

# Incorporating User provided Constraints into Document Clustering

Yanhua Chen    Manjeet Rege    Ming Dong  
Machine Vision and Pattern Recognition Lab

Jing Hua  
Graphics and Imaging Lab

Department of Computer Science, Wayne State University  
Detroit, MI 48202, USA

{chenyanh, rege, mdong, jinghua}@wayne.edu

## Abstract

*Document clustering without any prior knowledge or background information is a challenging problem. In this paper, we propose SS-NMF: a semi-supervised non-negative matrix factorization framework for document clustering. In SS-NMF, users are able to provide supervision for document clustering in terms of pairwise constraints on a few documents specifying whether they “must” or “cannot” be clustered together. Through an iterative algorithm, we perform symmetric tri-factorization of the document-document similarity matrix to infer the document clusters. Theoretically, we show that SS-NMF provides a general framework for semi-supervised clustering and that existing approaches can be considered as special cases of SS-NMF. Through extensive experiments conducted on publicly available data sets, we demonstrate the superior performance of SS-NMF for clustering documents.*

## 1. Introduction

Document clustering is the grouping of text documents into meaningful clusters in an unsupervised manner. It is one of the most important tasks in text mining and has received extensive attention in the data mining community recently [6, 19, 38].

Information retrieval (IR) needs range from a specific search at one end to an open ended browsing of the database at the other [8]. A keyword-based search, where the user is interested in retrieving all the documents that have an exact match with the query keyword, is an example of a specific search scenario. On the other hand in open-ended browsing, the user generally has a broader perspective of the information he/she is looking for and is interested in browsing and navigating through the database. While traditional IR techniques have been well developed for the specific search scenario, they are ill-suited for providing a browsing capability to the user. A good document clustering algorithm can provide a holistic view of the text corpus and hence overcome the limitations of traditional IR techniques.

Document clustering methods in general can be categorized into document partitioning (flat clustering) and agglomerative (hierarchical) clustering. Partitioning methods typically divide the documents in a given number of clusters directly. Hierarchical clustering aims to obtain a hierarchy of clusters by building a tree structure, that shows how the clusters are related to each other. The clustering result of the documents can be obtained by cutting the tree at a desired level. One of the popular hierarchical document clustering methods is the hierarchical agglomerative clustering (HAC) that proceeds in a bottom-up fashion by iteratively merging small clusters into larger ones [7, 35]. This is continued until all the documents get merged into one single cluster at the root node of the tree. Variations of HAC algorithm have been proposed that differ based on the method adopted to compute the similarities between the clusters. Some of the common methods to measure cluster similarity are single-linkage, complete-linkage, and group-average linkage. While, the first two use the maximum and minimum distance between the clusters, respectively, group-average linkage uses the cluster center distance. [14] has studied the different types of similarity measures and their effect on clustering accuracy.

Some of the widely applied methods in document partitioning include  $k$ -means [12], probabilistic clustering using the Naive Bayes or Gaussian mixture model [1, 28], etc.  $k$ -means produces clusters that minimizes the sum of squared distances between the data points and their corresponding cluster centers. On the other hand, both naive Bayes and Gaussian mixture model define a probabilistic cluster model and try to find the model by maximizing the likelihood of the data. The problems associated with these methods is that they make a strict assumption on the distribution of the document corpus.  $k$ -means assumes every document cluster has a compact shape, the Naive Bayes model assumes feature independence in the document corpus feature space, and the Gaussian mixture model assumes that the density of each cluster can be approximated by a Gaussian distribution. Since, the actual underlying distribution of the docu-

ment corpus can be different, these methods are susceptible to their *a priori* assumptions.

Recently, document clustering based on spectral clustering has emerged as a popular approach [9, 11]. These methods model the documents as vertices of a weighted graph with edge weights representing the similarity between two documents. Clustering is then obtained by “cutting” the graph vertices into different partitions. Partitioning of the graph is obtained by solving an eigenvalue problem where the clustering is inferred from the top eigenvectors. As can be seen from the above discussion, document clustering has been extensively studied and various methods proposed. However, accurately clustering documents without domain-dependent background information, is still a challenging task.

In this paper, we propose a non-negative matrix factorization (NMF) [23, 24] based framework to incorporate prior knowledge into document clustering. Under the proposed semi-supervised NMF (SS-NMF) methodology, user is able to provide pairwise constraints on a few documents specifying whether they “must” or “cannot” be clustered together. We derive an iterative algorithm to perform symmetric non-negative tri-factorization of the document-document similarity matrix. The correctness of the algorithm is proved by showing that the algorithm is guaranteed to converge. We also prove that SS-NMF is a general and unified framework for semi-supervised clustering by establishing the relationship between SS-NMF and other existing semi-supervised clustering algorithms. Experiments performed on publicly available text data sets demonstrate the effectiveness of the proposed work.

## 2. Related Work

There have been prior efforts on using user provided information to improve clustering. [17] proposed incorporating background knowledge into document clustering by enriching the text features using WordNet<sup>1</sup>. In [21], some words per class and a class hierarchy were sought from the user in order to generate labels and build an initial text classifier for the class. A similar technique was proposed in [27], where the user is made to select interesting words from automatically selected representative words for each class of documents. These user identified words were then used to re-train the text classifier. Active learning approaches have also found application in semi-supervised clustering. [13] has proposed to convert a user recommended feature into a mini-document which is then used to train an SVM classifier. This approach has been extended by [31] which adjusts SVM weights of the key features to a predefined value in binary classification tasks. Recently, [18] presented a probabilistic generative model to incorporate extended

<sup>1</sup><http://wordnet.princeton.edu>

feedback that allows the user and the algorithm to jointly arrive at coherent clusters that capture the categories of interest to the user. [5, 20, 30] proposed methods where the user provided class labels *a priori* to some of the documents. These algorithms use the labeled data to generate seed clusters that initialize a clustering algorithm, and use constraints generated from the labeled data to guide the clustering process. Proper seeding biases clustering towards a good region of the search space, while simultaneously producing a clustering similar to the specified labels.

