new representation (D'). This step of the algorithm is further explained in the next section.

IV. Document Clustering

Because of the dependency of the NMF algorithm with its initial values, a clustering of documents at the level of D' with a single run is not reliable. We propose to perform several runs of NMF with random initializations, and then a bisecting K-means clustering [11] for each NMF decomposition. The input to K-means are the columns of D' normalized such that each component represents the fraction of the contribution of the i-th theme to the total energy of a given document, i.e.

$$d''_{ij} = \frac{(d'_{ij})^2}{\sum\limits_{j=1}^{p} (d'_{ij})^2}, \tag{4}$$

where d'_{ij} is the ij-th element of the D' matrix obtained after NMF decomposition.

Bisecting K-means also depends on its initialization, although to a less extent than the standard K-means and, therefore, several runs with random initializations must be performed. We propose to collect all the different results (random NMFs and random K-means) in a single coocurrence matrix C. This matrix is a symmetric $m \times m$ matrix whose ij-th element represents how many times documents i and j were grouped in the same cluster.

The coocurrence matrix can be easily converted into a dissimilarity matrix by $C' = I - \frac{C}{\|C\|_{\max}}$, where I is the identity matrix of size $m \times m$ and $\|C\|_{\max}$ represents the matrix norm known as maximum. At this point, we propose to perform an agglomerative hierarchical clustering (AHC) [11] to finally obtain the document clusters. Because of the averaging in the construction of the coocurrence matrix, this final clustering tends to be independent of the initialization of the NMF and the K-means algorithms. Additionally, the coocurrence matrix allows the projection of the document structure in the low dimensional thematic space onto an even lower dimensional space (we use a 2D space) using Multidimensional Scaling (MDS), a classical multivariate data analysis technique employed for data visualization. In this way, the original documents as well as their relative distances can be easily visualized.

Finally, note that in our algorithm the number of themes used in the NMF decomposition, the number of clusters in the bisecting K-means, and the number of clusters in the final AHC need not to be the same. This fact gives us more flexibility for defining the document clusters.

V. Thematic decomposition of documents and clusters

Using the coocurrence matrix for the document clustering allows us to perform a reliable clustering with a smaller dependence on the particular initializations used for the NMF and the bisecting K-means. However, the final clusters lack of an intuitive interpretation of the results since we have lost the representation of each one of the documents in the cluster as a linear combination of themes. To avoid this problem we propose to use the NMF with the lowest representation error (Eq. (3)) among the different NMF decompositions performed to construct the coocurrence matrix. In this way, we can also attach a semantic meaning to each of the documents through the matrix D' and to each of the clusters (usually clusters are formed by documents with similar thematic energy composition). For each theme, we show the most important words as those having higher energy in the T matrix.

VI. Results

To test the efficacy of our method we have collected all journal papers from three different groups (the Biomedical Imaging Group of the Swiss Federal Institute of Technology, the Bioengineering Laboratory of the Univ. San Pablo - CEU, and the Biomedical Imaging Technologies of the Polytechnical Univ. of Madrid) working in biomedical imaging and with several papers in common. This dataset is specially challenging since the algorithm must discover subtle thematic differences since all of them work on biomedical imaging.

195 articles were collected from the three laboratories. At the beginning there were 25132 stemmed words. However, after removing those words not appearing in at least 10% of the documents, only 1876 stemmed words survived. An interesting result of the document preprocessing is the most relevant words along the dataset. Table I shows the ten most important words in our corpus.

We run our algorithm with 20 themes for the NMF (10 repetitions of the NMF decomposition), 32 clusters for the k-means (100 repetitions), and 25 clusters for the AHC. Table II shows the two most relevant words for each of the themes. It is interesting to note that these themes certainly cover the thematic spectrum of the articles analyzed.

Table III shows the different identified clusters. These clusters have been manually interpreted making use of the

Word	$\sum_{all\_documents} tr_{document}$
CTF	939.02
Velocity	876.98
Registration	832.66
Spline	748.41
Denoise	713.93
Deformable	671.71
PET	652.43
Wavelet	632.14
Detectors	619.18
Myocardium	615.59

TABLE I

SUM OF THE TERM RELEVANCE OVER ALL DOCUMENTS FOR MOST RELEVANT WORDS IN THE CORPUS ANALYZED.

Theme	Words
1	Contours, curve
2	Interpolation, kernels
3	Theorem, fractional (spline)
4	Denoising, wavelets
5	Detectors, pulse
6	CTF, electron
7	Dirac, Optical
8	Box, spline
9	Map, databases
10	Spline, interpolation
11	Reconstruction, tilting
12	Dots, detection
13	Patient, PET
14	Motion, elastic
15	Registration, deformation
16	Cell, protein
17	Particle, micrographs
18	Wavelet, resampling
19	Velocity, motion
20	Spline, wavelets

TABLE II

TWO MOST IMPORTANT (MAXIMUM ENERGY) WORDS OF EACH OF THE THEMES EXTRACTED BY THE MINIMUM ERROR (EQ. (3)) NMF.

thematic composition of the documents assigned to them. The articles assigned to the cluster have been analyzed and the number of incorrectly assigned articles is reported.

