Document Clustering Based On Non-negative Matrix Factorization

Wei Xu, Xin Liu, Yihong Gong

NEC Laboratories America, Inc.
10080 North Wolfe Road, SW3-350
Cupertino, CA 95014, U.S.A.

{xw,xliu,ygong}@ccrl.sj.nec.com

ABSTRACT

In this paper, we propose a novel document clustering method based on the non-negative factorization of the term-document matrix of the given document corpus. In the latent semantic space derived by the non-negative matrix factorization (NMF), each axis captures the base topic of a particular document cluster, and each document is represented as an additive combination of the base topics. The cluster membership of each document can be easily determined by finding the base topic (the axis) with which the document has the largest projection value. Our experimental evaluations show that the proposed document clustering method surpasses the latent semantic indexing and the spectral clustering methods not only in the easy and reliable derivation of document clustering results, but also in document clustering accuracies.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering*

General Terms

Algorithms

Keywords

Document Clustering, Non-negative Matrix Factorization

1. INDRODUCTION

Document clustering techniques have been receiving more and more attentions as a fundamental and enabling tool for efficient organization, navigation, retrieval, and summarization of huge volumes of text documents. With a good document clustering method, computers can automatically organize a document corpus into a meaningful cluster hierarchy, which enables an efficient browsing and navigation of the corpus. An efficient document browsing and navigation is a valuable complement to the deficiencies of traditional IR technologies. As pointed out in [3], the variety of information retrieval needs can be expressed by a spectrum where at one end is a narrowly specified search for documents matching the user's query, and at the other end is a broad information need such as what are the major international events in the year 2001, or a need without well defined goals but to learn more about general contents of the data corpus. Traditional text search engines fit well for covering one end of the spectrum, which is a keyword-based search for specific documents, while browsing through a cluster hierarchy is more effective for serving the information retrieval needs from the rest part of the spectrum.

In recent years, research on topic detection and tracking, document content summarization and filtering has received enormous attention in the information retrieval community. Topic detection and tracking (TDT) aim to automatically detect salient topics from either a given document corpus or an incoming document stream and to associate each document with one of the detected topics. The TDT problems can be considered as a special case of the document clustering problem and actually most of the TDT systems in the literature were realized by adapting various document clustering techniques. On the other hand, document summarization is intended to create a document abstract by extracting sentences/paragraphs that best present the main content of the original document. Again, many proposed summarization systems employed clustering techniques for identifying distinct content, and finding semantically similar sentences of the document.

Document clustering methods can be mainly categorized into two types: document partitioning (flat clustering) and agglomerative (bottom-up hierarchical) clustering. Although both types of methods have been extensively investigated for several decades, accurately clustering documents without domain-dependent background information, nor predefined document categories or a given list of topics is still a challenging task.

In this paper, we propose a novel document partitioning method based on the non-negative factorization of the term-document matrix of the given document corpus. In the latent semantic space derived by the non-negative matrix factorization (NMF) [7], each axis captures the base topic of a particular document cluster, and each document is represented as an additive combination of the base topics. The cluster membership of each document can be easily de-

SIGIR'03 July 28–August 1, 2003, Toronto, Canada.

termined by finding the base topic (the axis) with which the document has the largest projection value. Our method differs from the latent semantic indexing method based on the singular vector decomposition (SVD) and the related spectral clustering methods in that the latent semantic space derived by NMF does not need to be orthogonal, and that each document is guaranteed to take only non-negative values in all the latent semantic directions. These two differences bring about an important benefit that each axis in the space derived by the NMF has a much more straightforward correspondence with each document cluster than in the space derived by the SVD, and thereby document clustering results can be directly derived without additional clustering operations. Our experimental evaluations show that the proposed document clustering method surpasses SVD- and the eigenvector-based clustering methods not only in the easy and reliable derivation of document clustering results, but also in document clustering accuracies.

2. RELATED WORKS

Generally, clustering methods can be categorized as agglomerative and partitional. Agglomerative clustering methods group the data points into a hierarchical tree structure, or a dendrogram, by bottom-up approach. The procedure starts by placing each data point into a distinct cluster and then iteratively merges the two most similar clusters into one parent cluster. Upon completion, the procedure automatically generates a hierarchical structure for the data set. The complexity of these algorithms is $O(n^2 \log n)$ where n is the number of data points in the data set. Because of the quadratic order of complexity, bottom-up agglomerative clustering methods could become computationally prohibitive for clustering tasks that deal with millions of data points.

