Document Clustering with an Augmented Nonnegative Matrix Factorization Model

Zunyan Xiong, Yizhou Zang, Xingpeng Jiang, and Xiaohua Hu

College of Computing and Informatics, Drexel University, Philadelphia, USA
{zunyan.xiong,yizhou.zang,xingpeng.jiang,
xiaohua.hu}@drexel.edu

Abstract. In this paper, we propose an augmented NMF model to investigate the latent features of documents. The augmented NMF model incorporates the original nonnegative matrix factorization and the local invariance assumption on the document clustering. In our experiment, first we compare our model to baseline algorithms with several benchmark datasets. Then the effectiveness of the proposed model is evaluated using datasets from CiteULike. The clustering results are compared against the subject categories from Web of Science for the CiteULike dataset. Experiments of clustering on both benchmark data sets and CiteULike datasets outperforms many state of the art clustering methods.

Keywords: nonnegative matrix factorization, graph Laplacian, regularization, clustering, social tagging.

1 Introduction

Document clustering is an unsupervised machine learning technique aims to discover the classification of documents according to their similarities. So far many document clustering methods have been proposed, such as k-means [1], spectral clustering [2], non-negative matrix factorization (NMF) [3][4], and Probabilistic Latent Semantic Analysis (PLSA) [5] etc.

Besides, traditional clustering methods mostly apply linear dimensional reduction to extract features of the data set. However, recent research has shown that most data structures are nonlinear. To deal with this problem, researchers referred to the idea of manifold learning. Manifold learning [6] is a nonlinear matrix dimensionality reduction approach that tries to discover the low dimensional structure for the data in the high dimension. There are several popular manifold learning methods, such as Isomap [7][8][9], Locally Linear Embedding [10], and Laplacian Eigenmaps [11] etc. In [12], Cai et al. proposed a graph regularized non-negative matrix factorization (GNMF). GNMF embedded manifold learning into nonnegative matrix factorization by means of local invariance assumption. Experiments proved that the manifold learning can significantly improve the clustering results.

One defect of the aforementioned clustering methods is that they only focus on one dimension of the data. However, to better study the cluster dataset, it's important to explore the data structure from two dimensions, because the geometrical structures of

V.S. Tseng et al. (Eds.): PAKDD 2014, Part II, LNAI 8444, pp. 348–359, 2014.

vectors in two dimensions of the matrix are independent to each other. For example, in a document-term matrix, the similarities of the document vectors are independent of the term vectors.

Motivated by addressing this problem, we propose a novel model named augmented nonnegative matrix factorization (ANMF), which incorporates both matrix factorization and manifold learning on both dimensions of the data matrix. Then, we applied the method in a social tagging system, CiteULike. One of the biggest challenges here is how to establish a reliable clustering evaluation method. Since most contents of the social tagging systems are created by users, rarely any criterion exists to classify the contents, not to mention a gold standard for evaluation. In this paper, the dataset applied for the experiment is from CiteULike. In this paper, we solve the problem by using the subject classification from Web of Science for the CiteULike dataset. The classification provides an objective and reliable standard to test the effectiveness of algorithms.

2 Related Work

2.1 NMF

Nonnegative Matrix Factorization (NMF)[3][4][13] is widely applied in recent years and proved to be efficient and robust in various situations. NMF decomposes the original matrix to the product of two nonnegative matrices consisting of latent vectors. The factorized matrices consist of nonnegative components, which is convenient for data analysis. The relationship between the data points can be regarded as their distance of similarity on the graph.

Given a data matrix $\boldsymbol{X} = [X_1, \ldots, X_N] \in \mathbb{R}^{M \times N}$, the goal of nonnegative matrix factorization is to find two low-rank nonnegative matrices U and V whose product can best approximate the original matrix X:

$$X \approx UV^T$$

Xu et al. apply NMF method in document clustering, and the experiment results indicate that NMF method outperforms the latent semantic indexing and the spectral clustering methods in document clustering accuracies [14]. However, one defect of NMF is that it does not preserve the relational structure of the data during the factorization process. Two documents that are originally similar in their learned latent represents maybe dissimilar after factorization.

