Modern Management based on Big Data II and Machine Learning and Intelligent Systems III
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA210280

The Larger the Better: Analysis of a Scalable Spectral Clustering Algorithm with Cosine Similarity

Guangliang CHEN [a,1]

[a] *San José State University, San José, California, United States*

Abstract. Chen (2018) proposed a scalable spectral clustering algorithm for cosine similarity to handle the task of clustering large data sets. It runs extremely fast, with a linear complexity in the size of the data, and achieves state of the art accuracy. This paper conducts perturbation analysis of the algorithm to understand the effect of discarding a perturbation term in an eigendecomposition step. Our results show that the accuracy of the approximation by the scalable algorithm depends on the connectivity of the clusters, their separation and sizes, and is especially accurate for large data sets.

Keywords. spectral clustering, cosine similarity, perturbation analysis

1. Introduction

Spectral clustering [1,2] was introduced at the beginning of the century as a very effective clustering approach. Given a set of objects $O = \{o_1, o_2, \ldots, o_n\}$ (such as images or documents) and a notion of similarity $s(\cdot,\cdot)$ (e.g., Gaussian or cosine similarity), the first step of spectral clustering, as in other graph-based applications [3,4], is to construct a weighted graph $\mathscr{G}$ from the given data using the similarity function s,

$$\mathscr{G} = \{V, E, \mathbf{W}\}, \qquad \mathbf{W} = (w_{ij}), \quad w_{ij} = s(o_i, o_j),$$

where $V = O$ is the vertex set, E the edge set (there is an edge between objects o_i, o_j if and only if $w_{ij} > 0$), and $\mathbf{W}$ the weight matrix. Next, spectral clustering computes the eigenvectors of a normalized version of $\mathbf{W}$ to obtain a low dimensional embedding of the data. Lastly, simple clustering algorithms like k-means are applied to the low dimensional coordinates to effectively cluster the given data. See Algorithm 1 for the Ng-Jordan-Weiss (NJW) version of spectral clustering [2] .

[1] Corresponding Author: Department of Mathematics and Statistics, MH 308, San José State University, One Washington Square, San José, CA 95192-0103, United States; E-mail: guangliang.chen@sjsu.edu.

Algorithm 1 Spectral Clustering (NJW)

Input: Graph $\mathscr{G} = \{V, E, \mathbf{W}\}$, number of clusters k

Output: A partition of the vertices in V into k clusters

1: Construct a diagonal degree matrix $\mathbf{D}$ with $\mathbf{D}_{ii} = \sum_j w_{ij}$, and use it to normalize $\mathbf{W}$ to obtain $\widetilde{\mathbf{W}} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$.

2: Find the top k eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ of $\widetilde{\mathbf{W}}$ (corresponding to the largest k eigenvalues) and stack them as columns into a matrix $\mathbf{V} = [\mathbf{v}_1 \ldots \mathbf{v}_k]$.

3: Group the row vectors of $\mathbf{V}$ into k clusters by using the k-means algorithm.

Spectral clustering is a nonlinear clustering method due to the eigenvectors embedding step. As a result, it can easily handle non-convex geometries and accurately separate non-intersecting shapes. Spectral clustering has been successfully used in many applications, such as document clustering and image segmentation. However, spectral clustering is very slow on large data sets because of its high computational complexity associated to the $n \times n$ matrix $\mathbf{W}$, which requires $O(n^2)$ memory and $O(n^3)$ time (for performing eigenvalue decomposition). Consequently, there has been considerable effort in the machine learning and data mining communities to develop fast, approximate spectral clustering algorithms that are scalable to large data [5,6,7,8,9,10,11,12,13,14,15,16,17]. Among those published methods, many use a "sampling plus extension" strategy by first working on a small number of landmark points selected from the given data and later extending the result to the full data set. As a result, the quality of the selected landmark points is crucial for the clustering accuracy and effective sampling can be very challenging in the setting of large, complex data sets.

2. A Review of the Scalable Spectral Clustering Algorithm

In [13] we tried to use the entire data set for direct clustering in the special setting of spectral clustering with cosine similarity:

$$\mathbf{W} = \mathbf{X}\mathbf{X}^T - \mathbf{I}, \tag{1}$$

where $\mathbf{X} = [\mathbf{x}_1 \ldots \mathbf{x}_n]^T \in \mathbb{R}^{n \times d}$ represents a data matrix consisting of n unit vectors[2] in $\mathbb{R}^d$ and $\mathbf{I}$ is the identity matrix. To speed up spectral clustering on large data, we exploited the product form of the weight matrix $\mathbf{W}$ for efficient implementation of spectral clustering.

