Document Clustering Based on Spectral Clustering and Non-negative Matrix Factorization*

Lei Bao[1, 2], Sheng Tang[2], Jintao Li[2], Yongdong Zhang[2], and Wei-ping Ye[1]

[1] Beijing Normal University, Beijing 100875, China
[2] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China
{baolei, ts, jtli, zhyd}@ict.ac.cn, bnuywp@yahoo.com

Abstract. In this paper, we propose a novel non-negative matrix factorization (NMF) to the affinity matrix for document clustering, which enforces non-negativity and orthogonality constraints simultaneously. With the help of orthogonality constraints, this NMF provides a solution to spectral clustering, which inherits the advantages of spectral clustering and presents a much more reasonable clustering interpretation than the previous NMF-based clustering methods. Furthermore, with the help of non-negativity constraints, the proposed method is also superior to traditional eigenvector-based spectral clustering, as it can inherit the benefits of NMF-based methods that the non-negative solution is institutive, from which the final clusters could be directly derived. As a result, the proposed method combines the advantages of spectral clustering and the NMF-based methods together, and hence outperforms both of them, which is demonstrated by experimental results on TDT2 and Reuters-21578 corpus.

Keywords: Document Clustering, Spectral Clustering, Non-negative Matrix Factorization.

1 Introduction

Document clustering is to divide a collection of documents into different clusters based on similarities of content. It has been widely used as a fundamental and effective tool for efficient organization, summarization, navigation and retrieval of large amount of documents, and attracted a lot of attention in recent years [1,2,3].

Spectral clustering, which doesn't make assumption on data distributions, is one of the most popular modern clustering algorithms. It represents document corpus as an undirected graph, and the task of clustering is transformed to find the best cuts of graph optimizing certain criterion functions, such as the ratio cut [4], average association [5], normalized cut [5] and min-max cut [6]. It can be proved that the top eigenvectors matrix of the graph affinity matrix, or a matrix derived from it, is the

* This research was supported by National Basic Research Program of China (973 Program, 2007CB311100), National High Technology and Research Development Program of China (863 Program, 2007AA01Z416), Beijing New Star Project on Science & Technology (2007B071).

N.T. Nguyen et al. (Eds.): IEA/AIE 2008, LNAI 5027, pp. 149–158, 2008.

solution to these optimization problems under relaxed conditions [7]. The rigorous mathematical derivation ensures that the eigenvectors matrix encodes the cluster information, which provides a reasonable clustering interpretation for the eigenvectors matrix. However, the real-valued eigenvectors matrix is not intuitive and doesn't directly correspond to individual clusters [2]. Consequently, traditional clustering methods in the real-valued matrix are necessary to get the final clusters.

Recently, document clustering based on non-negative matrix factorization (NMF) [8,9] has become popular for the intuitiveness of its non-negative solution. Xu et al. [2] proposed to represent each document as an additive combination of base topics, which are learned by NMF. Ding et al. [11] proposed an extension of NMF to the affinity matrix for clustering. The non-negative solutions provided by these NMF-based methods are much more intuitive than the real-valued ones, as the final clusters can be directly derived from these solutions. However, these methods lack a rigorous theoretical derivation to provide a reasonable clustering interpretation for their solutions and sometimes lead to unsatisfactory performances.

Based on the previous works, we propose a novel NMF to the affinity matrix for document clustering with non-negativity and orthogonality constraints simultaneously. With the help of additional orthogonality constraints, this NMF provides a non-negative solution to spectral clustering, which not only inherits the intuitiveness of the NMF-based methods, but also inherits the reasonable clustering interpretation of spectral clustering. Consequently, our method could combine the advantages of spectral clustering and NMF-based methods together, and outperform both of them.

2 A Brief Review

This section reviews spectral clustering and the previous NMF-based methods.