2.2 GNMF

In[12], Cai et al. introduced Graph Regularized Nonnegative Matrix Factorization (GNMF) model, which is an extension of Nonnegative Matrix Factorization (NMF). The two nonnegative matrices represent the latent factors of the original matrix from the two dimensions respectively. GNMF model applies the local invariance assumption on the one of the two nonnegative matrices, which results in a regularization term to the NMF objective function. The main idea of the local invariance assumption is that if two data points are close to each other in the original geometry, they should still be close in the new representation after factorized [12]. This is formulated as follows:

$$O = \|X - UV^T\|^2 + \lambda Tr(V^T LV)$$

where $Tr(\cdot)$ denotes the trace of a matrix. L is a Laplacian matrix, which is defined by $L = D - W$ [15]. D is a diagonal matrix and its diagonal entries are the sum of columns or rows of the weight matrix W (for W is a symmetric matrix), i.e., $d_{ii} = \sum_{j=1}^{M} w_{ij}$.

In the experiment, two image data sets and one document corpus were selected for clustering and showed GNMF outperforms NMF model, k-means and SVD methods. A flaw of this model is that it only considers the local invariance assumption from one dimension of the data.

2.3 Document Clustering on the Web

Many document clustering methods including the aforementioned ones are developed on the relation between documents and terms [14][5][12]. However, for web applications other than textual resources, it's difficult to get direct content information. Sometimes even the textual documents are no longer available due to the instability of webpages. Therefore, it is necessary to exploit other information sources to improve the clustering effectiveness. Moreover, studies in [1] and [2] proved that compared to textual contents or keywords of Web pages, social tag information is more reliable for clustering. Besides, most traditional content-based clustering algorithms such as k-means and NMF ignore the semantic relations among terms. As a result, two documents with no common terms will be regarded as dissimilar even though they have many synonymic or semantically related terms. Therefore, it's natural to incorporate other useful information to benefit the document clustering research.

Ramage and Heymann proposed two methods in Web clustering that include both term and tag information [16]. One applied k-means in an extended vector that includes both term and tag information; the other used the term and tag information in a generative clustering algorithm based on latent Dirichlet Allocation. Their study shows that including tagging data can significantly improve the clustering quality.

Lu et al. exploited the tripartite information, i.e. resources, users and tags, for webpage clustering [17]. The authors integrated the tripartite information together for better clustering performance. Three different methods are proposed in the paper. The first method applies the structure of the tripartite network to cluster Web pages; the second method uses the tripartite information in k-means by combining two or three different vectors together; the third method utilizes the Link k-means algorithm in the tripartite information. Results indicated that all clustering methods incorporating tagging information significantly outperform the content-based clustering. Furthermore, compared to the other two methods, the tripartite network has better performance.

3 Augmented Nonnegative Matrix Factorization Model

In this section, we first propose the augmented nonnegative matrix factorization model, which simultaneously incorporates the geometric structures of both the data manifold and the feature manifold. Then, we introduce the model and its iterative algorithm in detail.

The objective function of the ANMF model is:

$$O = \|X - UV^T\|^2 + \lambda Tr(V^T LV) + \mu Tr(U^T \tilde{L} U)$$

3.1 Notations and Problem Formalization

Before describing the model, some useful definitions are introduced. Given a data set, $\mathcal{D} = \{d_1, d_2, \dots, d_M\}$, and $\mathcal{F} = \{f_1, f_2, \dots, f_N\}$ be the set of data features. Their relational matrix is denoted by $X = (x_{ij})_{M\times N}$, where x_{ij} denotes the weighting of the data feature f_j for the data point d_i.

For convenience, the meaning of notations used in the paper is summarized in Table 1.

Table 1. Important notation used in this paper

Notations	Description	Notations	Description
$\mathcal{D}$	data set	$\mathcal{F}$	data feature set
M	number of data points	N	number of features
X	data matrix of size $M \times N$	x_{ij}	data point in the matrix
U	data partition of size $M \times K$	V	feature partition of size $N \times K$
$\boldsymbol{u}_k$	kth column of U	$\boldsymbol{v}_k$	kth column of V
L	data graph Laplacian	D	data degree matrix
W	feature adjacency matrix	$\widetilde{W}$	data adjacency matrix
$w_{jj'}$	a cell in the feature adjacency matrix	$\widetilde{w}_{ii'}$	a cell in the data adjacency matrix
K	the number of latent component		

To analyze the similarities between the data points and the features respectively, the nonnegative matrix factorization can be applied to decompose the matrix X

$$X \approx UV^T,$$

such that the matrix U and V consists of the latent vectors associated to data points and features, i.e. the row vectors of U and V represent the latent vectors for the data d_i and feature f_j respectively. In this approach, the latent vectors are supposed to represent the factorized meaningful parts (or topic) of the data set and their features. In order to quantize the similarities of data and features, we next use the idea of local invariance assumption to obtain two regularization terms, Regularizer I and Regularizer II.