Specifically, we assumed that the given data, despite its large size (n), has some sort of low-dimensional structure in one of the following ways:

(a) **The dimension d is also large but X is sparse**. This assumption is often true for document data sets that are represented as document-term frequency matrices under the bag-of-words model [18]. An example is the well-known 20 newsgroups data set[3] stored in matrix format: $\mathbf{X} \in \mathbb{R}^{n \times d}$ where $n = 18,774$ (documents), and $d = 61,188$ (terms), but the average row sparsity is only 129.7 (that is, on average, each document only contains 129.7 distinct words).

[2] The original data are vectors in $\mathbb{R}^d$. They have been normalized to have unit length in order to compute the cosine similarity.

[3] Available at http://qwone.com/~jason/20Newsgroups/.

(b) $\mathbf{d} \ll \mathbf{n}$ ($\mathbf{X}$ does not need to be sparse). This assumption is true for many image data sets, such as the MNIST handwritten digits[4] ($n = 70,000, d = 784$).

Note that for high dimensional non-sparse data, one can always use principal component analysis (PCA) to embed the given data into a few hundred dimensions (such that $d \ll n$), which often prove to be sufficient.

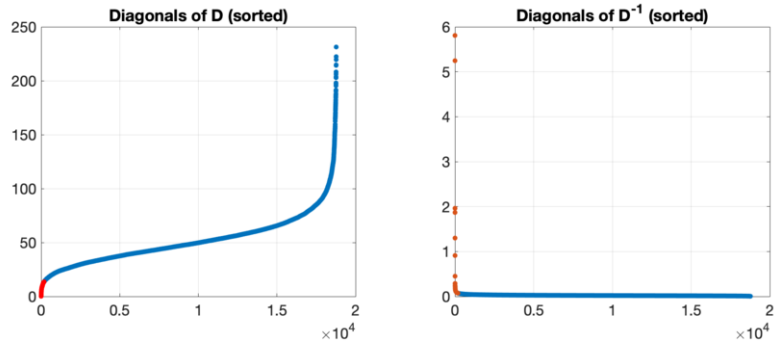

Figure 1. Sorted digonals of $\mathbf{D}$ and $\mathbf{D}^{-1}$ corresponding to the 20newsgroups data. The red part of each curve corresponds to the 1% of the data with the lowest degrees. The right plot shows that after that part is removed from the data, the remaining diagonals of $\mathbf{D}^{-1}$ are approximately constant (mean: 0.0229, standard deviation: 0.0102).

We then showed that in those scenarios, spectral clustering with cosine similarity may be performed directly through efficient operations on the data matrix $\mathbf{X}$ such as element-wise manipulation, matrix-vector multiplication, and low-rank singular value decomposition (SVD)[5], thus completely avoiding the $n \times n$ weight matrix $\mathbf{W}$:

(1) The diagonal degree matrix $\mathbf{D}$ is directly computed through matrix-vector multiplication

$$\mathbf{D} = \mathrm{diag}((\mathbf{X}\mathbf{X}^T - \mathbf{I})\mathbf{1}) = \mathrm{diag}(\mathbf{X}(\mathbf{X}^T\mathbf{1}) - \mathbf{1}), \tag{2}$$

where $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^n$ is a constant vector. The diagonals of $\mathbf{D}$ (degrees) are measures of the connectivity of the vertices in the graph. Afterwards, for some small $\alpha > 0$, a fraction α of the data in $\mathbf{X}$ with the lowest degrees is discarded as outliers such that for the remaining data (still denoted by $\mathbf{X}$), the diagonal of $\mathbf{D}^{-1}$ is approximately constant. See Figure 1 for a demonstration.

(2) For the correspondingly reduced weight matrix (still denoted by $\mathbf{W}$), write the normalized cosine similarity matrix $\widetilde{\mathbf{W}}$ as follows:

$$\widetilde{\mathbf{W}} = \mathbf{D}^{-1/2}\underbrace{(\mathbf{X}\mathbf{X}^T - \mathbf{I})}_{\mathbf{W}}\mathbf{D}^{-1/2} = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T - \mathbf{D}^{-1}, \quad \widetilde{\mathbf{X}} = \mathbf{D}^{-1/2}\mathbf{X}. \tag{3}$$

Disregard the $\mathbf{D}^{-1}$ term in (3) to use the left singular vectors of $\widetilde{\mathbf{X}}$ to approximate the corresponding eigenvectors of $\widetilde{\mathbf{W}}$ (note that $\widetilde{\mathbf{X}}$ is sparse or low dimensional, dependent on $\mathbf{X}$, and thus its SVD can be computed efficiently).