However, in certain applications, supervision in the form of class labels may be unavailable. For example, complete class labels may be unknown in the context of clustering for speaker identification in a conversation [2], or clustering GPS data for lane-finding [34]. In some domains, pairwise constraints occur naturally, e.g., the Database of Interacting Proteins (DIP) data set contains information about proteins co-occurring in processes, which can be viewed as *must-link* constraints during clustering. Similarly, for document clustering, user knowledge about which few documents are related or unrelated can be incorporated to improve the clustering results. Moreover, it is easier for a user who is not a domain expert to provide feedback in the form of pairwise constraints than class labels, since providing constraints does not require the user to have significant prior knowledge about the categories in the data set. Amongst the various methods proposed for utilizing user provided constraints for semi-supervised clustering [3, 4], two of the well-known include the semi-supervised kernel  $k$ -means (SS-KK) [22] and semi-supervised spectral clustering with normalized cuts (SS-SNC) [19]. While, SS-KK transforms the clustering distance measure by weighted kernel  $k$ -means with reward and penalty constraints to perform semi-supervised clustering of data given either as vectors or as a graph, SS-SNC utilizes supervision to change the clustering distance measure with pairwise information by spectral methods. The SS-NMF framework presented in this paper, allows the user to provide pairwise constraints on a small percentage of the documents. Specifically, these constraints specify whether two documents should belong to the same cluster or should strictly belong to different clusters.

## 3. Semi-supervised Non-negative Matrix Factorization for Document Clustering

### 3.1. Model Formulation

The entire document collection is typically represented using the vector space model [32] by a word-document matrix  $\mathbf{X} \in R^{m \times n}$  where columns index the documents and rows denote the words appearing in them. The documents are treated as vectors with words as their features such that an entry  $x_{fi}$  in the matrix signifies the relevance of word

$f$  for document  $\mathbf{d}_i$ , usually by the frequency of the word appearing in the document.

We propose a semi-supervised NMF (SS-NMF) model for document clustering. NMF has received much attention recently and proved to be very useful for applications such as pattern recognition, text mining, multimedia, and DNA gene expressions. It was initially proposed for “parts-of-whole” decomposition [23, 24], and later extended to a general framework for data clustering [10]. It can model widely varying data distributions and accomplish both hard and soft clustering simultaneously. When applied to the word-document matrix  $\mathbf{X}$ , NMF factorizes  $\mathbf{X}$  into two non-negative matrices [36],

$$\mathbf{X} \approx \mathbf{PQ}^T \quad (1)$$

where  $\mathbf{P} \in R^{m \times k}$  is cluster centroid,  $\mathbf{Q} \in R^{n \times k}$  is cluster indicator, and  $k$  is the number of clusters.

In the proposed model, we perform symmetric non-negative tri-factorization of the document-document similarity matrix  $\mathbf{A} = \mathbf{X}^T \mathbf{X} \in R^{n \times n}$  as,

$$\mathbf{A} \approx \mathbf{GSG}^T \quad (2)$$

where  $\mathbf{G} \in R^{n \times k}$  is the cluster indicator matrix. An entry  $g_{ih}$  in  $\mathbf{G}$  gives the degree of association of document  $\mathbf{d}_i$  with cluster  $h$ . The cluster membership of a document is given by finding the cluster with the maximum association value.  $\mathbf{S} \in R^{k \times k}$  is the cluster centroid matrix that gives a compact  $k \times k$  representation of  $\mathbf{X}$ .

Supervision is provided as two sets of pairwise constraints on the documents: *must-link* constraints  $C_{ML}$  and *cannot-link* constraints  $C_{CL}$ . Every pair of documents,  $(\mathbf{d}_i, \mathbf{d}_j) \in C_{ML}$  implies that  $\mathbf{d}_i$  and  $\mathbf{d}_j$  must belong to the same cluster. Similarly, all possible pairs  $(\mathbf{d}_i, \mathbf{d}_j) \in C_{CL}$  implies that the two documents should belong to different clusters. The constraints are accompanied by associated violation cost matrix  $\mathbf{W}$ . An entry  $w_{ij}$  in this matrix denotes the cost of violating the constraint between documents  $\mathbf{d}_i$  and  $\mathbf{d}_j$ , if such a constraint exists, that is, either  $(\mathbf{d}_i, \mathbf{d}_j) \in C_{ML}$  or  $(\mathbf{d}_i, \mathbf{d}_j) \in C_{CL}$ . The model relies on a distortion measure  $D : R^m \rightarrow R$ , to compute distance between documents. Assuming the text corpus consists of  $k$  semantic concepts, the goal is to partition the set of documents into  $k$  disjoint clusters  $\{\mathbf{X}_h\}_{h=1}^k$ , such that the total distortion between the documents and the corresponding cluster representatives is (locally) minimized according to the given distortion measure  $D$ , while constraint violations are kept to a minimum.

### 3.2. Algorithm Derivation

We define the objective function of SS-NMF as follows:

$$J_{SS-NMF} = \min_{\mathbf{S} \geq 0, \mathbf{G} \geq 0} \|\tilde{\mathbf{A}} - \mathbf{GSG}^T\|^2 \quad (3)$$

where  $\tilde{\mathbf{A}} = \mathbf{A} - \mathbf{W}_{reward} + \mathbf{W}_{penalty}$  is affinity or similarity matrix  $\mathbf{A}$  with constraints  $\mathbf{W}_{reward} = \{w_{ij} | (\mathbf{d}_i, \mathbf{d}_j) \in$

$C_{ML}, s.t. y_i = y_j\}$  and  $\mathbf{W}_{penalty} = \{w_{ij} | (\mathbf{d}_i, \mathbf{d}_j) \in C_{CL}, s.t. y_i = y_j\}$ ,  $w_{ij}$  is the penalty cost for violating a constraint between documents  $\mathbf{d}_i$  and  $\mathbf{d}_j$ , and  $y_i$  is the cluster label of  $\mathbf{d}_i$ .  $\mathbf{S} \in R^{k \times k}$  is the cluster centroid, and  $\mathbf{G} \in R^{n \times k}$  is the cluster indicator. Equation (3) can be re-written as:

$$J_{SS-NMF} = \min_{\mathbf{S} \geq 0, \mathbf{G} \geq 0} \|(\mathbf{A} - \mathbf{W}_{reward} + \mathbf{W}_{penalty}) - \mathbf{GSG}^T\|^2 \quad (4)$$