2.1 Spectral Clustering

Spectral clustering represents a document corpus $D=\{doc_1, doc_2, \cdots, doc_N\}$ as an undirected graph $G(V,E,W)$, where V,E,W denote the vertex set, the edge set, and the graph affinity matrix, respectively. Each vertex $v_i \in V$ represents a document doc_i, and each $edge(i,j) \in E$ is assigned an affinity score w_{ij} forming matrix W which reflects the similarity between doc_i and doc_j. The clustering task is consequently transformed to find the best cuts of the graph that optimize certain criterion functions. Here, we discuss two representative criterion functions: Average Association (AA) [5] and Normalized Cut (NC) [5]. The two criterion functions and their corresponding relaxed eigen-problems are summarized in Table 1, where $A_1, \cdots, A_K$ are the K disjoint sets (clusters) of V. $\overline{A}$ denotes the complement of $A \subset V$ and $|A|$ denotes the number of vertices in A. $cut(A,B)$ is the similarity between A and B calculated by $cut(A,B)=\sum_{i\in A, j\in B} w_{ij}$. $vol(A)$ is the weight of A calculated by $vol(A)=\sum_{i\in A} d_i$ where $d_i=\sum_{j=1}^{N} w_{ij}$. D is a diagonal matrix with $d_1, \cdots, d_N$ on the diagonal.

Table 1. Corresponding eigen-problems of AA and NC

Method	Average Association	Normalized Cut
Criterion Function	$\max_{A_1,\cdots,A_K} \sum_{i=1}^{K} \frac{cut(A_i, A_i)}{\lvert A_i \rvert}$	$\min_{A_1,\cdots,A_K} \sum_{i=1}^{K} \frac{cut(A_i, \overline{A_i})}{cut(A_i, V)}$
Indicator Matrix	$H \in \mathbb{R}^{N\times K}$ $h_{i,j} = \begin{cases} 1/\sqrt{\lvert A_j \rvert} & \text{if} \quad i \in A_j \\ 0 & \text{otherwise} \end{cases}$	$H \in \mathbb{R}^{N\times K}$ $h_{i,j} = \begin{cases} 1/\sqrt{vol(A_j)} & \text{if} \quad i \in A_j \\ 0 & \text{otherwise} \end{cases}$
Rewritten Problem	$\max_{A_1,\cdots,A_K} Tr(H^T W H)$, s.t. $H^T H = I$ H defined as Indicator Matrix	$\max_{A_1,\cdots,A_K} Tr(H^T W H)$, s.t. $H^T D H = I$ H defined as Indicator Matrix
Relaxed Problem	$\max_{H\in\mathbb{R}^{N\times K}} Tr(H^T W H)$ *s.t.* $H^T H = I$	$\max_{\widetilde{H}\in\mathbb{R}^{N\times K}} Tr(\widetilde{H}^T \widetilde{W} \widetilde{H})$ *s.t.* $\widetilde{H}^T \widetilde{H} = I$, where $\widetilde{H} = D^{1/2} H$, $\widetilde{W} = D^{-1/2} W D^{-1/2}$
Real-valued Solution	H : the eigenvectors of the K largest eigenvalues of W as columns	$\widetilde{H}$: the eigenvectors of the K largest eigenvalues of $\widetilde{W}$ as columns

From Table 1, we find that the relaxed problems for AA and NC are similar, and the only difference is that AA deals with W, while NC deals with $\widetilde{W}$. Here, we briefly explain some details of AA. The rewritten problem for AA is an NP-hard problem. However, when we relax it by allowing the elements of H to take arbitrary real values, the relaxed problem can be easily solved by applying Rayleigh Quotient Theorem [12]. The rigorous mathematical derivation ensure that as an approximation to indicator matrix, matrix H obtained by eigen-decomposition is interpretable and encodes the cluster information for the given document corpus. However, the solution H is not intuitive and doesn't directly indicate the cluster membership because of its real values. We have to use k-means on the rows of H to get the final cluster set.

2.2 Document Clustering by Non-negative Matrix Factorization

Given a document corpus with N documents and K clusters, [2] represents each document as a non-negative linear combination of K clusters centers which are also constrained to be non-negative. Translating this statement into mathematics, we have:

$$X \approx UV^T, \quad s.t.\ U \geq 0, V \geq 0 \tag{1}$$

where $X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]$ is the term-document matrix and $\mathbf{x}_i$ represents the term-frequency vector of doc_i, $U = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_K]$ and $\mathbf{u}_k$ represents the k'th cluster center, and $V^T = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_N]$ and $\mathbf{v}_i$ is the linear coefficients of doc_i. Intuitively, by the NMF to X, we can assume the coefficient matrix V encodes some cluster information. Based on the assumption, we can assign doc_i to cluster k when $k = \arg\max_j v_{ij}$.