[4] Available at http://yann.lecun.com/exdb/mnist/.

[5] All of these operations have a linear complexity in n, the number of data points.

See Algorithm 2 for the full algorithm.

Algorithm 2 Scalable spectral clustering with cosine similarity

Input: Data matrix $\mathbf{X}$ (sparse or of moderate dimension, with L_2 normalized rows), #clusters k, fraction of data to be removed α

Output: Clusters $C_1,\ldots,C_k$ and a set of outliers C_0

Steps:

1: Compute the degree matrix $\mathbf{D}$ via (2) and label the bottom $(100\alpha)\%$ points as outliers (stored in the set C_0). Remove C_0 from the input data.

2: Calculate $\widetilde{\mathbf{X}} = \mathbf{D}^{-1/2}\mathbf{X}$ for the remaining data and find its top k left singular vectors $\tilde{\mathbf{u}}_1,\ldots,\tilde{\mathbf{u}}_k \in \mathbb{R}^n$ by rank-k SVD. Let $\widetilde{\mathbf{U}}_k = [\tilde{\mathbf{u}}_1,\ldots,\tilde{\mathbf{u}}_k] \in \mathbb{R}^{n\times k}$.

3: Normalize the rows of $\widetilde{\mathbf{U}}_k$ to have unit length and apply k-means to find k clusters $C_1,\ldots,C_k$.

We tested our scalable algorithm for clustering large text and image data sets and obtained comparable accuracy with the plain implementation but our algorithm runs much faster [13]. Recently, we have successfully extended the work to deal with general similarity functions (such as Gaussian) [15,17].

3. Analysis

In this section we conduct careful and rigorous analysis of the effect of the term $\mathbf{D}^{-1}$ in (3) on the eigenvectors of $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T$.

3.1. Insights

We start by making the following observations:

- The matrix $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T$ in (3) is symmetric and positive semidefinite, and $\mathbf{D}^{-1}$ can be viewed as a perturbation to it. Thus, the research conducted here is along the direction of perturbation analysis of the eigenspace of a positive semidefinite matrix.
- The matrix $\widetilde{\mathbf{W}}$ is similar to a row-stochastic matrix $\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}$:

$$\widetilde{\mathbf{W}} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2} = \mathbf{D}^{1/2}\mathbf{P}\mathbf{D}^{-1/2} \tag{4}$$

  Therefore, the largest eigenvalue of $\widetilde{\mathbf{W}} = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T - \mathbf{D}^{-1}$ is 1. The next $k-1$ largest eigenvalues of $\widetilde{\mathbf{W}}$ are expected to be close to 1 and meanwhile, there should be a significant drop at the $(k+1)$-th eigenvalue.
- If $\mathbf{D}$ has a constant diagonal, then $\widetilde{\mathbf{W}}$ have the same eigenvectors with $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T$ (but not the same eigenvalues), thus discarding $\mathbf{D}^{-1}$ won't change the eigenvectors in such a case.
- Adding a constant multiple of the identity matrix to $\widetilde{\mathbf{W}}$ does not change its eigenvectors (but only shifts its eigenvalues by β):

$$\widetilde{\mathbf{W}} + \beta\mathbf{I} = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T + (\beta\mathbf{I} - \mathbf{D}^{-1}) \tag{5}$$

We will select β such that $\beta\mathbf{I}-\mathbf{D}^{-1}$ is as small as possible (with respect to appropriate matrix norm) while being positive semidefinite.

- The underlying function of $\mathbf{D}^{-1}$ is $f(x)=1/x$ which flattens out quickly as x increases. This implies that $\mathbf{D}^{-1}$ is often close to being constant diagonal for large data sets (in which case, most diagonals of $\mathbf{D}$ are expected to be large). We will conduct analysis to estimate the magnitude of the diagonals of $\mathbf{D}$ in such settings.