We can see that documents have been correctly clustered in 93.3% of the cases. There are some wide topics (e.g. Cardiac motion, Electron microscopy) that have been split into two different groups, with the most similar articles in the same group. Outlier articles have been assigned to three clusters so that these clusters cover those articles that are not similar to any of the rest groups.

This analysis has been carried out with a Java software that permits to perform all computations as well as navigating through the clusters, themes, theme words, open the corresponding articles, etc. (see Fig. 4). Thanks to the Java development, the software can run on any platform supporting the Java Virtual Machine. This software is freely available from `http://biolab.uspceu.com`.

#	Definition	Assigned	Errors
1	Cardiac motion	14	0
2	Cardiac motion	2	0
3	Cardiac reconstruction	3	0
4	Difficult to classify	3	0
5	Sampling with splines	5	1
6	Sampling	11	1
7	Databases	4	1
8	Sampling theory	13	2
9	Wavelet theory	5	0
10	Difficult to classify	1	0
11	Difficult to classify	3	0
12	Applications of wavelets	7	0
13	Interpolation	5	0
14	Applications to optical systems	7	3
15	Electron microscopy	5	2
16	Fluorescence microscopy	9	0
17	3D Reconstruction	20	0
18	Contour extraction	5	0
19	Electron microscopy	9	0
20	Spline applications	8	1
21	Denoising	11	1
22	PET and MRI scanners	11	0
23	Image registration	15	1
24	Clinical applications	10	0
25	Wavelet theory	9	0

TABLE III

LABEL MANUALLY ASSIGNED TO EACH ONE OF THE IDENTIFIED CLUSTERS, NUMBER OF ARTICLES ASSIGNED TO THAT CLUSTER BY THE ALGORITHM, AND NUMBER OF ARTICLES WRONGLY ASSIGNED TO THE CLUSTER.

VII. CONCLUSIONS

In this article we have presented an algorithm for clustering scientific articles that is able to produce document clusters as well as a thematic interpretation of each document. This thematic interpretation and clustering is based on NMF decomposition. We have taken care of the effects of the random initialization of the NMF as well as the intermediate bisecting K-means clustering. In our experiment, the algorithm was able to correctly identify the paper topics even if all of them were related to the very specific field of biomedical imaging.

