Spectral clustering and community detection in document networks

C. K. dos Santos, A. G. Evsukoff & B. S. L. P. de Lima
COPPE/UFRJ - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

Abstract

Document clustering is one of the most active research topics in text mining. In this work two approaches issued from very different fields are explored for document clustering: spectral clustering and community detection in complex networks. Both approaches are based on a representation of the document collection as a graph, of which the nodes represent the documents and the edges represent the similarities between each pair of documents, such that the two approaches have many issues in common. The results of the application of these two types of techniques to benchmark text mining problems show that they are complementary and are useful for finding structure in large collections of documents

Keywords: text mining, document clustering, spectral clustering, community detection, complex networks, modularity.

1 Introduction

Unstructured information in document databases presents intrinsic characteristics such that the classical machine learning algorithms must be adapted to solve text mining tasks. The most usual representations for text mining rely on the vector space information retrieval model of documents [1]. In such a model the order of words is not considered and each document in a collection is represented by a vector, of which the components are related to relevant words appearing in the document collection.

The high dimensionality issued from the vector space representation of document collections is one of the challenging problems in text mining research, especially for document clustering. Spectral clustering algorithms provide

WIT Transactions on Information and Communication Technologies, Vol 42, © 2009 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA090051

robustness to high dimensionality and sparsity of feature space defined over the vector space model used to represent documents. The main tools of spectral clustering are derived from spectral graph theory and, in recent years, it has been established as an efficient and meaningful alternative to traditional clustering techniques [2].

On the other hand, the representation of real systems as complex networks has been studied in widely different fields from molecular biology to internet and social sciences [3–5]. One of the most active areas in the study of complex network is the detection of community structure in networks [6–12], which has many applications in social and biological sciences.

The algorithms for community structure detection in complex networks and spectral clustering are both based on the graph theory, such that they present similarities and complementarities. The algorithms are usually formulated as graph partition problem where the weight of each edge is the similarity between points that correspond to vertices connected by the edge. The goal of that algorithm is to find the minimum weight cuts in the graph, but this problem can be addressed by means of linear algebra, in particular by the eigenvalues decomposition techniques.

These two types of algorithms are investigated in this work in the context of document clustering. The document collection can be viewed as a complex network, i.e. a graph, of which the nodes are the documents and the edges are weighted according to document similarities. A document cluster can be regarded as a community structure in complex network analysis, such that the best clustering assignments can be determined from the one resulting in the best community structure in the network. A robust approach to this problem is the maximization of the function known as "modularity", introduced by Newman and Girvan [8]. The main contribution of this work is to show that the two kinds of algorithms are complementary such that both techniques should be used together.

Next section presents the main issues of spectral clustering algorithms. Section 3 introduces an algorithm for community detection in complex networks that is very similar to spectral clustering approach. Section four presents results in two benchmark document collections. The paper ends with concluding remarks and future direction of this research project.

2 Spectral clustering of document networks

Document pre-processing and vector-space representation [1] result in a table, often called "Bag of Words" (BoW), of which the lines are related to the documents and the columns are related to the words (terms) that appear in the entire collection.

Although it is generally more interesting to store the BoW using special data structures due to dimensions involved, mathematically it can be considered as the $n \times m$ sparse matrix $\mathbf{X}$ of which the lines are related to the documents and the columns are related to the terms. An element x_{ij} accounts for how the term is related to the document, often computed by the tf-idf frequency [13].

A document network can be defined from $\mathbf{X}$ as a complete and weighted undirected graph $G(V,E)$, of which the set of vertices $V=\{v_1,...,v_n\}$ corresponds to the n documents and the set of edges E is defined through the symmetric adjacency matrix $\mathbf{A}\in\Re^{n\times n}$. Each element entry $a_{ij}\in\mathbf{A}$ represents the pair-wise similarity between the documents D_i and D_j, computed as:

$$a_{ij}=\begin{cases}h(\mathbf{x}_i,\mathbf{x}_j) & \text{if } i\neq j,\\ 0 & \text{otherwise.}\end{cases} \quad (1)$$

The function h measures the local neighbourhood relationships among their vertices and may be computed by different functions. In this work, this similarity function is computed by the Gaussian similarity:

$$h(\mathbf{x}_i,\mathbf{x}_j)=\exp\left(-\frac{\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2}{2\sigma^2}\right) \quad (2)$$

where the parameter σ controls the width of the neighbourhoods. This parameter plays an important role in the graph structure.