3.2 Local Invariance Assumption

The local invariance assumption is a general principle that can be interpreted in this context as following. If the data are close in some sense, after the NMF decomposition, they should still be close in the latent space. To measure the similarities between points of the original data, we construct two adjacency matrix W and $\widetilde{W}$ from X for both feature and data vectors. The metric of the adjacency (closeness) between the vectors $w_{jj'}$ (or $\widetilde{w}_{ii'}$) can be defined in different ways, such as 0-1 weighting, heat kernel weighting and dot-product weighting, the definitions of the three weighting modes can be referenced in [18]. For each data point, only the p nearest neighbors are considered.

In the NMF decomposition, let K be the number of latent component, and $K \ll \mathrm{M}$, $K \ll N$. The data and features are mapped to points in a lower K dimensional Euclidean space. From a geometric point of view, their similarity can be easily compared by Euclidean distance.

Consider the matrix $X = (x_{ij})_{M \times N}$, let $\boldsymbol{x}_j$ be the jth column vector, i.e. $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots \boldsymbol{x}_N]$. Then $\boldsymbol{x}_j$ can be regarded as the coordinates of feature $\boldsymbol{f}_j$ in the standard basis. Under the matrix decomposition,

$$X \approx UV^T,$$

Let the kth column vector of U be $\boldsymbol{u}_k{}^{(K)} = \langle u_{ik}{}^{(K)} | i = 1, 2, \dots, |M| \rangle$. Then the original vector $\boldsymbol{x}_j$ is approximated by the linear combination of vectors $\boldsymbol{u}$'s:

$$\boldsymbol{x}_j \approx \textstyle\sum_{k=1}^{K} \boldsymbol{u}_k \, v_{jk} \tag{1}$$

In this expression, now $\boldsymbol{u}_k$'s can be regarded as new basis vectors for the latent space, and the new coordinates for feature $\boldsymbol{f}_j$ hence are $\boldsymbol{v}_j{}^{(K)} = \langle v_{jk}{}^{(K)} | k = 1, 2, \dots, |K| \rangle$. According to the local invariance assumption, if feature vector $\boldsymbol{f}_j$ and $\boldsymbol{f}_{j'}$ are close in the original coordinates, they should still be close in the new coordinates. To quantize this information, we use the Euclidean distance in the latent space $\|\boldsymbol{f}_j - \boldsymbol{f}_{j'}\|$, weighted by their original closeness $w_{jj'}$. As stated previously, $w_{jj'}$ can be calculated by 0-1 weighting, heat kernel weighting or dot-product weighting. Also as in [12], we define the Regularizer I as

$$\mathcal{R}_1 = \frac{1}{2} \textstyle\sum_{j,j'=1}^{N} \left\| \boldsymbol{v}_j - \boldsymbol{v}_{j'} \right\|^2 w_{jj'}$$

It can be seen heuristically that for $\mathcal{R}_1$ bounded, if $w_{jj'}$ is large, meaning features j, j' are close in the adjacency, the Euclidean distance is forced to be small, which implies the factored feature $\boldsymbol{v}_j$, $\boldsymbol{v}_{j'}$ are close. For computational convenience, we simplify the regularizer as following.

$$\mathcal{R}_1 = \frac{1}{2} \textstyle\sum_{j,j'=1}^{N} \left\| \boldsymbol{v}_j - \boldsymbol{v}_{j'} \right\|^2 w_{jj'}$$

$$= \sum_{j=1}^{N} \boldsymbol{v}_j^T \boldsymbol{v}_j w_{jj} - \sum_{j,j'=1}^{N} \boldsymbol{v}_j^T \boldsymbol{v}_{j'} w_{jj'}$$

$$= Tr(V^T DV) - Tr(V^T WV) = Tr(V^T LV),$$

where $Tr(\cdot)$ denotes the trace of a matrix. D is a diagonal matrix and its diagonal entries are the sum of columns or rows of W (for W is a symmetric matrix), i.e., $d_{ii} = \sum_{j=1}^{N} w_{ij}$. The Laplacian matrix L is defined by $L = D - W$ [15].