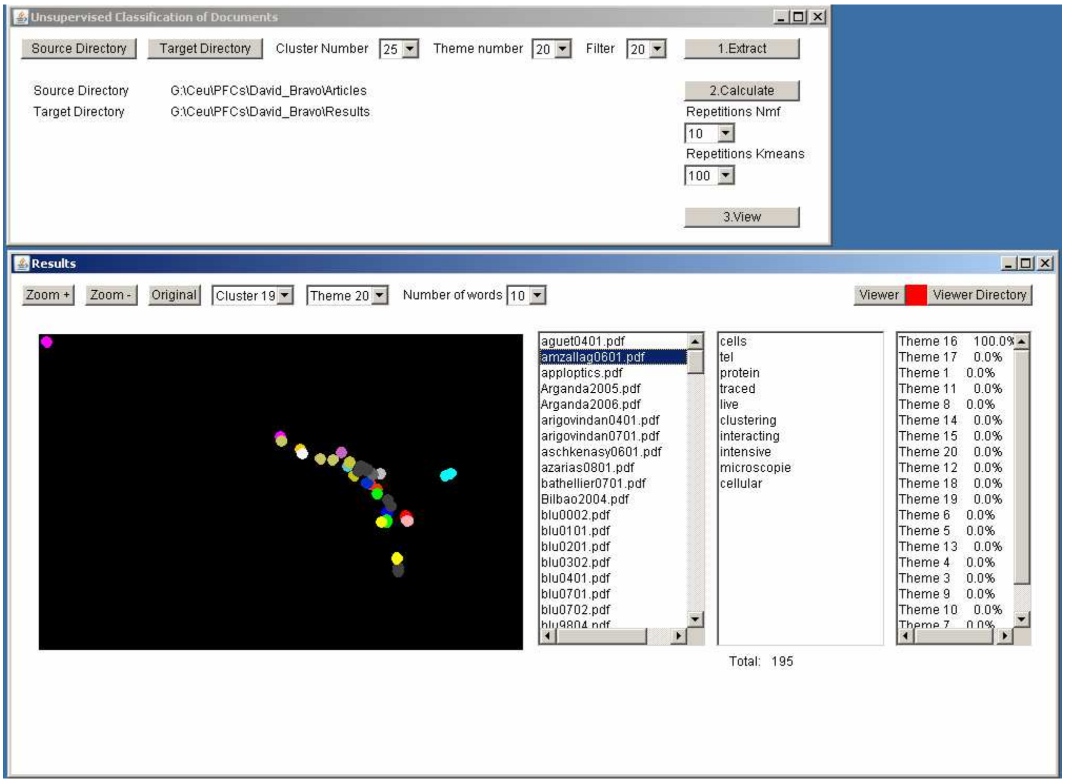
Acknowledgments

This work was funded by the European Union (3DEM Network of excellence FP6-502828 and UE-512092), the Spanish National Council of Scientific Research (CSIC, PIF08-020), the Spanish Ministerio de Ciencia e Innovación (CSD2006-0023, BIO2007-67150-C01, BIO2007-67150-C03, and BFU2007-60228), the Spanish Fondo de Investigaciones Sanitarias (04/0683), the Universidad San Pablo CEU (USP-PPC 04/07), and the Comunidad Autónoma de Madrid (S-GEN-0166-2006, CCG08-CSIC/SAL-3442) and by Award Number R01HL070472 from the National Heart, Lung, And Blood Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, And Blood Institute or the National Institutes of Health.

REFERENCES

[1] D. D. Lee and S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, pp. 788–791, 1999.

[2] W. Xu, W. Liu, and Y. Gong, “Document clustering based on non-negative matrix factorization,” in *Proc. of ACM Special Interest Group in Information Retrieval*, Toronto, Canada, August 2003.



Cluster Number 25 Theme number 20 Filter 20

Source Directory G:\Ceu\PFCs\David_Bravo\Articles

Target Directory G:\Ceu\PFCs\David_Bravo\Results

Repetitions Nmf 10

Repetitions Kmeans 100

Cluster 19 Theme 20 Number of words 10

aguet0401.pdf
amzallag0601.pdf
apploptics.pdf
Arganda2005.pdf
Arganda2006.pdf
arigovindan0401.pdf
arigovindan0701.pdf
aschkenasy0601.pdf
azarias0801.pdf
bathellier0701.pdf
Bilbao2004.pdf
blu0002.pdf
blu0101.pdf
blu0201.pdf
blu0302.pdf
blu0401.pdf
blu0701.pdf
blu0702.pdf
blu9804.pdf

cells
tel
protein
traced
live
clustering
interacting
intensive
microscopie
cellular

Theme 16	100.0%
Theme 17	0.0%
Theme 1	0.0%
Theme 11	0.0%
Theme 8	0.0%
Theme 14	0.0%
Theme 15	0.0%
Theme 20	0.0%
Theme 12	0.0%
Theme 18	0.0%
Theme 19	0.0%
Theme 6	0.0%
Theme 5	0.0%
Theme 13	0.0%
Theme 4	0.0%
Theme 3	0.0%
Theme 9	0.0%
Theme 10	0.0%
Theme 7	0.0%

Total: 195

Fig. 4. User interface of the program making the document clustering. The results window allows us to navigate through the different clusters, see their word and thematic compositions as well as observing the distance between the different clusters (each point in the black window of the results interface is the multidimensionally scaled projection of each article, relative distances in this projected space are related to relative distances in the coocurrence matrix).

[3] F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons, "Document clustering using nonngative matrix factorization," *Information processing and management*, vol. 42, pp. 373–386, 2006.

[4] H. Shinnou and M. Sasaki, "Ensemble document clustering using weighted hypergraph generated by nmf," in *Proc. of the 45th Annual Meeting of the Association of Computational Linguistic*, 2007.

[5] T. Li, C. Ding, and M. I. Jordan, "Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization," in *Proc. 7th IEEE Intl. Conf. on Data Mining*, 2007.

[6] C. Ding, T. Li, and W. Peng, "Nonnegative matrix factorization and probabilistic latent semantic indexing: equivalence, chi-square statistic, and a hybrid method," in *Proc. of the 21st National Conference on Artificial Intelligence (AAAI)*, 2006.

[7] S. Zhu, W. Yuan, and F. Wang, "Ensemble non-negative matrix factorization for clustering biomedical documents," in *Proc. of the 2nd Intl. Symposium on Optimization and Systems Biology*, 2008.

[8] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, pp. 613–620, 1975.

[9] C. Manning and H. Schütze, *Foundations of Statistical Natural LanguageProcessing*, MIT Press, Cambridge, MA, USA, 1999.

[10] R. W. Bellman, *Dynamic programming*, Princeton University Press, 1957.

[11] C. M. Bishop, *Pattern recognition and machine learning*, Springer, 2006.