On the other hand, document partitioning methods decompose a document corpus into a given number of disjoint clusters which are optimal in terms of some predefined criteria functions. Partitioning methods can also generate a hierarchical structure of the document corpus by iteratively partitioning a large cluster into smaller clusters. Typical methods in this category include K-Means clustering [12], probabilistic clustering using the Naive Bayes or Gaussian mixture model [1, 9], etc. K-Means produces a cluster set that minimizes the sum of squared errors between the documents and the cluster centers, while both the Naive Bayes and the Gaussian mixture models assign each document to the cluster that provides the maximum likelihood probability. The common drawback associated with these methods is that they all make harsh simplifying assumptions on the distribution of the document corpus to be clustered. K-Means assumes that each cluster in the document corpus has a compact shape, the Naive Bayes model assumes that all the dimensions of the feature space representing the document corpus are independent of each other, and the Gaussian mixture model assumes that the density of each cluster can be approximated by a Gaussian distribution. Obviously, these assumptions do not often hold true, and document clustering results could be terribly wrong with broken assumptions.

There have been research studies that perform document clustering using the latent semantic indexing method (LSI) [4]. This method basically projects each document into the singular vector space through the SVD, and then conducts document clustering using traditional data clustering algorithms (such as K-means) in the transformed space. Although it was claimed that each dimension of the singular vector space captures a base latent semantics of the document corpus, and that each document is jointly indexed by the base latent semantics in this space, negative values in some of the dimensions generated by the SVD, however, make the above explanation less meaningful.

In recent years, spectral clustering based on graph partitioning theories has emerged as one of the most effective document clustering tools. These methods model the given document set using a undirected graph in which each node represents a document, and each edge (i, j) is assigned a weight w_{ij} to reflect the similarity between documents i and j. The document clustering task is accomplished by finding the best cuts of the graph that optimize certain predefined criterion functions. The optimization of the criterion functions usually leads to the computation of singular vectors or eigenvectors of certain graph affinity matrices, and the clustering result can be derived from the obtained eigenvector space. Many criterion functions, such as the Average Cut [2], Average Association [11], Normalized Cut [11], Min-Max Cut [5], etc, have been proposed along with the efficient algorithms for finding their optimal solutions. It can be proven that under certain conditions, the eigenvector spaces computed by these methods are equivalent to the latent semantic space derived by the LSI method. As spectral clustering methods do not make naive assumptions on data distributions, and the optimization accomplished by solving certain generalized eigenvalue systems theoretically guarantees globally optimal solutions, these methods are generally far more superior than traditional document clustering approaches. However, because of the use of singular vector or eigenvector spaces, all the methods in this category have the same problem as LSI, i.e., the eigenvectors computed from the graph affinity matrices usually do not correspond directly to individual clusters, and consequently, traditional data clustering methods such as K-means have to be applied in the eigenvector spaces to find the final document clusters.

3. THE PROPOSED METHOD

Assume that a document corpus is comprised of k clusters each of which corresponds to a coherent topic. Each document in the corpus either completely belongs to a particular topic, or is more or less related to several topics. To accurately cluster the given document corpus, it is ideal to project the document corpus into a k-dimensional semantic space in which each axis corresponds to a particular topic. In such a semantic space, each document can be represented as a linear combination of the k topics. Because it is more natural to consider each document as an additive rather than subtractive mixture of the underlying topics, the linear combination coefficients should all take non-negative values. Furthermore, it is also quite common that the topics comprising a document corpus are not completely independent of each other, and there are some overlaps among them. In such a case, the axes of the semantic space that capture each of the topics are not necessarily orthogonal. Based on the above discussions, we propose to use non-negative matrix factorization (NMF) to find the latent semantic structure for the document corpus, and identify document clusters in the derived latent semantic space.

In fact document clustering methods based on the LSI and the spectral clustering, as described in Section 2, also strive to find the latent semantic structure for the document corpus by computing singular vectors or eigenvectors of certain matrices. The derived latent semantic space is orthogonal, and each document can take negative values in some directions in the space.

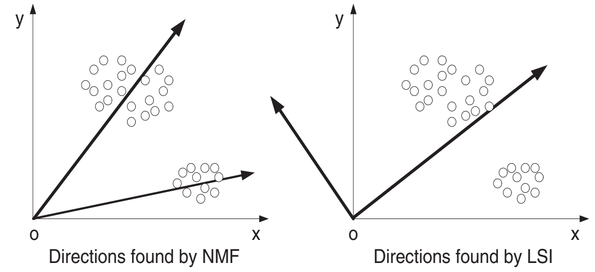


Figure 1: Illustration of the differences between NMF and LSI.

In contrast, NMF does not require the derived latent semantic space to be orthogonal, and it guarantees that each document takes only non-negative values in all the latent semantic directions. These two characteristics make the NMF superior to the LSI and spectral clustering methods because of the following reasons (See Figure 1). First, when overlap exists among clusters, NMF can still find a latent semantic direction for each cluster, while the orthogonal requirement by the SVD or the eigenvector computation makes the derived latent semantic directions less likely to correspond to each of the clusters. Second, with NMF, a document is an additive combination of the base latent semantics, which makes more sense in the text domain. Third, as the direct benefit of the above two NMF characteristics, the cluster membership of each document can be easily identified from NMF, while the latent semantic space derived by the LSI or the spectral clustering does not provide a direct indication of the data partitions, and consequently, traditional data clustering methods such as K-means have to be applied in this eigenvector space to find the final set of document clusters.