Incorporating this information to the NMF model, the objective function now becomes

$$\mathcal{O} = \|X - UV^T\|^2 + \lambda Tr(V^T LV) \tag{2}$$

Here λ is a regularization parameter that balances the effects of local invariance. This is the model considered in [1], called the graph regularized NMF method (GNMF).

At this point, it is important to notice that for our problem, the local invariance assumption applies to the other piece of data, the features.

To reflect the local invariance of the data, a second regularization term is added:

$$\mathcal{R}_2 = \frac{1}{2}\sum_{i,i'=1}^{N} \|\boldsymbol{u}_i - \boldsymbol{u}_{i'}\|^2 \widetilde{w}_{ii'}$$

$$= \sum_{i=1}^{N} \boldsymbol{u}_i^T \boldsymbol{u}_i d_{ii} - \sum_{i,i'=1}^{N} \boldsymbol{u}_i^T \boldsymbol{v} \boldsymbol{u}_{i'} \widetilde{w}_{ii'}$$

$$= Tr\left(U^T \widetilde{D} U\right) - Tr\left(U^T \widetilde{W} U\right)$$
$$= Tr\left(U^T \widetilde{L} U\right)$$

where $\widetilde{W}$ is the adjacency matrix for the data and $\widetilde{L} = \widetilde{D} - \widetilde{W}$. Now the final cost function can be defined as

$$\mathcal{O} = \|X - UV^T\|^2 + \lambda Tr(V^T LV) + \mu Tr(U^T \widetilde{L} U) \tag{3}$$

We call this new model the augmented NMF (ANMF). Here the two regularization parameters λ and μ are positive numbers to be chosen later. They balance the effects of local invariance and the original NMF. Heuristically, the larger the parameters, the stronger will the local invariance be reflected in the results. The optimal solution is obtained by minimizing $\mathcal{O}$ over all non-negative matrices U and V. We will discuss the algorithms in next section.

3.3 Iterative Algorithm

As in the original matrix factorization model [19] or the GNMF model [12], the cost function $\mathcal{O}$ is not convex in W and H jointly. Thus it is not possible to find global minima. However, it is convex in W for fixed H and vice versa. In fact, the Lagrange multiplier method used in [2] is also applicable here to give an iterative algorithm. However the updating rules could only be expected converge to a local (not global) minima.

The cost function can be rewritten as

$$O = Tr(X - UV^T)(X - UV^T)^T + \lambda Tr(V^T LV) + \mu Tr\left(U^T \tilde{L} U\right)$$

$$= Tr(XX^T) - 2Tr(XVU^T) + Tr(UV^T VU^T) + \lambda Tr(V^T LV) + \mu Tr\left(U^T \tilde{L} U\right)$$

Here the basic properties $Tr(A) = Tr(A^T)$ and $Tr(AB) = Tr(BA)$ are used for any matrices A and B. Next let ψ_{ik} be the Lagrange multiplier for the condition $u_{ik} \geq 0$, and ϕ_{jk} be the multiplier for the condition $v_{jk} \geq 0$. The augmented Lagrangian is

$$\mathcal{L} = Tr(XX^T) - 2Tr(XVU^T) + Tr(UV^T VU^T) + \lambda Tr(V^T LV) + \mu Tr\left(U^T \tilde{L} U\right) + Tr(\Psi \mathrm{U}) + Tr(\Phi V^T) \tag{4}$$

where $\Psi = (\psi_{ik})_{M \times K}$ and $\Phi = (\phi_{jk})_{N \times K}$. The partial derivatives are

$$\frac{\partial \mathcal{L}}{\partial U} = -2XV + 2UV^T V + 2\mu \tilde{L} U + \Psi \tag{5}$$

$$\frac{\partial \mathcal{L}}{\partial V} = -2XU + 2VU^T U + 2\lambda LV + \Phi \tag{6}$$

The derivatives vanish at local minima. Using the Karush-Kuhn-Tucker (KKT) condition, $\psi_{ik} u_{ik} = 0, \phi_{ik} v_{ik} = 0$. The equations (5) and (6) become

$$-(XV)_{iku_{ik}} + (UV^T V)_{iku_{ik}} + \mu\left(\tilde{L} U\right)_{iku_{ik}} = 0 \tag{7}$$

$$-(X^T U)_{jkv_{jk}} + (VU^T U)_{jkv_{jk}} + \lambda (LU)_{jku_{jk}} = 0 \tag{8}$$