We propose an iterative procedure for the minimization of equation (3) where we update one factor while fixing the others. The updating rules are,

$$\mathbf{S}_{ih} \leftarrow \mathbf{S}_{ih} \frac{(\mathbf{G}^T \tilde{\mathbf{A}} \mathbf{G})_{ih}}{(\mathbf{G}^T \mathbf{GSG}^T \mathbf{G})_{ih}} \quad (5)$$

$$\mathbf{G}_{ih} \leftarrow \mathbf{G}_{ih} \frac{(\tilde{\mathbf{A}} \mathbf{G})_{ih}}{(\mathbf{GSG}^T \mathbf{G})_{ih}} \quad (6)$$

Thus, the SS-NMF algorithm for document clustering can be illustrated in Algorithm 1.

---

#### Algorithm 1 SS-NMF Algorithm

---

**INPUT:** Document-document similarity matrix  $\mathbf{A}$ , number of clusters  $k$ , constraint penalty matrix  $\mathbf{W}_{penalty}$ , and constraint reward matrix  $\mathbf{W}_{reward}$

**OUTPUT:** Clusters  $\{\mathbf{X}_h\}_{h=1}^k$  with  $\mathbf{Y}_h = \{i | \mathbf{d}_i \in \mathbf{X}_h\}$

**METHOD:**

1. Initialize  $\mathbf{S}$  and  $\mathbf{G}$  with non-negative values.
2. Construct  $\tilde{\mathbf{A}} = \mathbf{A} - \mathbf{W}_{reward} + \mathbf{W}_{penalty}$
3. Iterate for each  $i$  and  $h$  until *convergence*

(a) Cluster centroid

$$\mathbf{S}_{ih} \leftarrow \mathbf{S}_{ih} \frac{(\mathbf{G}^T \tilde{\mathbf{A}} \mathbf{G})_{ih}}{(\mathbf{G}^T \mathbf{GSG}^T \mathbf{G})_{ih}}$$

(b) Cluster indicator

$$\mathbf{G}_{ih} \leftarrow \mathbf{G}_{ih} \frac{(\tilde{\mathbf{A}} \mathbf{G})_{ih}}{(\mathbf{GSG}^T \mathbf{G})_{ih}}$$


---

### 3.3. Algorithm correctness and convergence

We now prove the theoretical correctness and convergence of SS-NMF. Motivated by [29], we render the proof based on optimization theory, auxiliary function and several matrix inequalities.

#### 3.3.1 Correctness

First, we prove the correctness of algorithm.

- Following the standard theory of constrained optimization, we introduce the Lagrangian multipliers  $\lambda_1$  and  $\lambda_2$  to minimize the lagrangian function,

$$L(\mathbf{S}, \mathbf{G}, \lambda_1, \lambda_2) = \min_{\mathbf{S} \geq 0, \mathbf{G} \geq 0} \|\tilde{\mathbf{A}} - \mathbf{G}\mathbf{S}\mathbf{G}^T\|^2 - \text{Tr}(\lambda_1 \mathbf{S}^T) - \text{Tr}(\lambda_2 \mathbf{G}^T) \quad (7)$$

- Based on the Kuhn-Tucker complementarity condition,

$$\frac{\partial J}{\partial \mathbf{S}} = 0 \quad (8)$$

$$\frac{\partial J}{\partial \mathbf{G}} = 0 \quad (9)$$

$$\lambda_1 \odot \mathbf{S} = 0 \quad (10)$$

$$\lambda_2 \odot \mathbf{G} = 0 \quad (11)$$

where  $\odot$  denotes the Hadamard product of two matrices. Taking the derivatives, we obtain the following two equations from equation (8) and equation (9), respectively.

$$4\mathbf{G}^T \tilde{\mathbf{A}} \mathbf{G} - 4\mathbf{G}^T \mathbf{G} \mathbf{S} \mathbf{G}^T \mathbf{G} + \lambda_1 = 0 \quad (12)$$

$$4\tilde{\mathbf{A}} \mathbf{G} \mathbf{S} - 4\mathbf{G} \mathbf{S} \mathbf{G}^T \mathbf{G} \mathbf{S} + \lambda_2 = 0 \quad (13)$$

- Applying the Hadamard multiplication on both sides of equation (12) and equation (13) by  $\mathbf{S}$  and  $\mathbf{G}$ , respectively, and using conditions of equation (10) and equation (11), we can prove that : if  $\mathbf{S}$  and  $\mathbf{G}$  are a local minimizer of the objective function in equation (7), then the following equations are satisfied,

$$(\mathbf{G}^T \tilde{\mathbf{A}} \mathbf{G}) \odot \mathbf{S} - (\mathbf{G}^T \mathbf{G} \mathbf{S} \mathbf{G}^T \mathbf{G}) \odot \mathbf{S} = 0 \quad (14)$$

$$(\tilde{\mathbf{A}} \mathbf{G} \mathbf{S}) \odot \mathbf{G} - (\mathbf{G} \mathbf{S} \mathbf{G}^T \mathbf{G} \mathbf{S}) \odot \mathbf{G} = 0 \quad (15)$$

- Based on the above two equations, we derive the proposed updating rules of equation (5) and equation (6).

### 3.3.2 Convergence

Next, we prove the convergence. This can be done by making use of an auxiliary function similar to that used in [23]. Due to space constraints, we give an outline of the proof and omit the details.