The following subsections provide the detailed descriptions of the proposed document clustering method.

3.1 Document Representation

We use the weighted term-frequency vector to represent each document. Let $\mathcal{W} = \{f_1, f_2, \ldots, f_m\}$ be the complete vocabulary set of the document corpus after the stop-words removal and words stemming operations. The term-frequency vector X_i of document d_i is defined as

$$\begin{aligned} X_i &= [x_{1i}, x_{2i}, \ldots, x_{mi}]^T \\ x_{ji} &= t_{ji} \cdot \log\left(\frac{n}{idf_j}\right) \end{aligned}$$

where t_{ji}, idf_j, n denote the term frequency of word $f_j \in \mathcal{W}$ in document d_i, the number of documents containing word f_j, and the total number of documents in the corpus, respectively. In addition, X_i is normalized to unit Euclidean length. Using X_i as the i'th column, we construct the $m \times n$ term-document matrix $\mathbf{X}$. This matrix will be used to conduct the non-negative factorization, and the document clustering result will be directly obtained from the factorization result.

3.2 Document Clustering Based on NMF

NMF is a matrix factorization algorithm that finds the positive factorization of a given positive matrix [7, 8, 6]. Assume that the given document corpus consists of k document clusters. Here the goal is to factorize $\mathbf{X}$ into the non-negative $m \times k$ matrix $\mathbf{U}$ and the non-negative $k \times n$ matrix $\mathbf{V}^T$ that minimize the following objective function:

$$J = \frac{1}{2}\|\mathbf{X} - \mathbf{U}\mathbf{V}^T\| \tag{1}$$

where $\| \cdot \|$ denotes the squared sum of all the elements in the matrix. The objective function J can be re-written as:

$$\begin{aligned} J &= \frac{1}{2}\mathrm{tr}((\mathbf{X} - \mathbf{U}\mathbf{V}^T)(\mathbf{X} - \mathbf{U}\mathbf{V}^T)^T) \\ &= \frac{1}{2}\mathrm{tr}(\mathbf{X}\mathbf{X}^T - 2\mathbf{X}\mathbf{V}\mathbf{U}^T + \mathbf{U}\mathbf{V}^T\mathbf{V}\mathbf{U}^T) \\ &= \frac{1}{2}(\mathrm{tr}(\mathbf{X}\mathbf{X}^T) - 2\mathrm{tr}(\mathbf{X}\mathbf{V}\mathbf{U}^T) + \mathrm{tr}(\mathbf{U}\mathbf{V}^T\mathbf{V}\mathbf{U}^T)) \end{aligned} \tag{2}$$

where the second step of derivation uses the matrix property $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$. Let $\mathbf{U} = [u_{ij}]$, $\mathbf{V} = [v_{ij}]$, $\mathbf{U} = [U_1, U_2, \ldots, U_k]$. The above minimization problem can be restated as follows: minimize J with respect to $\mathbf{U}$ and $\mathbf{V}$ under the constraints of $u_{ij} \geq 0$, $v_{xy} \geq 0$, where $0 \leq i \leq m$, $0 \leq j \leq k$, $0 \leq x \leq n$, and $0 \leq y \leq k$. This is a typical constrainted optimization problem, and can be solved using the Lagrange multiplier method. Let α_{ij} and β_{ij} be the Lagrange multiplier for constraint $u_{ij} \geq 0$ and $v_{ij} \geq 0$, respectively, and $\alpha = [\alpha_{ij}]$, $\beta = [\beta_{ij}]$, the Lagrange L is,

$$L = J + \mathrm{tr}(\alpha\mathbf{U}^T) + \mathrm{tr}(\beta\mathbf{V}^T) \tag{3}$$

The derivatives of L with respect to $\mathbf{U}$ and $\mathbf{V}$ are:

$$\frac{\partial L}{\partial \mathbf{U}} = -\mathbf{X}\mathbf{V} + \mathbf{U}\mathbf{V}^T\mathbf{V} + \alpha \tag{4}$$

$$\frac{\partial L}{\partial \mathbf{V}} = -\mathbf{X}^T\mathbf{U} + \mathbf{V}\mathbf{U}^T\mathbf{U} + \beta \tag{5}$$

Using the Kuhn-Tucker condition $\alpha_{ij}u_{ij} = 0$ and $\beta_{ij}v_{ij} = 0$, we get the following equations for u_{ij} and v_{ij}:

$$(\mathbf{X}\mathbf{V})_{ij}u_{ij} - (\mathbf{U}\mathbf{V}^T\mathbf{V})_{ij}u_{ij} = 0 \tag{6}$$

$$(\mathbf{X}^T\mathbf{U})_{ij}v_{ij} - (\mathbf{V}\mathbf{U}^T\mathbf{U})_{ij}v_{ij} = 0 \tag{7}$$