In [11], Chris Ding proposed an extension of NMF to the graph affinity matrix:

$$W \approx HH^T, \quad s.t.\ H \geq 0 \tag{2}$$

Where $W \in \mathbb{R}_+^{N\times N}$, $H \in \mathbb{R}_+^{N\times K}$ and we denote this NMF as NMF-HHt. The matrix H could be considered as an approximate solution to spectral clustering. Hence the final clusters are obtained by assigning doc_i to cluster k when $k = \arg\max_j h_{ij}$. However, from the derivation process in [11], we notice that the equivalence between NMF-HHt and spectral clustering is based on strict orthogonality constraints $H^T H = I$, and the constraints are hard to retain by NMF-HHt. As a result, the equivalence is hardly convincing and the clustering interpretation for H is unreasonable.

The NMF-based methods mentioned above benefit a lot from the intuitiveness of the non-negative solution. However, these methods lack a rigorous theoretical derivation to provide a reasonable clustering interpretation for their solutions, and sometimes lead to unsatisfactory performances.

3 The Proposed Method

The previous work motivates us to find a new NMF with a reasonable clustering interpretation, which is not only intuitive but also interpretable. Here, we propose a novel NMF to the affinity matrix with non-negativity and orthogonality constraints simultaneously. With the additional orthogonality constraints, this NMF is similar to eigen-decomposition but adds non-negativity constraints. Intuitively, the new NMF can provide a non-negative solution to spectral clustering. Given a corpus with N documents and K clusters ($K < N$), the factorization is expressed as follows:

$$W \approx PSP^T, \ \ s.t.\ P \geq 0, S \geq 0, P^T P = I_K \tag{3}$$

where $W \in \mathbb{R}_+^{N\times N}$, $P \in \mathbb{R}_+^{N\times K}$, $S \in \mathbb{R}_+^{K\times K}$ is a diagonal matrix, and this NMF is denoted as NMF-PSPt in short. NMF-PSPt is transformed to the following optimization problem:

$$\min_{P\geq 0, S\geq 0} J_{NMF} = \min_{P\geq 0, S\geq 0} \left(\alpha \left\| W - PSP^T \right\|^2 + \beta \left\| P^T P - I_K \right\|^2 \right) \tag{4}$$

where α and β are positive constants. Since $\left\| W - PSP^T \right\|^2$ is the squared sum of $N \times N$ elements and $\left\| P^T P - I_K \right\|^2$ is the squared sum of $K \times K$ elements, we set $\alpha = \eta / (N \times N)$ and $\beta = (1-\eta)/(K \times K)$ to keep evenness, where η ($0 < \eta \leq 1$) is a parameter to control the weight of constraint $P^T P = I_K$. The relationship between NMF-PSPt and spectral clustering is shown in the following theorem.

Theorem 1. The matrix P obtained by NMF-PSPt is an approximate solution to the relaxed optimizing problems for spectral clustering $\max_{H\in\mathbb{R}^{N\times K}} Tr(H^T W H)\ s.t.\ H^T H = I$.

Proof. Firstly, according to Eq.(3), W can be written as: $W = PSP^T + \mathbf{\theta}$, where $\mathbf{\theta}$ is a N-by-N matrix and its elements are negligible compared with PSP^T, which is ensured by the objective function J_{NMF} in Eq.(4). Since $Tr\left(H^T W H\right) = Tr\left(H^T \left(PSP^T\right) H\right) + Tr(H^T \mathbf{\theta} H)$, we can relax the optimizing problems for spectral clustering as follows:

$$\max_{H\in\mathbb{R}^{N\times K}} Tr(H^T\left(PSP^T\right)H)\ s.t.\ H^TH=I \tag{5}$$