These equations give the following updating rules

$$u_{ik} \leftarrow u_{ik} \frac{(XV + \mu \widetilde{W} U)_{ik}}{(UV^T V + \mu \widetilde{D} U)_{ik}} \tag{9}$$

$$v_{jk} \leftarrow v_{jk} \frac{\left(X^T U + \lambda WV\right)_{jk}}{\left(VU^T U + \lambda DV\right)_{jk}} \tag{10}$$

The updating rules of our model actually lead to convergence sequences, which are justified by Theorem 1 and its proof below.

Theorem 1. Under the updating rules (9) and (10), the objective function (3) is non-increasing.

As in [12] and [18], the proof of Theorem 1 is essentially based on the existence of a proper auxiliary function for the ANMF. We give a simple proof on the ground of the following results from [12].

Lemma 2. Under the updating rule

$$v_{jk} \leftarrow v_{jk} \frac{\left(X^T U + \lambda WV\right)_{jk}}{\left(VU^T U + \lambda DV\right)_{jk}} \tag{11}$$

The cost function $\mathcal{O}_1$ in GNMF, i.e.

$$\mathcal{O}_1 = \|X - UV^T\|^2 + \lambda Tr(V^T L V) \tag{12}$$

is non-increasing.

Proof of Theorem 1. Consider the objective function $\mathcal{O}$ under the updating of V by (11). Then the last term $\mu Tr\left(U^T \tilde{L} U\right)$ in $\mathcal{O}$ will not change. It suffices to prove $\mathcal{O}_V = \mathcal{O} - \mu Tr\left(U^T \tilde{L} U\right) = \mathcal{O}_1$ is non-increasing, which is exactly given by Lemma 2. Next consider $\mathcal{O}$ under the updating of U. Since H is not changed, it suffices to consider

$$\begin{aligned}\mathcal{O}_U = \mathcal{O} - \lambda Tr(V^T L V) &= \|X - UV^T\|^2 + \mu Tr\left(U^T \tilde{L} U\right) \\ &= \|X^T - VU^T\|^2 + \mu Tr\left(U^T \tilde{L} U\right)\end{aligned}$$

Now interchange U, V and replace μ by λ, X by X^T in Lemma 2, $\mathcal{O}_U$ is not increasing under the updating of W by (9). ■

One problem of the objective function of ANMF is that the solutions U and V are not unique. If U and V are the solutions, then UD and VD^{-1} can also be the solutions of the objective function. To obtain unique solutions, we refer to the approach from [12] that enforces the Euclidean distance of the column vectors in matrix U as one. This approach can be achieved by

$$u_{ik} \leftarrow \frac{u_{ik}}{\sqrt{\sum_i u_{ik}^2}} \tag{13}$$

$$v_{jk} \leftarrow v_{jk}\sqrt{\sum_i u_{ik}^2} \tag{34}$$

Table 2 shows the simple algorithm of ANMF model.

Table 2. Algorithm of ANMF

```
Input: the data matrix $X$, regularization parameter $\lambda$ and $\mu$.
Output: the data-topic matrix $U$, and the topic-feature matrix $V$.
Method:
Construct weighting matrix $W$ and $\widetilde{W}$, compute the diagonal matrix $D$ and $\widetilde{D}$;
Random initialize U and V;
Repeat (9) and (10) until convergence;
Normalize U and V using (13) and (14).
```

3.4 Complexity Analysis

In this section, the computational cost of NMF, GNMF and ANMF algorithms are discussed. Supposing the algorithm stops after t iterations, the overall cost for NMF is $O(tMNK)$. For GNMF, the adjacency matrix needs $O(N^2M)$ to construct, so the overall cost for GNMF is $O(tMNK + N^2M)$. As ANMF adds one more adjacency matrix on the other dimension, so the overall cost for ANMF is $O(tMNK + N^2M + M^2N)$.

4 Experiments

4.1 Data Sets and Evaluation Metrics

Before applying CiteULike data set, four data sets were chosen as the benchmark, which were Coil20, ORL, TDT2, and Reuters-21578. Two of them are image data and the other two are text data.

The results of our experiments were evaluated by Clustering Accuracy [20] and normalized mutual information (NMI) [21]. Both of the evaluation metrics range from zero to one, and a high value indicates better clustering result.