These equations lead to the following updating formulas:

$$u_{ij} \leftarrow u_{ij}\frac{(\mathbf{X}\mathbf{V})_{ij}}{(\mathbf{U}\mathbf{V}^T\mathbf{V})_{ij}} \tag{8}$$

$$v_{ij} \leftarrow v_{ij}\frac{(\mathbf{X}^T\mathbf{U})_{ij}}{(\mathbf{V}\mathbf{U}^T\mathbf{U})_{ij}} \tag{9}$$

It is proven by Lee [8] that the objective function J is non-increasing under the above iterative updating rules, and that the convergence of the iteration is guaranteed. Note that the solution to minimizing the criterion function J is not unique. If $\mathbf{U}$ and $\mathbf{V}$ are the solution to J, then, $\mathbf{UD}$, $\mathbf{VD}^{-1}$ will also form a solution for any positive diagonal matrix $\mathbf{D}$. To make the solution unique, we further require that the Euclidean length of the column vector in matrix $\mathbf{U}$ is one.

This requirement of normalizing $\mathbf{U}$ can be achieved by[1]:

$$v_{ij} \leftarrow v_{ij}\sqrt{\sum_i u_{ij}^2} \quad (10)$$

$$u_{ij} \leftarrow \frac{u_{ij}}{\sqrt{\sum_i u_{ij}^2}} \quad (11)$$

There is an analogy with the SVD in interpreting the meaning of the two non-negative matrices $\mathbf{U}$ and $\mathbf{V}$. Each element u_{ij} of matrix $\mathbf{U}$ represents the degree to which term $f_i \in \mathcal{W}$ belongs to cluster j, while each element v_{ij} of matrix $\mathbf{V}$ indicates to which degree document i is associated with cluster j. If document i solely belongs to cluster x, then v_{ix} will take on a large value while rest of the elements in i'th row vector of $\mathbf{V}$ will take on a small value close to zero.

In summary, our document clustering algorithm is composed of the following steps:

1. Given a document corpus, construct the term-document matrix $\mathbf{X}$ in which column i represents the weighted term-frequency vector of document d_i.

2. Perform the NMF on $\mathbf{X}$ to obtain the two non-negative matrices $\mathbf{U}$ and $\mathbf{V}$ using Eq.(8) and Eq.(9).

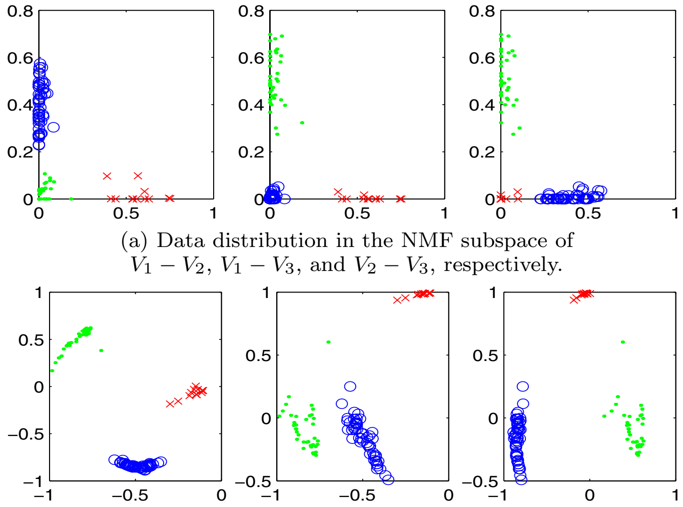
3. Normalize $\mathbf{U}$ and $\mathbf{V}$ using Eq.(11) and Eq.(10).

4. Use matrix $\mathbf{V}$ to determine the cluster label of each data point. More precisely, examine each row i of matrix $\mathbf{V}$. Assign document d_i to cluster x if $x = \arg\max_j v_{ij}$.

The computation complexity for Eq.(11) and Eq.(10) is $O(kn)$ and the total computation time is $O(tkn)$, where is t is number of iterations performed.

3.3 NMF VS. SVD

At the beginning of Section 3, we discussed the characteristics of the NMF and its differences with the methods based on the SVD and eigenvector computations. Here we further illustrate these differences using experiments. We have applied both the NMF and the SVD to a data set that consists of three clusters, and plotted the data set in the spaces derived from the NMF and the SVD, respectively. Figure 2(a) and (b) show the data distributions in the two spaces in which data points belonging to the same cluster are depicted by the same symbol. The three figures in (a) plot the data points in the space of V_1–V_2, V_1–V_3, and V_2–V_3, respectively, where V_1, V_2, V_3 are the three row vectors of $\mathbf{V}$ from the NMF, while the three figures in (b) plot the data points in the space of E_1–E_2, E_1–E_3, and E_2–E_3, respectively, where E_1, E_2, E_3 are the first three singular vectors of the SVD. Clearly, in the NMF space, every document takes non-negative values in all three directions, while in the SVD space, each document may take negative values in some of the directions. Furthermore, in the NMF space, each axis corresponds to a cluster, and all the data points belonging to the same cluster spread along the same axis. Determining the cluster label for each data point is as simple as finding the axis with which the data point has the largest projection value. However, in the SVD space, although the three axes do separate the data points belonging to the different clusters, there is no direct relationship between the axes (eigenvectors) and the clusters. Traditional data clustering methods such as K-means have to be applied in this eigenvector space to find the final set of data clusters.