Then, since $P^TP=I_K$, PSP^T can be written as: $PSP^T=\sum_{i=1}^{K}s_ip_ip_i^T$, where $s_1,\cdots,s_K$ are diagonal elements of S, $p_1,\cdots,p_K$ are columns of P and satisfy: $p_i^T\cdot p_i=1$, $p_i^T\cdot p_j=1$. This means $p_1,\cdots,p_K$ are the K eigenvectors corresponding to the K largest eigenvalues $s_1,\cdots,s_K$ of PSP^T. Finally, based on Rayleigh Quotient Theorem, we can prove that P is a solution to Eq.(5) and an approximate solution to spectral clustering. □

With Theorem 1, the solutions to spectral clustering can be equivalently carried out by NMF-PSPt, which ensures that the proposed method inherits the reasonable clustering interpretation of spectral clustering and is more interpretable and reliable than previous NMF-based methods. Furthermore, as the intuitiveness of non-negative solution, we can directly deduce the final clusters from P without an additional clustering operation, which is necessary for eigenvector-based spectral clustering. These mean NMF-PSPt combines the interpretation of spectral clustering and the intuitiveness of NMF-based methods together, overcomes the disadvantages of them.

3.1 Algorithm for Computing NMF-PSPt

The optimization for J_{NMF} can be completed by alternatively updating P and S until convergence. With fixed value of S, J_{NMF} becomes a quadratic form of $P: J_{NMF}(P)$. With fixed value of P, we get the quadratic form of $S: J_{NMF}(S)$. By referring to the algorithm derivation for symmetric convex coding [13], we deduced the following updating rules for P and S.

Theorem 2. The objective function $J_{NMF}(P)$ is decreasing under the updating rule,

$$P_{ik}=\tilde{P}_{ik}\left(\frac{\alpha\left[W\tilde{P}S\right]_{ik}+\beta\tilde{P}_{ik}}{\alpha\left[\tilde{P}S\tilde{P}^T\tilde{P}S\right]_{ik}+\beta\left[\tilde{P}^T\tilde{P}\right]_{ik}}\right)^{\frac{1}{4}} \tag{6}$$

Proof. Firstly, with the help of Jensen's inequality, convexity of the quadratic function and inequalities: $x^2+y^2\geq 2xy$, $x\geq 1+\log x$, we can prove that the following equation is the auxiliary function (defined in [9]) for $J_{NMF}(P)$.

$$\begin{aligned}G\left(P,\tilde{P}\right)=&\alpha\left(\sum_{ij}\left(W_{ij}^2-2W_{ij}\left[\tilde{P}S\tilde{P}^T\right]_{ij}\right)+4\sum_{ip}\tilde{P}_{ip}\log\frac{P_{ip}}{\tilde{P}_{ip}}\left[W\tilde{P}S\right]_{ip}+\sum_{ip}\frac{P_{ip}^4}{\tilde{P}_{ip}^3}\left[\tilde{P}S\tilde{P}^T\tilde{P}S\right]_{ip}\right)\\&+\beta\left(\sum_{pq}\left(I_{pq}^2-2I_{pq}\left[\tilde{P}^T\tilde{P}\right]_{pq}\right)-4\sum_{ip}\tilde{P}_{ip}\log\frac{P_{ip}}{\tilde{P}_{ip}}\left[\tilde{P}I_K\right]_{ip}+\sum_{ip}\frac{P_{ip}^4}{\tilde{P}_{ip}^3}\left[\tilde{P}\tilde{P}^T\tilde{P}\right]_{ip}\right)\end{aligned} \tag{7}$$

Then, taking the derivative of $G(P,P^t)$ w.r.t P_{ip}, and solving $\partial G\left(P,\tilde{P}\right)\big/\partial P_{ip}=0$, we get that Eq.(6). Based on the Lemma of auxiliary function in [9], we can prove $J_{NMF}(P)$ is decreasing under the updating rule in Eq.(6). □

Theorem 3. The objective function $J_{NMF}(S)$ is decreasing under the updating rule,

$$S_{kk} = \tilde{S}_{kk} \frac{\left[P^T W P\right]_{kk}}{\left[P^T P \tilde{S} P^T P\right]_{kk}} \tag{8}$$

Due to the space limit, we omit the proof of Theorem 3, which also can be analogously completed by introducing auxiliary function of $J_{NMF}(S)$.