The degree d_i of a vertex $v_i\in V$ is the number of edges incident to the vertex and is defined as:

$$d_i=\sum_{j=1}^{n}a_{ij} \quad (3)$$

The graph cut problem aims to separate a subset of vertices $S\subset V$ from its complement $V-S$ denoted by $\overline{S}$ [14]. The graph cut problem can be formulated in several different ways, depending on the choice of the objective function to be optimized [2]. One of which is the cut function [14], whose minimization favours partitions containing isolated vertices, and is defined as follows:

$$cut(S,\overline{S})=\sum_{i\in S,j\in\overline{S}}a_{ij} \quad (4)$$

To overcome the weakness of the cut function and achieve better balance between partitions it is suggested to use its normalized version, that is, the normalized cut function [14]:

$$Ncut(S,\overline{S})=cut(S,\overline{S})\left(\frac{1}{vol(S)}+\frac{1}{vol(\overline{S})}\right) \quad (5)$$

where $vol(S)$ is the volume of S, computed as:

$$vol(S)=\sum_{i\in S}d_i \quad (6)$$

The minimization of the function (5) is a NP-hard problem [15], which can be relaxed by introducing the graph Laplacian matrix [2, 14].

$$\mathbf{L}=\mathbf{D}-\mathbf{A} \quad (7)$$

where the degree matrix $\mathbf{D}$ is defined as the diagonal matrix of the degrees $d_1,...,d_n$.

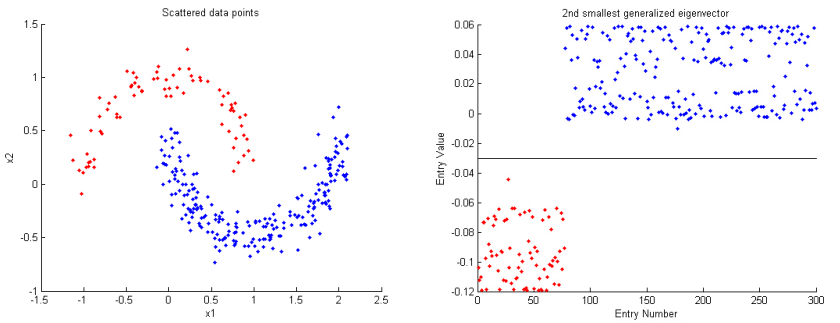
The graph Laplacian is a positive semi-definite matrix, such that its eigenvalues are always positive real-valued. Shi and Malik [15] suggested an approach based on thresholding of the second smallest eigenvector of the generalized eigenvalue problem:

$$\mathbf{LU} = \mathbf{\Lambda DU} \tag{8}$$

where $\mathbf{\Lambda}$ is the diagonal matrix of the eigenvalues, which are ordered in ascending order, $0 \le \lambda_1 \le \lambda_2 \le ... \le \lambda_n$. The orthogonal matrix $\mathbf{U}$ is the matrix of which the columns are the generalized eigenvectors.

The second smallest eigenvector computed as the solution of (9) can be used to split the graph into two clusters by using a threshold value.

The result of this approach in the 2D two moons problem is shown in Figure 1 with 300 data points generated with Gaussian noise of variance 0.01. It can be seen that for this problem, the solution of the approximate normalized cut problem (8) allows to separate the two clusters with facility.



(a) (b)

Figure 1: (a) Scattered data points; (b) entry values of the second smallest generalized eigenvector ordered according to the classes.

Recent approaches use more than one eigenvector to build an embedded of data and then cluster the points with the *k*-means algorithm [16]. Shi and Malik [15] propose an algorithm that clusters the graph in *k* groups as:

Build the affinity matrix $\mathbf{A}$ using (1);

Construct the graph Laplacian matrix $\mathbf{L}$ as (8);

Compute the first *k* eigenvectors $u_1,...,u_k$ of the generalized Eigen-problem (9);

Let $\mathbf{Y} \in \Re^{n \times k}$ be the matrix containing the vectors $u_1,...,u_k$ as vector column;

Cluster the points of $\mathbf{Y}$ with the *k*-means algorithm.