(a) Data distribution in the NMF subspace of $V_1 - V_2$, $V_1 - V_3$, and $V_2 - V_3$, respectively.

(b) Data distribution in the SVD subspace of $E_1 - E_2$, $E_1 - E_3$, and $E_2 - E_3$, respectively.

Figure 2: Data distribution in the NMF and LSI spaces. Documents belonging to the same cluster are depicted by the same symbol.

4. PERFORMANCE EVALUATIONS

In this section, we describe the document corpora used for the performance evaluations, unveil the document clustering accuracies of the proposed method and its variations, and compare the results with the representative spectral clustering methods.

4.1 Data Corpora

We conducted the performance evaluations using the TDT2[2] and the Reuters[3] document corpora. These two document corpora have been among the ideal test sets for document clustering purposes because documents in the corpora have been manually clustered based on their topics and each document has been assigned one or more labels indicating which topic/topics it belongs to. The TDT2 corpus consists of 100 document clusters, each of which reports a major news event occurred in 1998. It contains a total of 64527 documents from six news agencies such as ABC, CNN, VOA, NYT, PRI and APW, among which 7803 documents have a unique category label. The number of documents for different news events is very unbalanced, ranging from 1 to 1485. In our experiments, we excluded those events with less than 5 documents, which left us with a total of 56 events. The final test set is still very unbalanced, with some large clusters more than 100 times larger than some small ones.

[1] When normalizing matrix $\mathbf{U}$, matrix $\mathbf{V}$ needs to be adjusted accordingly so that $\mathbf{U}\mathbf{V}^T$ does not change.

[2] Nist topic detection and tracking corpus at http://www.nist.gov/speech/tests/tdt/tdt98/index.htm

[3] Reuters-21578, distribution 1.0 corpus at http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

Table 1: Statistics of TDT2 and Reuters corpora.

	TDT2	Reuters
No. documents	64527	21578
No. docs. used	7803	9494
No. clusters	100	135
No. clusters used	56	51
Max. cluster size	1485	3945
Min. cluster size	1	5
Med. cluster size	48	30
Avg. cluster size	137	186

On the other hand, Reuters corpus contains 21578 documents which are grouped into 135 clusters. Compared with TDT2 corpus, the Reuters corpus is more difficult for clustering. In TDT2, each document has a unique category label, and the content of each cluster is narrowly defined, whereas in Reuters, many documents have multiple category labels, and documents in each cluster have a broader variety of content. In our test, we discarded documents with multiple category labels, and removed the clusters with less than 5 documents. This has lead to a data set that consists of 51 clusters with a total of 9494 documents. Table 1 provides the statistics of the two document corpora.

4.2 Evaluation Metrics

The testing data used for evaluating the proposed document clustering method are formed by mixing documents from multiple clusters randomly selected from the document corpus. At each run of the test, documents from a selected number k of topics are mixed, and the mixed document set, along with the cluster number k, are provided to the clustering process. The result is evaluated by comparing the cluster label of each document with its label provided by the document corpus. Two metrics, the accuracy (AC) and the normalized mutual information metric $\widehat{\text{MI}}$ are used to measure the document clustering performance. Given a document d_i, let l_i and α_i be the cluster label and the label provided by the document corpus, respectively. The AC is defined as follows:

$$\text{AC} = \frac{\sum_{i=1}^{n} \delta(\alpha_i, \text{map}(l_i))}{n} \tag{12}$$

where n denotes the total number of documents in the test, $\delta(x, y)$ is the delta function that equals one if $x = y$ and equals zero otherwise, and $\text{map}(l_i)$ is the mapping function that maps each cluster label l_i to the equivalent label from the document corpus. The best mapping can be found by using the Kuhn-Munkres algorithm [10].