Finally, we summarize the algorithm for computing NMF-PSPt as follows:

Step 1. Given an N-by-N graph affinity matrix W and a positive integer K, initialize N-by-K non-negative matrix P and K-by-K non-negative diagonal matrix.
Step 2. Update P and S by Eq.(6) and Eq.(8).
Step 3. Repeat Step 2 until convergence.

3.2 NMF-PSPt vs. Spectral Clustering and NMF-HHt

In this subsection, we use a simple and small document corpus to illustrate the document distributions derived by three different methods: the eigenvector-based AA method and two NMF-based methods based on NMF-HHt and NMF-PSPt. All the methods are focused on original affinity matrix W, and provide approximate solution to Average Association function. The small corpus is constructed by three topics from TDT2 corpus, which consist of 26, 15, and 43 documents respectively. Fig.1(a) shows matrix W, where each element is the cosine similarity between documents, and (b), (c) and (d) illustrate the data distributions derived by the three methods, where data points belonging to the same cluster are depicted by the same symbol. E_1, E_2, E_3 are the top three eigenvectors of W, H_1, H_2, H_3 are the column vectors of H obtained by NMF-HHt, and P_1, P_2, P_3 are the column vectors of P obtained by NMF-PSPt.

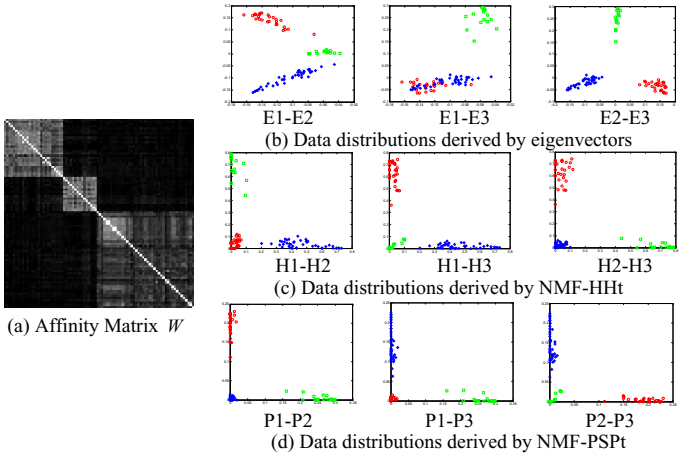

Fig. 1. Data distributions derived by eigenvectors, NMF-HHt and NMF-PSPt

From Fig. 1 we can observe that:

1. The small corpus can be easily separated as shown in Fig. 1(a). The within-cluster similarity is large and the between-cluster similarity is small. In fact, the three clusters are well separated by all the three methods as shown in Fig. 1(b), (c) and (d).
2. In Fig. 1(c) and (d), each data point is non-negative in all three directions, and the data points belonging to the same cluster spread along the same axis. However, in Fig. 1(b), each data point may take negative values in some directions, and the axes don't directly correspond to the clusters. As also discussed in [2], it shows the intuitiveness of non-negative solution.
3. Comparing Fig. 1(c) and (d), we find that the data points in Fig. 1(d) are much easier to be separated, as the data points of same cluster in Fig. 1(d) are more compact in same axis than those in Fig. 1(c). This fact verifies that, with the help of orthogonality constraints, the solution derived by NMF-PSPt is indeed more reliable and interpretable than that derived by NMF-HHt as mentioned in previous analysis.

4 Experimental Results

This section provides experimental evidences to show the effectiveness of our proposed algorithm in comparison with eigenvector-based spectral clustering algorithms [5] and previous NMF-based methods [2, 11].

4.1 Data Corpora

The standard corpora used in our experiments are TDT2[1] and Reuter-21578[2], which are benchmarks for document clustering. The TDT2 consists of English and Chinese documents from 9 news sources during the first half year of 1998 in 96 categories. The Reuters has 21,578 English documents in 135 categories. In our experiments, we used the English documents with unique category label, and excluded categories with less than 10 and more than 1,000 documents, leaving 5,333 documents in 52 categories for TDT2 and 2,710 documents in 40 categories for Reuters.