Good clustering algorithms for graph clustering depend on the quality of the objective function being used. Recently, a cost function was proposed by Newman and Girvan, named modularity function [8], to overcome limitations of the previous measures for measuring community structure, as discussed in next section.

3 Modularity-based community detection

3.1 The modularity and the community structure in networks

A community structure in a network $G(V,E)$ is defined as a partition P_K of the set of vertices in K subsets $C_j, j=1...K$, such that $\bigcap_{j=1...K} C_j = \varnothing$ and $\bigcup_{j=1...K} C_j = V$.

One can think about group structure in graph clustering problems as clusters with high density of edges within them, and a lower density of edges among them.

Newman and Girvan [8] defined a quantitative measure to evaluate an assignment of nodes into communities called modularity, which can be used to compare different assignments of nodes into communities. The network modularity Q is defined over a network partition P_K as:

$$Q(P_k) = \sum_{j=1}^{K} \left(\frac{R(C_j, C_j)}{R(V,V)} - \left(\frac{R(C_j, V)}{R(V,V)} \right)^2 \right) \tag{9}$$

where $R(C', C'') = \sum_{i \in C', j \in C''} a_{ij}$ measures the association among the nodes of the subsets C' and C''. Thus, $R(C_j, C_j)$ measures the within-community sum of edge weights, $R(C_j, V)$ measures the sum of weights over all edges attached to nodes in community C_j and $R(V,V)$ is the normalization term that measures the sum over all edge weights in the entire network [9]. Considering binary weighs, the first term $R(C_j, C_j)/R(V,V)$ is the empirical probability that the both vertices of a randomly selected edge falls in subset C_j. The second term $\left(R(C_j, V)/R(V,V)\right)^2$ is the empirical probability that only one of the ends (either one) of a randomly selected edge falls in subset C_j. Thus, the modularity measures the deviation between observed cluster structure and what it could be expected under an independent random model. If the number of within-community edges is no better than random, then the value $Q = 0$. A value $Q = 1$, which is the maximum, indicates strong community structure. In practice however, values typically fall in the range from about 0.3 to 0.7 [8].

In the next subsection, the modularity function will be reformulated as a spectral optimization problem, according to Newman [17].

3.2 The spectral modularity optimization method

Consider a document network represented by the graph $G(V,E)$, as described above, and suppose a particular partition of the G into two groups $S \subset V$ and

its complement $V - S$, denoted by $\overline{S}$. The partition is defined by the partition vector $\mathbf{q} = (q_1, \ldots, q_n)$, such that $q_i = 1$ if vertex $v_i \in S$ and $q_i = -1$ if vertex $v_i \in \overline{S}$.

The expected edge weight p_{ij} between vertices v_i and v_j when edges are placed at random is computed by [17]:

$$p_{ij} = \frac{d_i d_j}{2m} \tag{10}$$

where d_i and d_j are the degrees of the vertices v_i and v_j as defined by (3) and m measures the sum of all edge weights in the entire network:

$$m = \frac{1}{2} \sum_{i=1..n} d_i \tag{11}$$

The modularity measure Q can be written as the sum of the differences between a_{ij} and p_{ij} over all pairs of vertices v_i and v_j that fall in the same community:

$$Q = \frac{1}{4m} \sum_{\substack{i=1\ldots n \\ j=1\ldots n}} (a_{ij} - p_{ij}) q_i q_j \tag{12}$$

which is written in the matrix forma as:

$$Q = \frac{1}{4m} \mathbf{q}^{\mathrm{T}} \mathbf{B} \mathbf{q} \tag{13}$$

where $\mathbf{B}$ is a real and symmetric matrix, called modularity matrix, of which the elements are computed as:

$$b_{ij} = a_{ij} - p_{ij} \tag{14}$$

The maximization of the modularity (13) is equivalent to a graph cut problem such that an approximate solution can be computed by the spectral decomposition of $\mathbf{B}$:

$$\mathbf{B}\mathbf{z} = \mathbf{z}\beta \tag{15}$$

where $\mathbf{z}$ is the eigenvector corresponding to the largest eigenvalue β. The approximate solution corresponds thus to the maximization of the Rayleigh quotient:

$$\hat{Q} = \frac{\mathbf{z}^{\mathrm{T}} \mathbf{B} \mathbf{z}}{\mathbf{z}^{\mathrm{T}} \mathbf{z}} \tag{16}$$

A partition of the network is computed by maximizing the modularity Q by choosing appropriated values for the partition vector $\mathbf{q}$ according to the sign of the components of the eigenvector $\mathbf{z}$:

$$q_i = \begin{cases} +1 & if\ z_i \geq 0 \\ -1 & if\ z_i < 0 \end{cases} \tag{17}$$

The partition vector defined by (17) divides the network in only two communities. However, many networks contain more than two communities,

such that the approach can be applied recursively to find a partition of the network in more than two communities.