On the other hand, given the two sets of document clusters $\mathcal{C}$, $\mathcal{C}'$, their mutual information metric $\text{MI}(\mathcal{C}, \mathcal{C}')$ is defined as:

$$\text{MI}(\mathcal{C}, \mathcal{C}') = \sum_{c_i \in \mathcal{C}, c_j' \in \mathcal{C}'} p(c_i, c_j') \cdot \log_2 \frac{p(c_i, c_j')}{p(c_i) \cdot p(c_j')} \tag{13}$$

where $p(c_i)$, $p(c_j')$ denote the probabilities that a document arbitrarily selected from the corpus belongs to the clusters c_i and c_j', respectively, and $p(c_i, c_j')$ denotes the joint probability that this arbitrarily selected document belongs to the clusters c_i as well as c_j' at the same time. $\text{MI}(\mathcal{C}, \mathcal{C}')$ takes values between zero and $\max(\text{H}(\mathcal{C}), \text{H}(\mathcal{C}'))$, where $\text{H}(\mathcal{C})$ and $\text{H}(\mathcal{C}')$ are the entropies of $\mathcal{C}$ and $\mathcal{C}'$, respectively. It reaches the maximum $\max(\text{H}(\mathcal{C}), \text{H}(\mathcal{C}'))$ when the two sets of document clusters are identical, whereas it becomes zero when the two sets are completely independent. Another important character of $\text{MI}(\mathcal{C}, \mathcal{C}')$ is that, for each $c_i \in \mathcal{C}$, it does not need to find the corresponding counterpart in $\mathcal{C}'$, and the value keeps the same for all kinds of permutations. To simplify comparisons between different pairs of cluster sets, instead of using $\text{MI}(\mathcal{C}, \mathcal{C}')$, we use the following normalized metric $\widehat{\text{MI}}(\mathcal{C}, \mathcal{C}')$ which takes values between zero and one:

$$\widehat{\text{MI}}(\mathcal{C}, \mathcal{C}') = \frac{\text{MI}(\mathcal{C}, \mathcal{C}')}{\max(\text{H}(\mathcal{C}), \text{H}(\mathcal{C}'))} \tag{14}$$

4.3 Performance Evaluations and Comparisons

To demonstrate how our method improves the document clustering accuracy in comparison to the best contemporary methods, we implemented two representative spectral clustering methods: Average Association (AA in short) [13], and Normalized Cut (NC in short) [11], and conducted the performance evaluations on the two methods using the same data corpora. These methods model the given document set using a undirected graph in which each node represents a document, and each edge (i, j) is assigned a weight w_{ij} to reflect the similarity between documents i and j. The document clustering task is accomplished by finding the graph's best cuts that optimize certain predefined criterion functions. Let $G = G(V, E)$ be a weighted graph with the node set V and edge set E, $\mathbf{W} = [w_{ij}]$ be the graph weight matrix, A and B be two subgraphs of G. The criterion functions adopted by the AA and NC methods are defined as:

$$AA = \frac{cut(A, A)}{|A|} + \frac{cut(B, B)}{|B|} \tag{15}$$

$$NC = \frac{cut(A, B)}{cut(A, V)} + \frac{cut(A, B)}{cut(B, V)} \tag{16}$$

where $|A|$ is the size of A and

$$cut(A, B) = \sum_{i \in A, j \in B} w_{ij}. \tag{17}$$

It has been proven that with certain relaxation, the minimization of the above two criterion functions can be approximated by solving the following eigenvalue systems:

$$\begin{array}{ll} \text{AA:} & \mathbf{W}Y = \lambda Y \\ \text{NC:} & \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}Y = \lambda Y \end{array} \tag{18}$$

where $\mathbf{D} = \text{diag}(\mathbf{W}\mathbf{e})$, $\mathbf{e} = [1, 1, \ldots, 1]^T$, and Y is the cluster indication vector that can be used to determine the cluster label of each document.

Interestingly, Zha et al has shown that the AA criterion function (Eq.(15)) is equivalent to that of the LSI followed by the K-means clustering method [13] if the inner product $\langle X_i, X_j \rangle$ is used to measure the document similarity. We can prove that when the weight $1/\sqrt{d_{ii}}$ (the i'th diagonal element of the matrix $\mathbf{D}$) is applied to column vector i of the matrix $\mathbf{W}$, the AA method becomes exactly the same as the NC method. In other words, the essential difference between the AA and the NC methods is that NC applies the weights to $\mathbf{W}$ while AA does not. (See Appendix A).

Table 2: Performance comparisons using TDT2 corpus

	Mutual Information				Accuracy			
k	AA	NC	NMF	NMF-NCW	AA	NC	NMF	NMF-NCW
2	0.834	0.954	0.854	0.972	0.934	0.990	0.946	0.993
3	0.754	0.890	0.790	0.931	0.863	0.951	0.899	0.981
4	0.743	0.846	0.786	0.909	0.830	0.918	0.866	0.953
5	0.696	0.802	0.740	0.874	0.758	0.857	0.812	0.925
6	0.663	0.761	0.701	0.823	0.712	0.802	0.773	0.880
7	0.679	0.756	0.704	0.816	0.707	0.783	0.750	0.857
8	0.624	0.695	0.651	0.782	0.641	0.717	0.697	0.824
9	0.663	0.741	0.683	0.804	0.664	0.754	0.708	0.837
10	0.656	0.736	0.681	0.812	0.638	0.729	0.685	0.835
average	0.701	0.798	0.732	0.858	0.750	0.833	0.793	0.898