- Assuming  $L(\mathbf{S}, \mathbf{S}')$  is an auxiliary function of  $J(\mathbf{S})$  if  $L(\mathbf{S}, \mathbf{S}') \geq J(\mathbf{S})$  and  $L(\mathbf{S}, \mathbf{S}) = J(\mathbf{S})$ , we minimize a lower bound, set  $\mathbf{S}^{(t+1)} = \arg \min_{\mathbf{S}} L(\mathbf{S}, \mathbf{S}^{(t)})$ , then  $J(\mathbf{S}^{(t)}) = L(\mathbf{S}^{(t)}, \mathbf{S}^{(t)}) \geq L(\mathbf{S}^{(t+1)}, \mathbf{S}^{(t)}) \geq J(\mathbf{S}^{(t+1)})$ . Thus  $J(\mathbf{S})$  is monotonically decreasing and is bounded from up.
- Similarly, assuming  $L(\mathbf{G}, \mathbf{G}')$  is an auxiliary function of  $J(\mathbf{G})$  if  $L(\mathbf{G}, \mathbf{G}') \geq J(\mathbf{G})$  and  $L(\mathbf{G}, \mathbf{G}) = J(\mathbf{G})$ , we minimize a lower bound, set  $\mathbf{G}^{(t+1)} = \arg \min_{\mathbf{G}} L(\mathbf{G}, \mathbf{G}^{(t)})$ , then  $J(\mathbf{G}^{(t)}) = L(\mathbf{G}^{(t)}, \mathbf{G}^{(t)}) \geq L(\mathbf{G}^{(t+1)}, \mathbf{G}^{(t)}) \geq J(\mathbf{G}^{(t+1)})$ . Thus  $J(\mathbf{G})$  is monotonically decreasing and is bounded from up.

### 3.4. Equivalence of SS-NMF and other semi-supervised clustering methods

We now show that SS-NMF is a general and unified framework for semi-supervised clustering by establishing the relationship between SS-NMF and other well-known semi-supervised clustering algorithms, i.e., semi-supervised kernel  $k$ -means (SS-KK) [22] and semi-supervised spectral clustering with normalized cuts (SS-SNC) [19]. In fact, both these algorithms can be considered to be special cases of SS-NMF.

**Proposition 1.** Orthogonal SS-NMF clustering is equivalent to SS-KK clustering.

**Proof.** The SS-NMF objective function is,

$$J_{SS-NMF} = \min_{\mathbf{S} \geq 0, \mathbf{G} \geq 0} \|\tilde{\mathbf{A}} - \mathbf{G}\mathbf{S}\mathbf{G}^T\|^2 \quad (16)$$

The equation can be written as,  $J_{SS-NMF} = \|\tilde{\mathbf{A}} - \mathbf{G}\mathbf{S}\mathbf{G}^T\|^2 = \|\tilde{\mathbf{A}} - \mathbf{G}'\mathbf{G}'^T\|^2 = \text{Tr}(\tilde{\mathbf{A}}^T \tilde{\mathbf{A}} - 2\mathbf{G}'^T \tilde{\mathbf{A}} \mathbf{G}' + \mathbf{G}'^T \mathbf{G}')$  if let  $\mathbf{S} = \mathbf{Q}^T \mathbf{Q}$  and  $\mathbf{G}' = \mathbf{G}\mathbf{Q}^T$ . Since  $\text{Tr}(\tilde{\mathbf{A}}^T \tilde{\mathbf{A}} + \mathbf{G}'^T \mathbf{G}')$  is a constant, the minimization of  $J$  becomes a maximization problem as,

$$\max_{\mathbf{G}' \geq 0} \text{Tr}(\mathbf{G}'^T \tilde{\mathbf{A}} \mathbf{G}') \text{ s.t. } \mathbf{G}'^T \mathbf{G}' = \mathbf{I} \quad (17)$$

The SS-KK objective function is [22],

$$J_{SS-KK} = \sum_{h=1}^k \sum_{i \in \mathbf{X}_h} \|\phi(\mathbf{d}_i) - \overline{\phi}_h\|^2 - \sum_{\substack{(\mathbf{d}_i, \mathbf{d}_j) \in C_{ML}, \\ \text{s.t. } y_i = y_j}} w_{ij} + \sum_{\substack{(\mathbf{d}_i, \mathbf{d}_j) \in C_{CL}, \\ \text{s.t. } y_i = y_j}} w_{ij} \quad (18)$$

where  $\phi(\cdot)$  is the kernel function and  $\overline{\phi}_h$  the centroid. Let  $\mathbf{E}$  be the matrix of pairwise squared Euclidean distances among the data points,  $\mathbf{W}$  the constraint matrix and  $\mathbf{G}$  the cluster indicator. Equation (18) becomes the minimization of the following function,

$$\min_{\mathbf{G} \geq 0} \text{Tr}(\mathbf{G}^T (\mathbf{E} - 2\mathbf{W}) \mathbf{G}) \text{ s.t. } \mathbf{G}^T \mathbf{G} = \mathbf{I} \quad (19)$$

We can convert the minimization of equation (19) to a maximization of the problem,

$$\max_{\mathbf{G} \geq 0} \text{Tr}(\mathbf{G}^T \mathbf{K} \mathbf{G}) \text{ s.t. } \mathbf{G}^T \mathbf{G} = \mathbf{I} \quad (20)$$

where  $\mathbf{K} = \mathbf{A} + \mathbf{W}$  and  $\mathbf{A}$  the similarity matrix.

It is clear that the objective function of SS-NMF (equation (17)) is equivalent to that of SS-KK (equation (20)) if  $\mathbf{K} = \tilde{\mathbf{A}}$ . The  $\mathbf{G}'$  in equation (17) represents the same clustering as  $\mathbf{G}$  of equation (20) does.