3.2. Analysis

Let $\widetilde{\mathbf{U}}_k \in \mathbb{R}^{n\times k}$ be the matrix consisting of the top k eigenvectors of $\widetilde{\mathbf{W}}$, which is used by the exact Ng-Jordan-Weiss spectral clustering algorithm. Denote by $\widehat{\mathbf{U}}_k \in \mathbb{R}^{n\times k}$ the matrix consisting of the top k eigenvectors of $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T$ (which are also left singular vectors of $\widetilde{\mathbf{X}}$), which is used by the scalable spectral clustering algorithm as an approximation to $\widetilde{\mathbf{U}}_k$. Our goal here is to relate the Grassmann distance between $\widehat{\mathbf{U}}_k$ and $\widetilde{\mathbf{U}}_k$, i.e., $\|\widehat{\mathbf{U}}_k\widehat{\mathbf{U}}_k^T - \widetilde{\mathbf{U}}_k\widetilde{\mathbf{U}}_k^T\|_F$, to the perturbation term $\mathbf{D}^{-1}$.

Using [19, Theorem A.1] we can prove the following result.

Theorem 3.1. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ be the eigenvalues of $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T$, and define $\delta_k = \lambda_k - \lambda_{k+1} > 0$. Then

$$\|\widetilde{\mathbf{U}}_k\widetilde{\mathbf{U}}_k^T - \widehat{\mathbf{U}}_k\widehat{\mathbf{U}}_k^T\|_F \leq \frac{2\sqrt{n}}{\delta_k}\|\mathbf{D}^{-1}\|_2. \tag{6}$$

Proof. Write the degree matrix as

$$\mathbf{D} = \mathrm{diag}(d_1,\ldots,d_n), \quad 0 < d_1 \leq \cdots \leq d_n$$

Let $\beta = {d_1}^{-1}$, and define

$$\widetilde{\mathbf{W}}^{(\beta)} = \widetilde{\mathbf{W}} + \beta\mathbf{I} = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T + (\beta\mathbf{I}-\mathbf{D}^{-1})$$

Clearly, both $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T$ and $\beta\mathbf{I}-\mathbf{D}^{-1}$ are symmetric and positive semidefinite. It follows that $\widetilde{\mathbf{W}}^{(\beta)}$ is also positive semidefinite. Additionally, the eigenvectors of $\widetilde{\mathbf{W}}^{(\beta)}$ are the same with those of $\widetilde{\mathbf{W}}$ (so the top k eigenvectors of $\widetilde{\mathbf{W}}^{(\beta)}$ are still $\widetilde{\mathbf{U}}_k$). Thus, $\beta\mathbf{I}-\mathbf{D}^{-1}$ can be viewed as a positive semidefinite perturbation matrix to the positive semidefinite matrix $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T$.

Using the perturbation result in [19, Theorem A.1], we obtain that

$$\|\widetilde{\mathbf{U}}_k\widetilde{\mathbf{U}}_k^T - \widehat{\mathbf{U}}_k\widehat{\mathbf{U}}_k^T\|_F \leq \frac{2}{\delta_k}\|\beta\mathbf{I}-\mathbf{D}^{-1}\|_F$$

provided that $\|\beta\mathbf{I}-\mathbf{D}^{-1}\|_F \leq \delta_k/4$.

Since

$$\|\beta\mathbf{I}-\mathbf{D}^{-1}\|_F^2 = \sum_{i=1}^{n}\left(\frac{1}{d_1}-\frac{1}{d_i}\right)^2 = \frac{1}{d_1^2}\sum_{i=1}^{n}\left(\frac{d_i-d_1}{d_i}\right)^2 \leq \frac{1}{d_1^2}\sum_{i=1}^{n}1 = \frac{n}{d_1^2}$$

we have

$$\|\widetilde{\mathbf{U}}_k\widetilde{\mathbf{U}}_k^T - \widehat{\mathbf{U}}_k\widehat{\mathbf{U}}_k^T\|_F \le \frac{2}{\delta_k}\cdot\sqrt{\frac{n}{d_1^2}} = \frac{2\sqrt{n}}{\delta_k}\|\mathbf{D}^{-1}\|_2.$$

This completes the proof. □

Next, we estimate the magnitude of the diagonals of **D** in the setting of image or document data (where the data matrix has nonnegative entries). For this purpose, we need to assume that the data is a sample from a mixture distribution in the first orthant of $\mathbb{R}^d$. Specifically, we suppose that each cluster C_j is a random sample of n_j points from a cone that is within an angle of θ around a unit vector $\mathbf{t}_j$, where $n_j \ge \gamma n$ for some fixed constant $\gamma > 0$. See Figure 2 for an illustration. The data used in Algorithm 2, $\mathbf{X} \in \mathbb{R}^{n\times d}$, represent the normalized version of the sample.