Table 3: Performance comparisons using Reuters corpus

	Mutual Information				Accuracy			
k	AA	NC	NMF	NMF-NCW	AA	NC	NMF	NMF-NCW
2	0.399	0.484	0.437	0.494	0.784	0.821	0.824	0.837
3	0.482	0.536	0.489	0.574	0.709	0.765	0.731	0.803
4	0.480	0.581	0.487	0.604	0.629	0.734	0.655	0.758
5	0.565	0.590	0.587	0.600	0.655	0.695	0.686	0.722
6	0.537	0.627	0.559	0.650	0.611	0.678	0.650	0.728
7	0.560	0.599	0.575	0.624	0.584	0.654	0.624	0.696
8	0.559	0.592	0.578	0.606	0.581	0.613	0.618	0.651
9	0.603	0.633	0.614	0.659	0.599	0.640	0.634	0.692
10	0.607	0.647	0.626	0.661	0.600	0.634	0.634	0.677
average	0.532	0.588	0.550	0.608	0.639	0.693	0.673	0.729

Inspired by the above observations, we have conducted performance evaluations on the proposed method with its standard form as well as the NC weighted variation:

Standard form (NMF in short): Conduct document clustering using the original data matrix $\mathbf{X}$.

NC weighted form (NMF-NCW in short): Calculate $\mathbf{D} = \text{diag}(\mathbf{X}^T\mathbf{X}\mathbf{e})$, conduct document clustering using the weighted data matrix $\mathbf{X}' = \mathbf{X}\mathbf{D}^{-1/2}$ (See Appendix B).

Table 2 and 3 show the evaluation results using the TDT2 and the Reuters corpus, respectively. The evaluations were conducted for the cluster numbers ranging from two to ten. For each given cluster number k, 50 test runs were conducted on different randomly chosen clusters, and the final performance scores were obtained by averaging the scores from the 50 tests. Because NMF algorithm is not guaranteed to find the global optimum, it is beneficial to perform NMF algorithm a few times with different initial values and choose the trial with minimal square error J. In reality, if the data-set has reasonable clusters, usually a very few number of trials is enough to find a satisfactory solution. In all of our experiments, 10 trials of NMF are performed in each test run.

Our finding can be summarized as follows: regardless of the document corpora, the performance ranking is always in the order of AA, NMF, NC, and NMF-NCW. Applying the NC weighting always brings positive effects for both the spectral clustering (NC vs. AA) and the NMF methods (NMF-NCW vs. NMF). The improvement becomes more obvious for the TDT2 corpus than the Reuters corpus. As described in Section 4.1, document clusters in TDT2 are generally more compact and focused than the clusters in Reuters. The above experimental results for the two document corpora are mostly in line with the expectations because document clustering methods generally produce better results for document corpora comprised of compact and well-focused clusters.

5. SUMMARY

In this paper, we have presented a novel document partitioning method based on the non-negative factorization of the term-document matrix of the given document corpus. Our method differs from the latent semantic indexing method based on the singular vector decomposition (SVD) and the related spectral clustering methods in that the latent semantic space derived by NMF does not need to be orthogonal, and that each document is guaranteed to take only non-negative values in all the latent semantic directions. As evidenced by the experiment in Section 3.3, these two differences bring about an important benefit that each axis in the space derived by the NMF has a much more straightforward correspondence with each document cluster than in the space derived by the SVD, and thereby document clustering results can be directly derived without additional clustering operations. Our experimental evaluations show that the proposed document clustering method surpasses SVD- and the eigen decomposition clustering methods not only in the easy and reliable derivation of document clustering results, but also in document clustering accuracies.

6. REFERENCES

[1] L. Baker and A. McCallum. Distributional clustering of words for text classification. In *Proceedings of ACM SIGIR*, 1998.

[2] P. K. Chan, D. F. Schlag, and J. Y. Zien. Spectral k-way ratio-cut partitioning an clustering. *IEEE Trans. Computer-Aided Design*, 13:1088–1096, Sep. 1994.

[3] D. Cutting, D. Karger, J. Pederson, and J. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of ACM SIGIR*, 1992.

[4] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[5] C. Ding, X. He, H. Zha, M. Gu, and H. D. Simon. A min-max cut algorithm for graph partitioning and data clustering. In *Proceedings of IEEE ICDM 2001*, pages 107–114, 2001.

[6] P. O. Hoyer. Non-negative sparse coding. In *Proc. IEEE Workshop on Neural Networks for Signal Processing*, Martigny, Switzerland, 2002.

[7] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[8] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, volume 13, pages 556–562, 2001.