**Proposition 2.** Orthogonal SS-NMF clustering is equivalent to SS-SNC clustering.

**Proof.** The objective function of SS-SNC is [19],

$$J_{SS-SNC} = \sum_{h=1}^k \frac{\mathbf{g}_h^T (\tilde{\mathbf{D}} - \tilde{\mathbf{A}}) \mathbf{g}_h}{\mathbf{g}_h^T \tilde{\mathbf{D}} \mathbf{g}_h} = \sum_{h=1}^k \mathbf{z}_h^T (\mathbf{I} - \dot{\mathbf{A}}) \mathbf{z}_h \quad (21)$$

where  $\tilde{\mathbf{A}} = \mathbf{A} - \mathbf{W}_{reward} - \mathbf{W}_{penalty}$  is the pairwise similarity matrix with constraints,  $\tilde{\mathbf{D}} = \text{diag}(\tilde{\mathbf{d}}_1, \dots, \tilde{\mathbf{d}}_n)$  is the diagonal matrix,  $\mathbf{g}_h$  is the cluster indicator, scaled cluster indicator vector  $\mathbf{z}_h = \tilde{\mathbf{D}}^{1/2} \mathbf{g}_h / \|\tilde{\mathbf{D}}^{1/2} \mathbf{g}_h\|$ , and  $\dot{\mathbf{A}} = \mathbf{D}^{-1/2} \tilde{\mathbf{A}} \mathbf{D}^{-1/2}$ .

It can be shown that the minimization of equation (21) becomes a maximization problem as,

$$\max_{\mathbf{Z} \geq 0} \text{Tr}(\mathbf{Z}^T \dot{\mathbf{A}} \mathbf{Z}) \quad s.t. \quad \mathbf{Z}^T \mathbf{Z} = \mathbf{I} \quad (22)$$

Also, it can be seen that equation (17) is equivalent to equation (22) if  $\tilde{\mathbf{A}} = \dot{\mathbf{A}}$ . Moreover, the  $\mathbf{G}'$  in equation (17) represents the same clustering as  $\mathbf{Z}$  of equation (22) does.

From the above two proofs, we can see that the SS-NMF, SS-KK, and SS-SNC are mathematically equivalent. However, notice that in SS-NMF, the matrix  $\tilde{\mathbf{A}}$  might have some negative values, which is not permitted in traditional NMF [23, 24]. In this case, one possible solution is to perform some normalization techniques to guarantee non-negative values. Alternatively, we can simply relax the non-negative constraint to allow negative values as in Semi-NMF [26]. In either of the approaches, the clustering result will not get affected. In SS-NMF, the cluster indicator  $\mathbf{G}'$  is near-orthogonal and can produce soft clustering results. The cluster centroid  $\mathbf{S}$  can provide good characterization of the quality of data clustering because the residue of the matrix approximation  $J = \min \|\tilde{\mathbf{A}} - \mathbf{G} \mathbf{S} \mathbf{G}^T\|$  is smaller than  $J = \min \|\tilde{\mathbf{A}} - \mathbf{G} \mathbf{G}^T\|$ . On the other hand, for SS-KK and SS-SNC, if input matrix is added with constraint weight  $\mathbf{W}$ , in order to ensure positive definiteness, certain additive constraints need to be enforced. Moreover, these constraints are difficult to be relaxed. Also, the cluster indicator  $\mathbf{G}$  or  $\mathbf{Z}$  is required to be orthogonal, leading to only hard clustering results. Hence, both SS-KK and SS-SNC can be viewed as special cases of SS-NMF with orthogonal space constraints. Thus, SS-NMF essentially provides a general and flexible mathematical framework for semi-supervised data clustering.

### 3.5. Advantages of SS-NMF

In this Section, we further illustrate the advantages of SS-NMF using a toy data set shown in Figure 1a, which follows an extreme distribution consisting of 20 data points forming two natural clusters: two circular rings with 10 data points each. Traditional unsupervised clustering methods, such as (kernel)  $k$ -means, spectral normalized cut or NMF,

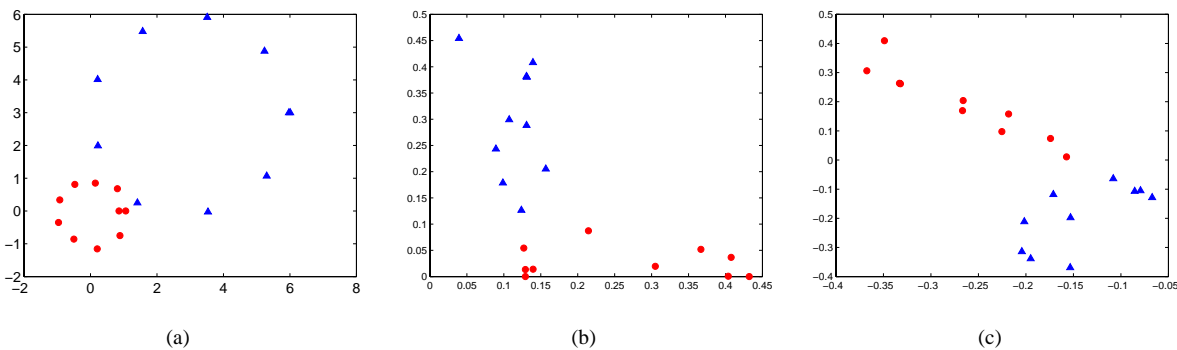
**Table 1.** Cluster indicator  $\mathbf{G}$  of SS-KK and SS-NMF for the toy data set.

$\mathbf{G}$	SS-KK		SS-NMF	
$\mathbf{g}_1$	1	0	0.2778	0.0820
$\mathbf{g}_2$	1	0	0.2977	0.0486
$\mathbf{g}_3$	1	0	0.4301	0.0009
$\mathbf{g}_4$	1	0	0.1295	0.0494
$\mathbf{g}_5$	1	0	0.1377	0.0021
$\mathbf{g}_6$	1	0	0.3845	0.0000
$\mathbf{g}_7$	1	0	0.1281	0.0001
$\mathbf{g}_8$	1	0	0.1426	0.0097
$\mathbf{g}_9$	1	0	0.3119	0.0023
$\mathbf{g}_{10}$	1	0	0.4691	0.0080
$\mathbf{g}_{11}$	0	1	0.0651	0.3959
$\mathbf{g}_{12}$	0	1	0.0599	0.4449
$\mathbf{g}_{13}$	0	1	0.1161	0.4108
$\mathbf{g}_{14}$	0	1	0.0978	0.2985
$\mathbf{g}_{15}$	0	1	0.0592	0.2506
$\mathbf{g}_{16}$	1	0	0.1220	0.1233
$\mathbf{g}_{17}$	0	1	0.1047	0.1735
$\mathbf{g}_{18}$	0	1	0.1503	0.2028
$\mathbf{g}_{19}$	0	1	0.1233	0.2866
$\mathbf{g}_{20}$	0	1	0.1181	0.3800

are unable to produce satisfactory results on this data set. However, after incorporating knowledge from the user in the form of constraints, we are able to achieve much better results.