The idea of the recursive algorithm is to evaluate the gain in the modularity function if a community is further divided. For each group C' generated by a partition like (17), the additional contribution to the modularity ΔQ is computed as:

$$\Delta Q = \frac{1}{4m}\mathbf{q}'^{\mathrm{T}}\hat{\mathbf{B}}'\mathbf{q}' \tag{18}$$

where $\hat{\mathbf{B}}'$ is the matrix corresponding to the vertices that belongs to the group C' and $\mathbf{q}'$ is the partition vector that will subdivide the group C'.

The group modularity matrix $\hat{\mathbf{B}}'$ is computed as a fraction of $\mathbf{B}'$, the sub-matrix of $\mathbf{B}$ corresponding to the vertices that belong to the group C' as:

$$\hat{b}'_{ij} = b_{ij} - \delta_{ij}\sum_{k\in C'} b_{ik} \tag{19}$$

where $b_{ij} \in \mathbf{B}$ are the elements of the modularity matrix computed as (14), $b'_{ij} \in \hat{\mathbf{B}}'$ are the elements of the modularity matrix corresponding to the partition of the group C', of which the (i, j) indexes refers to the nodes of the entire network, and δ_{ij} stands for the Kronecker δ.

The recursive process is halted if there is no further division of a sub network that will increase the modularity of the network, such that there is no gain in further divide the network. In practice, the test $\Delta Q > \varepsilon$ is used as stopping criterion.

4 Results

In order to evaluate the effectiveness of the techniques presented above, two benchmarks documents collection were used: the 7sectors corpus (7SEC) and the 20 newsgroups corpus (20NG) taken from the CMU World Wide Knowledge Base Project (http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-11/www/wwkb/).

The 7sectors (http://www.cs.cmu.edu/afs/cs/project/theo-20/www/data/bootstrappingIE/7sectors.tar.gz) is a collection of web pages belonging to companies from various economic sectors and contains 4,581 html articles partitioned in a hierarchical order. Each document in this collection was labelled with the label in the first level. The 20 newsgroups (http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/news20.tar.gz) contains 20,000 Usenet articles taken from twenty newsgroups at random (1,000 documents per newsgroup) and it is labelled by their newsgroup name. In the experiments described below, 20% of documents of each newsgroup were extracted to form a new corpus of the 4,000 documents.

After the usual text pre-processing, removing all html tags, headers, non alpha-numeric characters, a stop list, etc. and stemming, a feature selection

method was used to discard terms with weak influence on the document collection resulting in 1,994 terms for the 7sectors corpus and 1,964 terms for the 20 newsgroups.

The clustering results were evaluated by the modularity (9) and the following metrics:

- *Entropy:*

$$\mathrm{E}_j = \frac{1}{\log H} \sum_{t=1}^{H} \frac{n_{jt}}{n_t} \log \frac{n_{jt}}{n_t} \tag{20}$$

- *Purity:*

$$\Pi_j = \frac{1}{n_j} \mathrm{Max}_t (n_{jt}) \tag{21}$$

where the index j runs over all K communities. H is the number of available classes. n_j is the total number of data points in the community C_j and n_{jt} is the number of data points of class t (i.e., labels) positioned in community C_j.

The entropy measure deals with the distribution of classes in a community. A zero-value entropy indicates that community consists exclusively of instances belonging to a single class. In the other hand, the entropy values close to one mean that community contains uniform mixture of classes, hence produces a bad clustering.

The purity function (21), in turns, computes the ratio of dominant class size in the community to the total size of community. High purity value implies better quality of clustering.

The both approaches have been applied to the two corpora and the results of the cluster metrics are shown in Table 1. The Newman method has been computed first, using a threshold value $\varepsilon = 0.01$ to halt the recursive division. The number of groups found by the Newman algorithm is used in the spectral algorithm. The results show that each one of the algorithms has better values in each example.