4.2 Parameter Settings

In this section, we compared our proposed method with the following methods, K-means [22], NMF [4], and GNMF [12]. For both GNMF and ANMF, we normalized the vectors on columns of W and H.

To fairly compare algorithms, each algorithm was run under different parameter settings, and the best results were selected to compare with each other. The number of clusters was set equal to the true number of standard categories for all the data sets and clustering algorithms.

The 0-1 weighting was applied in GNMF and ANMF algorithms for convenience. Here we set the nearest neighborhood p as 7 for both the algorithms. The value of p determines the construction of the adjacency matrix for both GNMF and ANMF, which lies on the assumption that the neighboring data points share the same topic. So the performance of GNMF and ANMF are supposed to decrease as p increases, which was verified by [12] for GNMF. There is only one regularization parameter in GNMF, the parameter was set by the grid $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}, 10^{3}, 10^{4}\}$. For ANMF algorithm, there are two regularization parameters λ and μ. Both of them were set by the grid $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}, 10^{3}, 10^{4}\}$.

The aforementioned algorithms were repeated 20 times for each parameter setting, the average results were computed. The best average results are shown in Table 3 and Table 4.

4.3 Clustering Results

Table 3 and 4 display the Accuracy and NMI of all algorithms on the four data sets respectively. We can see that overall both GNMF and ANMF performed much better than K-means and NMF algorithms. Note that both GNMF and ANMF consider the geometrical structure of the data through the local invariance assumption, the results imply the importance of the geometrical structure in mining the latent features of the data. Besides, ANMF shows the best performance in all the four data sets, which indicates that by adding the geometrical structure for the two dimensions of the data, the algorithm can achieve better performance.

Table 3. Clustering Accuracy (%)

Data Sets	*k*-means	NMF	GNMF	ANMF
Coil20	95.56%	95.90%	97.80%	**97.82%**
ORL	96.40%	96.01%	96.61%	**97.19%**
TDT2	90.92%	90.11%	95.09%	**95.35%**
Reuters-21578	74.04%	73.68%	74.68%	**75.13%**

Table 4. Normalized Mutual Information (%)

Data Sets	*k*-means	NMF	GNMF	ANMF
Coil20	73.86%	74.36%	89.17%	**90.14%**
ORL	71.82%	66.80%	72.01%	**75.24%**
TDT2	64.54%	58.75%	83.49%	**84.65%**
Reuters-21578	33.90%	29.98%	34.41%	**36.31%**

5 Study on the CiteULike Data Set

5.1 Data Processing

CiteULike is a social bookmarking platform that allows researchers to share scientific references, so nearly all the bookmarks in CiteULike are academic papers. The CiteULike data was crawled during January-December 2008. We extracted the article id, journal name of the articles, user id and tag information from the original data. The journal name of the articles was used for setting evaluation standard. Before processing the dataset, we unified the format of the tags. Tags such as “data_mining”, “data-mining”, “data.mining”, “datamining”, etc. were all considered as the same one. Here we excluded the articles, users and tags with less than four bookmarks. To evaluate the CiteULike dataset, we utilized the subject categories in Web of Science [23]. There are a total of 176 top-level subject categories for science journals. Under each subject category, they display a list of the afflicted journals. By overlapping the journals of all articles from CiteULike with the journals under the categories in Web of Science, we could discover the subject categories of the articles in CiteULike dataset. Under the 176 subject categories, we only kept the 44 biggest subject categories with the largest articles numbers. Finally, we had 3,296 bookmarks with 2406 articles, 1220 users and 4593 tags.

5.2 Clustering Results

We construct two matrices for CiteULike data set, article-user matrix and article-tag matrix. Besides, in order to test if combining the article vectors from article-user and article-tag vectors can get a better performance, we also construct a new matrix that consists of the linear combination of the article-user vectors and article-tag vectors. Just as the experiments in section 4, we compare the clustering results of ANMF with GNMF, NMF and k-means based on the Clustering Accuracy and NMI. The settings for the parameters and the value of the nearest neighborhood p are all the same as in section 4.