Unlike SS-SNC, SS-NMF maps the documents into a non-negative latent semantic space. Moreover, SS-NMF does not require the derived space to be orthogonal. Figures 1b and c show the data distributions in the two spaces for SS-NMF and SS-SNC, respectively. Data points belonging to the same cluster are depicted by the same symbol. For SS-NMF, we plot the data points in the space of two column vectors of  $\mathbf{G}$ , while for SS-SNC the first two singular vectors are used. Clearly, in the SS-NMF space, every data point takes non-negative values in both the directions. Furthermore, in SS-NMF space, each axis corresponds to a cluster, and all the data points belonging to the same cluster are nicely spread along the axis. The cluster label for a data point can be determined by finding the axis with which the data point has the largest projection value. However, in the SS-SNC space, there is no direct relationship between the axes (singular vectors) and the clusters.

Table 1 shows the difference of cluster indicator between the hard clustering of SS-KK and soft clustering of SS-NMF. An exact orthogonality in SS-KK means that each row of cluster indicator  $\mathbf{G}$  has only one nonzero element, which implies that each data object belongs to only 1 cluster. The near-orthogonality of cluster indicator  $\mathbf{G}$  in SS-NMF relaxes this a bit, i.e., each data object could be



**Figure 1.** (a) An artificial toy data set consisting of two natural clusters (b) Data distribution in the SS-NMF subspace of the two column vectors of  $G$ . The data points from the two clusters get distributed along the two axes. (c) Data distribution in the SS-SNC subspace of the first two singular vectors. There is no relationship between the axes and the clusters.

long fractionally to more than 1 cluster. This can help in knowledge discovery in the cases where the data point is evenly projected along the different axes. For instance,  $\mathbf{g}_{16} = \{0.1220, 0.1233\}$  indicates that this data point may belong to any one of the two clusters.

SS-NMF uses an efficient iterative algorithm instead of solving a computationally expensive constrained eigen decomposition problem as in SS-SNC. The time complexity of SS-NMF is  $\mathcal{O}(tkn^2)$  where  $k$  is the number of clusters,  $n$  is the number of documents, and  $t$  is the number of iterations. In fact, the time complexity is similar to that of the classical SS-KK clustering algorithm. However, compared to SS-KK, SS-NMF algorithm is simple as it only involves some basic matrix operations and hence can be easily deployed over a distributed computing environment when dealing with large data sets. Another advantage in favor of SS-NMF is that a partial answer can be obtained at intermediate stages of the solution by specifying a fixed number of iterations.

## 4. Experiments and Results

In this Section, we empirically demonstrate the performance of SS-NMF in clustering documents by comparing it with well-established unsupervised and semi-supervised clustering algorithms.

### 4.1. Data Description

We primarily utilize the data set used in [15]<sup>2</sup>. Data sets *oh0* and *oh5* are from OHSUMED collection [16], a subset of MEDLINE database, which contains 233,445 documents indexed using 14,321 unique categories. Data set *re0* is from Reuters-21578 text categorization collection Distribution 1.0 [25]. Data set *Fbis* is from the Foreign Broadcast Information Service data of TREC-5 [33].

<sup>2</sup><http://www.cs.umn.edu/~han/data/tmdata.tar.gz>

For the experiments, we mixed some of the data sets mentioned above. Table 2 shows the details. These data sets were created as follows:

1. Classes *Graft-Survival* and *Phospholipids* from *oh5* were mixed to form the *Graft-Phos* data set.
2. Data set *England-Heart* was created by mixing classes *England* and *Heart-Valve-Prosthesis* from *oh0*.
3. *Interest-Trade* was formed by mixing *Interest* and *Trade* classes of *re0* data set.
4. We randomly selected 2, 3, 4, and 5 classes from *Fbis* to form data sets *Fbis2*, *Fbis3*, *Fbis4* and *Fbis5*, respectively.

We performed feature selection on the words according to [37] by retaining the top 10% of the words based on mutual information in each of the data sets.

**Table 2. Summary of data sets used in the experiments.**

Data sets	No. of clusters	No. of words	No. of docs
<i>Graft-Phos</i>	2	2432	293
<i>England-Heart</i>	2	2504	375
<i>Interest-Trade</i>	2	2682	438
<i>Fbis2</i>	2	2000	200
<i>Fbis3</i>	3	2000	300
<i>Fbis4</i>	4	2000	400
<i>Fbis5</i>	5	2000	500

### 4.2. Methodology and Evaluation Metrics

We evaluate the clustering results using confusion matrix and the accuracy metric AC. Each entry  $(i, j)$  in the confusion matrix represents the number of documents in cluster  $i$  that belong to true class  $j$ . The AC metric measures how accurately a learning method assigns labels  $\hat{y}_i$  to the ground truth  $y_i$ , and is defined as,

$$AC = \frac{\sum_{i=1}^n \delta(y_i, \hat{y}_i)}{n}. \quad (23)$$

where  $n$  denotes the total number of documents in the experiment, and  $\delta$  is the delta function that equals one if  $\hat{y}_i = y_i$ , else its zero. Since iterative algorithm is not guaranteed to find the global minimum, it is beneficial to run the algorithm several times with different initial values and choose one trial with a minimal objective value. In reality, usually a few number of trials is sufficient. In the case of NMF and  $k$ -means, for a given  $k$ , we conducted 20 test runs. 3 trials are performed in each of the 20 test runs and final accuracy value is the average of all the test runs.