The modularity values for both problems are within the range from about 0.3 to 0.7 usually found in other real networks [8], which validate the representation of documents as a complex network. The other cluster metrics values are very similar and each of the algorithms has obtained a better performance in one problem.

Table 1: Clustering results where the clusters number *K* of the spectral method is defined by the communities number returned by the Newman method. Best values are in bold face.

Corpus	*K*	Purity		Entropy		Modularity	
		Newman	Spectral	Newman	Spectral	Newman	Spectral
7SEC	7	**0.3713**	0.3222	**0.8550**	0.8915	**0.4628**	0.4455
20NG	10	0.2030	**0.2169**	0.8547	**0.8049**	0.5730	**0.5806**

5 Conclusion

This work has presented two approaches for documents clustering based on the representation of a document collection as a network of documents, of which the nodes represent the documents and the edges represent pair-wise similarities between documents.

The first approach is the spectral clustering approach that has been recently investigated within the pattern recognition community as an approach capable to deal efficiently with high dimensionality problems. The second approach is issued from the theory of complex networks and has been studied mainly by the physics community in a wide range of applications.

Both approaches are based on the same similarity or affinity and are computed by the solution of an eigenvalue problem formulated on a transformation of the affinity matrix. The community structure approach allows the definition the number of clusters, which is a recurrent problem in cluster analysis. Moreover it is recursive, such that it can deal with hierarchical structure, usually found in document clustering problems.

The application on two benchmark problems show that the two approaches has similar results; each one has had better performance in one of the problems. Nevertheless, the Newman algorithm is able to compute the number of groups, what is an input value for the spectral algorithm. These preliminary results indicate that there is much to gain in further exploring kind of representation of document collections and connections of these kinds of algorithms.

Acknowledgements

The authors are grateful to the Brazilian Research Agencies, CNPq, FINEP and FAPERJ, for the financial support for this research.

References

[1] W.B. Michael, D. Zlatko, and R.J. Elizabeth, “Matrices, Vector Spaces, and Information Retrieval,” SIAM Rev.1999, pp. 335-362.
[2] U. von Luxburg, A tutorial on spectral clustering, Statistics and Computing 17(4), arXiv:0711.0189v1, pp. 395-416, 2007.
[3] A.-L. Barabási, Linked: how everything is connected to everything else and what it means for business, science, and everyday life, Plume, 2003.
[4] M.E.J. Newman, “The Structure and Function of Complex Networks,” SIAM Review, pp. 167-256, 2003.
[5] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.U. Hwang, “Complex networks: Structure and dynamics,” Physics Reports, pp. 175-308, 2006.
[6] M.E.J. Newman, “Finding community structure in networks using the eigenvectors of matrices,” doi: 10.1103/PhysRevE.74.036104, 2006.
[7] M.E.J. Newman, “Modularity and community structure in networks,” PNAS 0601602103, 2006.

[8] M.E.J. Newman, and M. Girvan, "Finding and evaluating community structure in networks," Phys. Rev. E 69, 026113, 2004.
[9] S. White, and P. Smyth, "A spectral clustering approach to finding communities in graphs," SIAM International Conference on Data Mining, pp. 76-84, 2005
[10] E.A. Leicht, and M.E.J. Newman, "Community structure in directed networks," arXiv:0709.4500v1 2007.
[11] B. Karrer, E. Levina, and M.E.J. Newman, "Robustness of community structure in networks," arXiv:0709.2108v1, 2007.
[12] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," Nature 435, pp. 814-818, 2005.
[13] M. Berry, Survey of Text Mining: Clustering, Classification, and Retrieval, Springer, 2003.
[14] F.R.K. Chung, Spectral Graph Theory, CBMS Regional Conf. Series in Mathematics, no. 92. American Mathematic Society, 1997.
[15] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) 888–905, 2000.
[16] Weiss Y. Segmentation using eigenvectors: a unifying view. Proceedings IEEE International Conference on Computer Vision p. 975-982, 1999.
[17] M.E.J. Newman, Modularity and community structure in networks, PNAS 0601602103, 2006.
[18] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Defining and identifying communities in networks, Proc. Natl. Acad. Sci. USA 101, 2658-2663, arXiv:cond-mat/0309488v2, 2004.