[9] X. Liu and Y. Gong. Document clustering with cluster refinement and model selection capabilities. In *Proceedings of ACM SIGIR 2002*, Tampere, Finland, Aug. 2002.

[10] L. Lovasz and M. Plummer. *Matching Theory.* Akadémiai Kiadó, North Holland, Budapest, 1986.

[11] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[12] P. Willett. Document clustering using an inverted file approach. *Journal of Information Science*, 2:223–231, 1990.

[13] H. Zha, C. Ding, M. Gu, X. He, and H. Simon. Spectral relaxation for k-means clustering. In *Advances in Neural Information Processing Systems*, volume 14, 2002.

APPENDIX

A. NORMALIZED CUT CRITERION VS. WEIGHTED K-MEANS ALGORITHM

In this appendix, we prove that the weighted K-means criterion is the same as that of the normalized-cut criterion. Let each data point has weight γ_i, the weighted sum of the squared error of each data point to its corresponding cluster center is:

$$\begin{aligned} J &= \sum_k \sum_{i \in C_k} \gamma_i \|X_i - \mu_k\|^2 \\ &= \sum_k \sum_{i \in C_k} \gamma_i X_i^T X_i - \sum_k \left(\sum_{i \in C_k} \gamma_i \right) \mu_k^T \mu_k \\ &= \sum_i \gamma_i X_i^T X_i - \sum_k \left(\sum_{i \in C_k} \gamma_i \right) \mu_k^T \mu_k \end{aligned} \tag{19}$$

where μ_k is the center of k-th cluster,

$$\mu_k = \frac{\sum_{i \in C_k} \gamma_i X_i}{\sum_{i \in C_k} \gamma_i} \tag{20}$$

Let S_k be the indicator vector of cluster k, i.e., the i-th element of S_k is equal to 1 of $i \in C_k$ and equal to 0 otherwise, and $\mathbf{\Gamma}$ be the diagonal matrix consists of γ_i, then we have the following identities:

$$\sum_{i \in C_k} \gamma_i = S_k^T \mathbf{\Gamma} S_k \tag{21}$$

$$\sum_{i \in C_k} \gamma_i X_i = S_k^T \mathbf{\Gamma} \mathbf{X} \tag{22}$$

So J can be re-written as:

$$\begin{aligned} J &= \sum_i \gamma_i X_i^T X_i - \sum_k \frac{S_k^T \mathbf{\Gamma} \mathbf{X}^T \mathbf{X} \mathbf{\Gamma} S_k}{S_k^T \mathbf{\Gamma} S_k} \\ &= \sum_i \gamma_i X_i^T X_i - \sum_k Y_k^T \mathbf{\Gamma}^{1/2} \mathbf{W} \mathbf{\Gamma}^{1/2} Y_k \end{aligned} \tag{23}$$

where $\mathbf{W} = \mathbf{X}^T \mathbf{X}$ and $Y_k = \frac{\mathbf{\Gamma}^{1/2} S_k}{\|\mathbf{\Gamma}^{1/2} S_k\|}$

Noting that the first term of Eq.(23) does not depend on the partition, so minimizing Eq.(23) is equivalent to maximizing the second term of Eq.(23). If we allow the optimal solution S_k to be any real values instead of insisting them to be binary, using the fact $\mathbf{Y}^T \mathbf{Y} = \mathbf{I}$, then Y can be found by solving the following eigenproblem:

$$\mathbf{\Gamma}^{1/2} \mathbf{W} \mathbf{\Gamma}^{1/2} Y = \lambda Y \tag{24}$$

Now, if we let each weight $\gamma_i = 1/d_{ii}$, then above eigenproblem is exactly same as the eigenproblem of normalized-cut criterion.

B. WEIGHTED NMF

In this appendix, we derive the mathematical form of the NMF weighted by the normalized cut weighting scheme. Following the notation in Appendix A, the weighted sum of the squared error is:

$$\begin{aligned} J &= \frac{1}{2} \sum_i \gamma_i (X_i - \mathbf{U} V_i^T)^T (X_i - \mathbf{U} V_i^T) \\ &= \frac{1}{2} \sum_i (\gamma_i^{1/2} X_i - \gamma_i^{1/2} \mathbf{U} V_i^T)^T (\gamma_i^{1/2} X_i - \gamma_i^{1/2} \mathbf{U} V_i^T) \\ &= \frac{1}{2} \mathrm{tr}((\mathbf{X}' - \mathbf{U}\mathbf{V}'^T)(\mathbf{X}' - \mathbf{U}\mathbf{V}'^T))^T \end{aligned}$$

where $\mathbf{X}' = \mathbf{X}\mathbf{\Gamma}^{\mathbf{1/2}}$ and $\mathbf{V}' = \mathbf{\Gamma}^{\mathbf{1/2}}\mathbf{V}$

The above equation has the same form as Eq.(2) in Section 3.2, so the same algorithm can be used to find the solution.