Table 5. The Evaluation Results for CiteULike Data Set

Data Sets	***k*-means**	**NMF**	**GNMF**	**ANMF**
	Clustering Accuracy (%)			
article-user matrix	30.04%	44.22%	86.57%	**87.17%**
article -tag matrix	73.42%	76.24%	**88.46%**	88.43%
the combination matrix	68.65%	68.94%	85.48%	**87.60%**
	Normalized Mutual Information (%)			
article -user matrix	10.83%	19.03%	27.24%	**28.55%**
article -tag matrix	25.00%	27.72%	32.07%	**36.85%**
the combination matrix	26.24%	23.85%	36.91%	**42.35%**

Table 5 displays the evaluation scores of the four algorithms with CiteULike dataset. The experiments reveal several interesting points:

- ANMF still performs the best among the four algorithms. Specifically, the improvement is significant in NMI results. This shows that ANMF is efficient not only in image and text data, but also in the data from social tagging systems, which suggests the potential of ANMF in collaborative filtering area.
- The evaluation results of the combination matrix are rather poor for k-means and NMF algorithms for the article-user matrix and article-tag matrix. For GNMF and ANMF, their NMI scores are better than the other two matrices, while the Clustering Accuracy scores are a little lower.

6 Conclusion

In this paper, we have explored a graph regularized nonnegative matrix factorization model for document clustering. First, we applied our algorithm in four benchmark data sets, and compared it with three canonical algorithms to evaluate its performance in clustering. Then the algorithm was used in CiteULike dataset by applying user and tag information for analysis. The experiment results demonstrate that our algorithm outperforms GNMF, NMF and k-means models in both benchmark data sets and CiteULike data set.

References

1. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. CSUR 31(3), 264–323 (1999)
2. Guy, I., Carmel, D.: Social recommender systems. In: Proceedings of the 20th International Conference Companion on World Wide Web, pp. 283–284 (2011)
3. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401(6755), 788–791 (1999)
4. Seung, D., Lee, L.: Algorithms for non-negative matrix factorization. Adv. Neural Inf. Process. Syst. 13, 556–562 (2001)

5. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57 (1999)
6. Law, M.H., Jain, A.K.: Incremental nonlinear dimensionality reduction by manifold learning. Pattern Anal. Mach. Intell. IEEE Trans. 28(3), 377–391 (2006)
7. Balasubramanian, M., Schwartz, E.L.: The isomap algorithm and topological stability. Science 295(5552), 7–7 (2002)
8. Bengio, Y., Paiement, J.-F., Vincent, P., Delalleau, O., Le Roux, N., Ouimet, M.: Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. Adv. Neural Inf. Process. Syst. 16, 177–184 (2004)
9. Samko, O., Marshall, A.D., Rosin, P.L.: Selection of the optimal parameter value for the Isomap algorithm. Pattern Recognit. Lett. 27(9), 968–979 (2006)
10. Tenenbaum, J.B., De Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290(5500), 2319–2323 (2000)
11. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. 15(6), 1373–1396 (2003)
12. Cai, D., He, X., Han, J., Huang, T.S.: Graph regularized nonnegative matrix factorization for data representation. Pattern Anal. Mach. Intell. IEEE Trans. 33(8), 1548–1560 (2011)
13. Gaussier, E., Goutte, C.: Relation between PLSA and NMF and implications. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 601–602 (2005)
14. Xu, W., Liu, X., Gong, Y.: Document clustering based on non-negative matrix factorization. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 267–273 (2003)
15. Chung, F.R.: Spectral graph theory, vol. 92. AMS Bookstore (1997)
16. Ramage, D., Heymann, P., Manning, C.D., Garcia-Molina, H.: Clustering the tagged web. In: Proceedings of the Second ACM International Conference on Web Search and Data Mining, pp. 54–63 (2009)
17. Lu, C., Hu, X., Park, J.: Exploiting the social tagging network for Web clustering. Syst. Man Cybern. Part Syst. Humans IEEE Trans. 41(5), 840–852 (2011)
18. Matlab Codes and Datasets for Feature Learning, `http://www.cad.zju.edu.cn/home/dengcai/Data/data.html` (accessed: September 18, 2013)
19. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer 42(8), 30–37 (2009)
20. Gu, Q., Zhou, J.: Co-clustering on manifolds. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 359–368 (2009)
21. Strehl, A., Ghosh, J.: Cluster ensembles—a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. 3, 583–617 (2003)
22. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, p. 14 (1967)
23. Journal Search - IP & Science - Thomson Reuters, `http://www.thomsonscientific.com/cgi-bin/jrnlst/jlsubcatg.cgi?PC=D` (accessed: October 01, 2013)