4.2 Performances Evaluations and Comparisons

As shown in Table 1, the essential difference between AA and NC is that NC applies the weights to W while AA does not. Therefore, we separate the two eigenvector-based methods, previous NMF-based methods and our proposed methods into two groups: one is non-normalized; the other is normalized, as showed in Table 2 and 3.

Non-normalized methods contain AA, NMF, AA-HHt and AA-PSPt. AA is eigenvector-based methods for Average Association function. NMF is the method proposed by [2] based on $X \approx UV^T$. AA-HHt and AA-PSPt are based on $W \approx HH^T$ and $W \approx PSP^T$, where w_{ij} is the cosine similarity between doc_i and doc_j.

Normalized methods contain NC, NMF-NCW, NC-HHt and NC-PSPt. NC is eigenvector-based methods for Normalized Cut function. NMF-NCW is the

[1] TDT2 corpus is at http://projects.ldc.upenn.edu/TDT2/
[2] Reuters-21578 corpus at http://www.daviddlewis.com/resources/testcollections/reuters21578

normalized NMF based on $XD^{-1/2} \approx UV^T$ [2]. NC-HHt and NC-PSPt are respectively based on $\widetilde{W} \approx HH^T$ and $\widetilde{W} \approx PSP^T$, where $\widetilde{W} = D^{-1/2}WD^{-1/2}$.

For performance measures, we selected the two metrics used in [2]: the accuracy (AC) and the normalized mutual information (MI) between the resulting clusters set and the ground truth set. The values of AC and MI are both within the range [0, 1], and the larger value represents the better performance.

As the evaluations in [2], we conducted our experiments for the cluster numbers ranging from 2 to 10. For each cluster number k, 20 tests were conducted on different randomly chosen clusters, and the final performance scores were the average of the 20 tests. For AA and NC, we applied k-means 10 times with different initial points and recorded the best result in terms of the objective function of k-means. Since NMF-based algorithms are affected by initialization, 10 trials were performed with different initial values and the trial with minimal criterion function value was chosen. In addition, the parameter η for AA-PSPt and NC-PSPt was set to 0.8. Table 2 and 3 respectively show the evaluation results of eight methods in two groups on TDT2 and Reuters. For each k, the best score in each group is shown in bold font.

Table 2. Performance comparisons on TDT2 corpus

K	Accuracy							
	Non-normalized algorithms				Normalized algorithms			
	AA	NMF	AA-HHt	AA-PSPt	NC	NMF-NCW	NC-HHt	NC-PSPt
2	0.8595	0.8465	0.8429	**0.9996**	**0.9987**	**0.9987**	**0.9987**	**0.9987**
3	0.8597	0.8404	0.8515	**0.9951**	0.9687	0.9413	0.9710	**0.9997**
4	0.7141	0.6884	0.6972	**0.9778**	0.9761	0.9532	0.9514	**0.9955**
5	0.7483	0.7280	0.7665	**0.9666**	0.9367	0.9503	0.9535	**0.9908**
6	0.7327	0.7279	0.7175	**0.9510**	0.8671	0.8958	0.8985	**0.9892**
7	0.7797	0.7588	0.7700	**0.9372**	0.9037	0.9280	0.9338	**0.9611**
8	0.7573	0.7413	0.7753	**0.9401**	0.8819	0.9253	0.9192	**0.9882**
9	0.7535	0.7546	0.7690	**0.9237**	0.8344	0.8696	0.8728	**0.9624**
10	0.7228	0.7426	0.7793	**0.9004**	0.9011	0.9310	0.9321	**0.9585**
avg.	0.7697	0.7587	0.7744	**0.9546**	0.9187	0.9326	0.9368	**0.9827**
K	Mutual Information							
	Non-normalized algorithms				Normalized algorithms			
	AA	NMF	AA-HHt	AA-PSPt	NC	NMF-NCW	NC-HHt	NC-PSPt
2	0.6953	0.6643	0.6640	**0.9961**	**0.9891**	**0.9882**	**0.9882**	**0.9882**
3	0.7598	0.7314	0.7545	**0.9497**	0.9426	0.9076	0.9430	**0.9968**
4	0.6377	0.6119	0.6171	**0.8790**	0.9486	0.9056	0.9035	**0.9676**
5	0.7163	0.6982	0.7179	**0.8986**	0.9240	0.9373	0.9352	**0.9776**
6	0.7254	0.7132	0.7084	**0.8871**	0.8598	0.8852	0.8840	**0.9749**
7	0.7952	0.7796	0.7758	**0.8906**	0.8959	0.9281	0.9272	**0.9473**
8	0.7973	0.7831	0.8003	**0.8970**	0.8911	0.9253	0.9168	**0.9756**
9	0.7954	0.8070	0.8114	**0.8842**	0.8538	0.8808	0.8859	**0.9473**
10	0.7876	0.7955	0.8083	**0.8689**	0.8834	0.9327	0.9308	**0.9699**
avg.	0.7456	0.7316	0.7397	**0.9057**	0.9098	0.9212	0.9238	**0.9717**