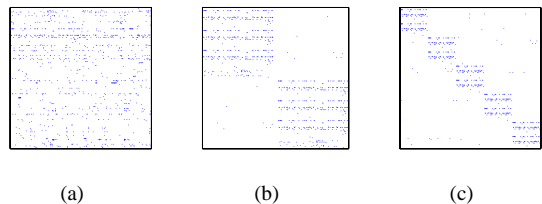
### 4.3. Clustering Results

We compare the performance of SS-NMF model on all the 7 data sets with the following 6 clustering methods: (1)  $k$ -means, (2) kernel  $k$ -means, (3) spectral normalized cuts, (4) NMF, (5) SS-KK, (6) SS-SNC. The first four methods are the most popular unsupervised data clustering methods, whereas SS-KK and SS-SNC are the representative semi-supervised ones. Through these comparison studies, we demonstrate the relative position of SS-NMF with respect to unsupervised and semi-supervised approaches to document clustering.

We first perform comparison of the 4 unsupervised clustering approaches with SS-NMF having pairwise constraints on only 3% pairs of all the possible document pairs, which is  $\binom{\text{total docs}}{2}$ . Each of the constraints were generated by randomly selecting a pair of documents. If both the documents have the same class label (*must-link*), then the constraint is assigned maximum weight in the document-document similarity matrix. On the other hand, if they belong to different classes (*cannot-link*), then the minimum weight in the similarity matrix is used for the constraint. For kernel  $k$ -means, we used a Gaussian (exponential) kernel  $K(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2/2\sigma^2)$ , with variance  $\sigma = 0.00001$  for 2 clusters and  $\sigma = 0.01$  for more than 2 clusters. In Table 3, we compare the algorithms on all the data sets using AC values. The performance of the first three methods is similar with NMF proving to be the best amongst the unsupervised methods. However, the accuracy of NMF greatly deteriorates and is unable to produce meaningful results on data sets having more than 2 clusters. On the other hand, the superior performance of SS-NMF is evident across all the data sets. We can see that in general a semi-supervised method can greatly enhance the document clustering results by benefitting from the user provided knowledge. Moreover, SS-NMF is able to generate significantly better results by quickly learning from the few pairwise constraints provided. Table 4, demonstrates the performance of SS-NMF when varying amounts of pairwise constraints were available *a priori*. We report the results in terms of the confusion matrix  $\mathbf{C}$  and the clus-

ter centroid matrix  $\mathbf{S}$ . As the available prior knowledge increases from 0% to 5%, we can make the following two key observations. Firstly, the confusion matrices tend to become perfectly diagonal indicating higher clustering accuracy. Second observation pertains to the cluster centroid matrix  $\mathbf{S}$  which represents the similarity or distance between the clusters. Increasing values of the diagonal elements of  $\mathbf{S}$  indicate higher inter-cluster similarities. As expected, when the amount of prior knowledge available is more, the performance of the algorithm clearly gets better.

In Figure 2a, the sparsity pattern of a typical document-document matrix  $\mathbf{A} = \mathbf{X}^T \mathbf{X}$  (*England-Heart* in the figure) before clustering is shown. The SS-NMF algorithm is applied to the modified similarity matrix  $\hat{\mathbf{A}}$ . Document clustering leads to re-ordering of the rows and columns of the matrix. Figures 2b and c, show the  $\hat{\mathbf{A}}$  matrices for *England-Heart* and *Fbis5* data sets after clustering with 5% pairwise constraints. Document clusters are indicated by the dense sub-matrices in these matrices.



**Figure 2.** (a) Typical document-document matrix (shown here *England-Heart*) before clustering (b) *England-Heart* similarity matrix after clustering with SS-NMF (c) *Fbis5* similarity matrix after clustering with SS-NMF.

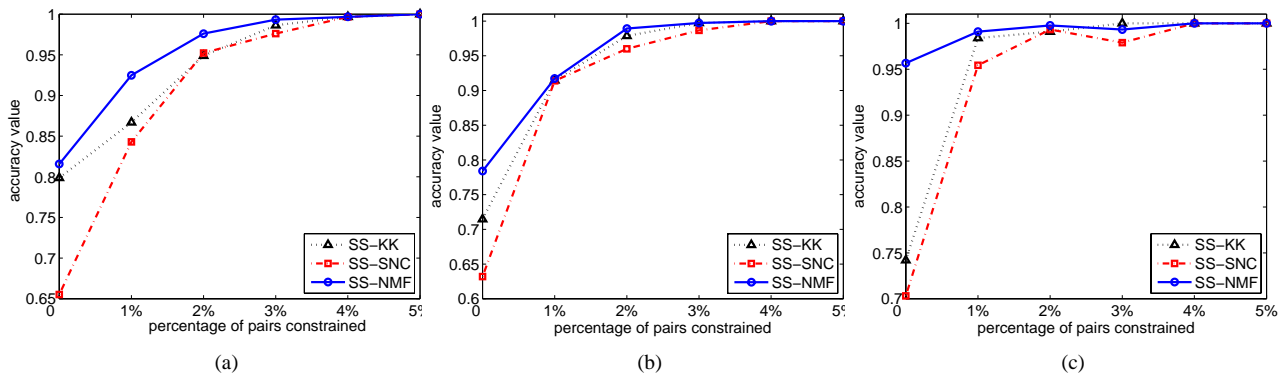
We now compare SS-NMF with the other two semi-supervised clustering approaches. As before, for SS-KK, a Gaussian kernel was used. In Figures 3 and 4, we plot the AC values against increasing percentage of pairwise constraints available, for the algorithms on all the data sets. On the whole, all three algorithms perform better as the percentage of pairwise constraints increases. While the performance of SS-KK is close to that of SS-SNC on the data sets in Figure 3, it is clearly left out of the race completely in Figure 4. This is mainly because of the fact that SS-KK is unable to maintain its accuracy when producing more than 2 clusters. While, the performance of SS-SNC is head-to-head with SS-NMF on *Fbis2* and *Fbis3*, it is consistently outperformed by SS-NMF on the rest of the data sets. Another noticeable fact is that the curve for SS-KK and SS-SNC might take a slow rise in some cases indicating that they need more amount of prior knowledge to improve the performance. Comparatively, SS-NMF gets better accuracy than the other two algorithms even for minimum percentage of pairwise constraints.