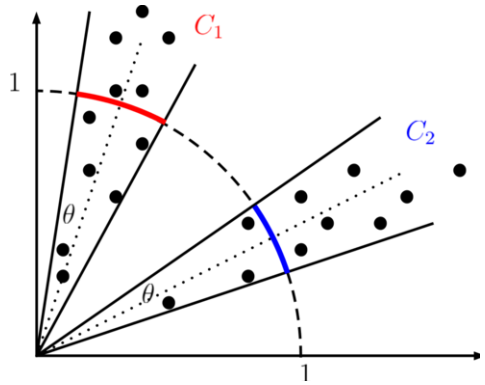

Figure 2. Mixture of cones model (when $k = 2$). Each cluster is a collection of samples from a cone concentrated around its axis within angle θ. The original observations are projected onto the unit sphere (represented by the dashed curve) to have unit length for computing the cosine similarity used by Algorithm 2.

We can prove the following result on the magnitude of the degrees of the points.

Theorem 3.2. Under the above assumptions,

$$d_i \ge \gamma\cos^2\theta - 1, \quad \text{for all } i = 1,\ldots,n. \tag{7}$$

Proof. Consider the data point $\mathbf{x}_i$ and suppose that it comes from C_j.

By (2), we have

$$d_i = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{1}) - 1 = \mathbf{x}_i^T\left(\sum_{\ell=1}^{n}\mathbf{x}_\ell\right) - 1$$

Since all the data points are in the first quadrant,

$$d_i \ge \mathbf{x}_i^T\left(\sum_{\mathbf{x}_\ell\in C_j}\mathbf{x}_\ell\right) - 1 = \mathbf{x}_i^T(n_j\mathbf{m}_j) - 1$$

where

$$\mathbf{m}_j = \frac{1}{n_j} \sum_{\mathbf{x}_\ell \in C_j} \mathbf{x}_\ell$$

is the centroid of C_j.

Using elementary geometry, we have

$$\|\mathbf{m}_j\| \geq \cos\theta.$$

It follows that

$$\begin{aligned} d_i &\geq n_j \left(\mathbf{x}_i^T \mathbf{m}_j\right) - 1 \\ &\geq n_j \cdot (1 \cdot \|\mathbf{m}_j\| \cdot \cos\theta) - 1 \\ &\geq (\gamma n) \cdot \cos\theta \cdot \cos\theta - 1 \\ &= \gamma n \cos^2\theta - 1 \end{aligned}$$

This completes the proof. □

Combining the above two theorems immediately give the following result.

Corollary 3.1. Under the same assumptions as in Theorems 3.1 and 3.2,

$$\|\widetilde{\mathbf{U}}_k \widetilde{\mathbf{U}}_k^T - \widehat{\mathbf{U}}_k \widehat{\mathbf{U}}_k^T\|_F \leq \frac{2\sqrt{n}}{\delta_k (\gamma n \cos^2\theta - 1)}. \tag{8}$$

4. Discussions

Theorem 3.1 shows that the closeness of $\widehat{\mathbf{U}}_k$ to $\widetilde{\mathbf{U}}_k$ is bounded by the kth eigengap δ_k (inversely) and the spectral norm of the matrix $\mathbf{D}$. The size of eigengap δ_k corresponds to the separation between the different clusters. If the k clusters are all well separated, then δ_k will be considerably bigger than zero.

Theorem 3.2 shows that the magnitude of the degrees of the data is $O(n)$. For large data sets, the degrees of the points are also large, on the same order with the size of the data. As a result, the diagonals of $\mathbf{D}^{-1}$ will be very small, and thus the perturbation to $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T$ will be small.

Theorem 3.1 combines Theorems 3.1 and 3.2 to relate the Grassmann distance of $\widehat{\mathbf{U}}_k$ to $\widetilde{\mathbf{U}}_k$ to the following quantities:

- n: size of the data set.
- γ: fraction of the smallest cluster in the data.
- θ: tightness of the clusters. The smaller θ (and correspondingly the larger $\cos^2\theta$), the more connected within each cluster.
- δ_k: eigengap which measures separation among the clusters.

Overall, the Grassmann distance of $\widehat{\mathbf{U}}_k$ to $\widetilde{\mathbf{U}}_k$ is $O(n^{-1/2})$.