Table 3. Performance comparisons on Reuters corpus

K	Accuracy							
	Non-normalized algorithms				Normalized algorithms			
	AA	NMF	AA-HHt	AA-PSPt	NC	NMF-NCW	NC-HHt	NC-PSPt
2	0.7976	0.8400	0.8242	**0.9071**	0.8998	**0.9069**	0.8881	**0.9041**
3	0.7110	0.7335	0.7333	**0.8752**	**0.8714**	**0.8748**	0.8686	**0.8745**
4	0.7147	0.7091	0.7124	**0.7846**	0.7757	0.7824	0.7775	**0.8019**
5	0.6424	0.6608	0.6652	**0.7719**	0.7805	0.7723	0.7728	**0.8000**
6	0.5875	0.6113	0.6041	**0.7604**	0.6713	0.6866	0.6906	**0.7424**
7	0.5688	0.5982	0.6006	**0.7225**	0.7181	0.7339	0.7257	**0.8031**
8	0.5400	0.5635	0.5528	**0.6639**	0.6431	0.6637	0.6631	**0.7275**
9	0.5272	0.5444	0.5444	**0.6716**	0.6048	0.6137	0.6170	**0.7023**
10	0.4984	0.5476	0.5302	**0.6627**	0.6026	0.6380	0.6257	**0.6975**
avg.	0.6208	0.6454	0.6408	**0.7578**	0.7297	0.7414	0.7366	**0.7837**
K	Mutual Information							
	Non-normalized algorithms				Normalized algorithms			
	AA	NMF	AA-HHt	AA-PSPt	NC	NMF-NCW	NC-HHt	NC-PSPt
2	0.4564	0.5017	0.4716	**0.5736**	0.6183	**0.6380**	0.6085	0.6173
3	0.4419	0.4587	0.4504	**0.6163**	0.6847	**0.7069**	**0.7015**	**0.7027**
4	0.5771	0.5402	0.5395	**0.6015**	0.5761	0.5887	0.5841	**0.5989**
5	0.5471	0.5420	0.5397	**0.5829**	**0.6492**	0.6365	0.6389	**0.6484**
6	0.5281	0.5137	0.5120	**0.5802**	0.5941	0.6032	0.6011	**0.6287**
7	0.5066	0.4853	0.4856	**0.5249**	0.6324	0.6354	0.6277	**0.7032**
8	**0.5077**	0.4970	0.4848	0.4921	0.5883	0.5952	0.6000	**0.6291**
9	0.5104	0.4944	0.4950	**0.5347**	0.5588	0.5564	0.5595	**0.6079**
10	**0.4971**	0.4890	**0.4901**	**0.4925**	0.5718	0.5840	0.5768	**0.6234**
avg.	0.5080	0.5024	0.4965	**0.5554**	0.6082	0.6160	0.6108	**0.6400**

From Table 2 and 3, we observed that:

1. Most of normalized methods performed better than their corresponding non-normalized methods. This means that the normalization brings positive effects for spectral clustering (NC vs. AA) and the NMF-based methods (NMF-NCW vs. NMF, NC-HHt vs. AA-HHt, and NC-PSPt vs. AA-PSPt).
2. Comparing eigenvector-based spectral clustering with previous NMF-based methods in the same group (AA vs. NMF and AA-HHt, NC vs. NMF and NC-HHt), we notice that: 1) the intuitiveness of solution is important: although the eigenvector-based methods is more interpretable than the previous NMF-based methods, however the latter perform slightly better than the former on the whole because the real-valued solution of the former is not intuitive and leads to unsatisfactory performances; 2) a reasonable clustering interpretation is necessary: although the previous NMF-based methods are more intuitive than the eigenvector-based methods, they aren't obviously superior to the eigenvector-based methods and sometimes are inferior; the possible reason is that they lack a reasonable clustering interpretation. The above two facts motivate us to find a NMF with a reasonable clustering interpretation, which can combine both intuitiveness and interpretation together.
3. Obviously, our proposed methods (AA-PSPt and NC-PSPt) perform much better than other methods in the same group on most data sets. The results indicate that our method indeed combines the interpretation of spectral clustering and the intuitiveness

of NMF-based methods together, and brings outstanding improvements to the performances. Actually, the methods AA, AA-HHt and AA-PSPt (or NC, NC-HHt and NC-PSPt) all provide a solution to AA (or NC) criterion function under relaxed conditions; however, the non-negative and orthogonal solution provided by AA-PSPt (or NC-PSPt) may be much closer to the indicator matrix for spectral clustering, which is also non-negative and orthogonal simultaneously. That also explains the outstanding performances of our proposed methods.

5 Conclusions

In this paper, we have proposed a novel NMF to the affinity matrix with non-negativity and orthogonality constraints simultaneously, and also have presented the iterative algorithm for computing the new factorization. With the help of additional orthogonality constraints, the novel NMF provides a non-negative solution to spectral clustering, which combines the interpretation of spectral clustering and the intuitiveness of NMF-based methods together. The experimental evaluations demonstrate that the proposed method is much more efficient and effective than the eigenvector-based spectral clustering and the previous NMF-based methods.

References

1. Li, T., Ma, S., Ogihara, M.: Document Clustering via Adaptive Subspace Iteration. In: Proceedings of the 27th ACM SIGIR Conference, pp. 218–225 (2004)
2. Xu, W., Liu, X., Gong, Y.: Document Clustering Based on Non-Negative Matrix Factorization. In: Proceedings of the 26th ACM SIGIR Conference, pp. 267–273 (2003)
3. Xu, W., Liu, X., Gong, Y.: Document Clustering by Concept Factorization. In: Proceedings of the 27th ACM SIGIR Conference, pp. 202–209 (2004)
4. Chan, P.K., Schlag, D.F., Zien, J.Y.: Spectral K-way Ratio-cut Partitioning and Clustering. IEEE Trans. on CAD 13, 1088–1096 (1994)
5. Shi, J., Malik, J.: Normalized Cuts and Image Segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence 22, 888–905 (2000)
6. Ding, C., He, X., Zha, H., et al.: A Min-max Cut Algorithm for Graph Partitioning and Data Clustering. In: Proceedings of the 2001 IEEE ICDM Conference, pp. 107–114 (2001)
7. von Luxburg, U.: A Tutorial on Spectral Clustering. Technical Report No. TR-149, Max Planck Institute for Biological Cybernetics (2006)
8. Lee, D.D., Seung, H.S.: Learning the Parts of Objects by Non-negative Matrix Factorization. Nature 401, 788–791 (1999)
9. Lee, D.D., Seung, H.S.: Algorithms for Non-negative Matrix Factorization. Advances in Neural Information Processing Systems 13, 556–562 (2001)
10. Deerwester, S.C., Dumais, S.T., Landauer, T.K., et al.: Indexing by Latent Semantic Analysis. Journal of the American Society of Information Science 41, 391–407 (1990)
11. Ding, C., He, X., Simon, H.D.: On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering. In: Proceedings of the 2005 SIAM Data Mining Conference, pp. 606–610 (2005)
12. Lütkepohl, H.: Handbook of Matrices. Wiley, Chichester (1997)
13. Long, B., Zhang, A., Wu, X., et al.: Relational Clustering by Symmetric Convex Coding. In: Proceeding of the 24th International Conference on Machine Learning, pp. 680–687 (2007)