5. Conclusion

We showed through a matrix perturbation analysis that, for large data sets that have well connected clusters and sufficient separation between them, the scalable spectral clustering algorithm (Algorithm 2) provides a close approximation to the plain algorithm (Algorithm 1). The larger the data set, the better the approximation!

References

[1] Shi J, Malik J. Normalized cuts and image segmentation. IEEE Trans Pattern Anal Mach Intell. 2000;22(8):888–905.

[2] Ng A, Jordan M, Weiss Y. On spectral clustering: analysis and an algorithm. In: Advances in Neural Information Processing Systems 14; 2001. p. 849–856.

[3] Pietrabissa A, Celsi LR, Cimorelli F, Suraci V, Priscoli FD, Giorgio AD, et al. Lyapunov-based design of a distributed wardrop load-balancing algorithm with application to software-defined networking. IEEE Transactions on Control Systems Technology. 2019;27(5):1924–1936.

[4] Celsi LR, Giorgio AD, Gambuti R, Tortorelli A, Priscoli FD. On the many-to-many carpooling problem in the context of multi-modal trip planning. In: Proceedings of the 25th Mediterranean Conference on Control and Automation (MED); 2017. p. 303–309.

[5] Fowlkes C, Belongie S, Chung F, Malik J. Spectral grouping using the Nyström method. IEEE Trans Pattern Analysis and Machine Intelligence. 2004;26(2):214–225.

[6] Yan D, Huang L, Jordan M. Fast approximate spectral clustering. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2009. p. 907–916.

[7] Wang L, Leckie C, Ramamohanarao K, Bezdek J. Approximate spectral clustering. vol. 5476 of Advances in Knowledge Discovery and Data Mining. PAKDD 2009, Lecture Notes in Computer Science. Berlin, Heidelberg: Springer; 2009.

[8] Wang L, Leckie C, Kotagiri R, Bezdek J. Approximate pairwise clustering for large data sets via sampling plus extension. Pattern Recognition. 2011;44:222–235.

[9] Tasdemir K. Vector quantization based approximate spectral clustering of large datasets. Pattern Recognition. 2012;45(8):3034–3044.

[10] Choromanska A, Jebara T, Kim H, Mohan M, Monteleoni C. Fast spectral clustering via the Nyström method. vol. 8139 of Algorithmic Learning Theory. ALT 2013. Lecture Notes in Computer Science. Jain S, Munos R, Stephan F, Zeugmann T, editors. Berlin, Heidelberg: Springer; 2013.

[11] Cai D, Chen X. Large scale spectral clustering via landmark-based sparse representation. IEEE Transactions on Cybernetics. 2015;45(8):1669–1680.

[12] Moazzen Y, Tasdemir K. Sampling based approximate spectral clustering ensemble for partitioning data sets. In: Proceedings of the 23rd International Conference on Pattern Recognition; 2016. .

[13] Chen G. Scalable spectral clustering with cosine similarity. In: Proceedings of the 24th International Conference on Pattern Recognition (ICPR), Beijing, China; 2018. .

[14] Pham K, Chen G. Large-scale spectral clustering using diffusion coordinates on landmark-based bipartite graphs. In: Proceedings of the 12th Workshop on Graph-based Natural Language Processing (TextGraphs-12). New Orleans, Louisiana: Association for Computational Linguistics; 2018. p. 28–37.

[15] Chen G. A scalable spectral clustering algorithm based on landmark-embedding and cosine similarity. In: Bai X., Hancock E., Ho T., Wilson R., Biggio B., Robles-Kelly A. (eds) Structural, Syntactic, and Statistical Pattern Recognition. S+SSPR 2018. Lecture Notes in Computer Science. vol. 11004. Cham: Springer; 2018. .

[16] Chen G. Matlab implementation details of a scalable spectral clustering algorithm with cosine similarity. In: Proceedings of the 2nd Workshop on Reproducible Research in Pattern Recognition. Cham: Springer; 2018. .

[17] Chen G. A general framework for scalable spectral clustering based on document models. Pattern Recognition Letters. 2019;125:488–493.

[18] Aggarwal C, Zhai C. In: A survey of text clustering algorithms. Boston, MA: Springer US; 2012. p. 77–128.

[19] Chen G, Lerman G. Foundations of a multi-way spectral clustering framework for hybrid linear modeling. Found Comput Math. 2009;DOI 10.1007/s10208-009-9043-7.