**Table 3. Comparison of document clustering accuracy between  $k$ -means, kernel  $k$ -means, spectral normalized cuts (SNC), NMF and, SS-NMF with 3% constraints.**

Data set	Graft-Phos	England-Heart	Interest-Trade	Fbis2	Fbis3	Fbis4	Fbis5
$k$ -means	0.6849	0.7108	0.7228	0.5650	0.4728	0.4620	0.4180
kernel $k$ -means	0.7986	0.7147	0.7420	0.5700	0.5533	0.5525	0.5140
SNC	0.6553	0.6320	0.7032	0.9900	0.6367	0.5975	0.5420
NMF	0.8157	0.7840	0.9566	0.9950	0.6533	0.6125	0.5900
SS-NMF	0.9932	0.9973	1.0000	1.0000	0.8833	0.8775	0.7520

**Table 4. The comparison of confusion matrix C and cluster centroid matrix S of SS-NMF for different percentages of document pairs constrained.**

% of constraints	Comparison matrix	Graft-Phos data set	England-Heart data set	Interest-Trade data set	Fbis5 data set				
0%	<b>C</b>	116 21 33 123	181 81 0 113	215 15 4 204	1 1 84 95 14 1 0 0 1 3	4 1 0 0 11 1 0 96 85 2	1 4 0 0 1 0 96 3 2 92		
	<b>S</b>	0.7771 0 0 0.7733	1.0364 0 0 1.1500	2.2788 0 0 2.0855	1.0695 0 0 0.8690 0 0 0 0 0 0	0 0 0 1.0392 0 0.87 0 0	0 0 0 0 0 0 0 0 0 1.0416		
1%	<b>C</b>	130 3 19 141	181 31 0 163	216 1 3 218	92 17 0 0 0 0 0 0 8 83	0 8 22 0 64 0 1 89 13 3	0 0 0 0 0 1 89 0 3 99		
	<b>S</b>	0.9143 0 0 0.9442	1.2164 0 0 1.5346	2.6920 0 0 2.4075	2.5203 0 0 2.4751 0 0 0 0 0 0	0 0 0 2.4251 0 2.6532 0 0	0 0 0 0 0 0 2.6532 0 0 2.8233		
3%	<b>C</b>	147 0 2 144	193 0 1 181	219 0 0 219	55 0 33 99 0 0 0 0 72 1	0 7 0 0 0 0 90 89 10 4	0 0 0 0 0 0 89 0 4 100		
	<b>S</b>	1.2317 0 0 1.3005	2.5813 0 0 2.7989	3.3250 0 0 3.7290	4.2578 0 0 4.6787 0 0 0 0 0 0	0 0 0 4.2349 0 4.0898 0 0	0 0 0 0 0 0 4.0898 0 0 4.0951		
5%	<b>C</b>	149 0 0 144	194 0 0 181	219 0 0 219	100 0 0 100 0 0 0 0 0 0	0 0 0 100 0 100 0 100 0 0	0 0 0 0 0 0 100 0 0 100		
	<b>S</b>	1.6094 0 0 1.5981	3.4279 0 0 2.5649	4.1829 0 0 4.5167	6.5171 0 0 6.3111 0 0 0 0 0 0	0 0 0 6.0427 0 6.7312 0 0	0 0 0 0 6.7312 0 0 5.9222		



**Figure 3.** Comparison of document clustering accuracy between SS-KK, SS-SNC, and SS-NMF for different percentages of document pairs constrained (a) *Graft-Phos* (b) *England-Heart* (c) *Interest-Trade* data set.

## 5. Conclusions

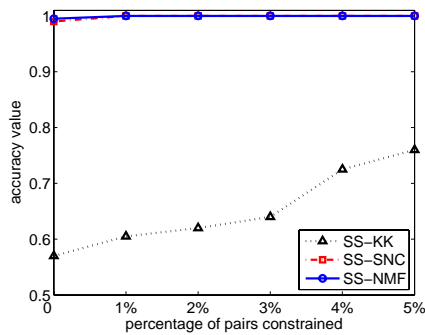
We presented SS-NMF: a semi-supervised approach for document clustering based on non-negative matrix factorization. In the proposed framework, users are able to provide supervision in terms of *must-link* and *cannot-link* pairwise constraints on the documents. We derived an iterative algorithm to perform symmetric tri-factorization of the document-document similarity matrix. We have proved that SS-NMF provides a general framework for semi-supervised clustering and that existing approaches can be considered as special cases of SS-NMF. Empirically, we showed that SS-NMF outperforms 6 well-established unsupervised and semi-supervised clustering methods in clustering documents using publicly available text data sets.

## Acknowledgements

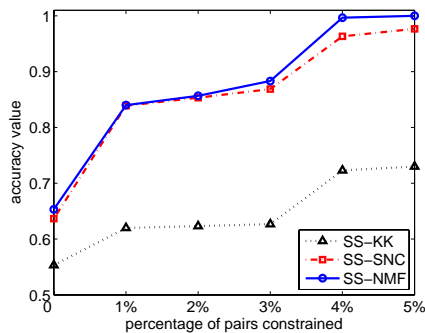
This research was partially funded by the 21<sup>st</sup> Century Jobs Fund Award, State of Michigan, under grant: 06-1-P1-0193.

## References

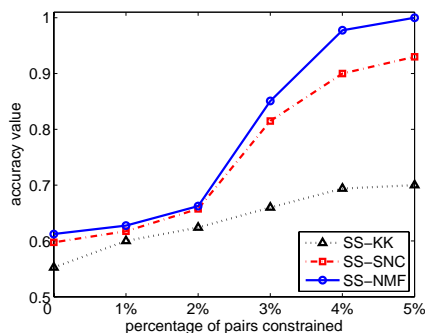
- [1] L. Baker and A. McCallum. Distributional clustering of words for text classification. In *proc. of ACM SIGIR*, 1998.
- [2] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. In *proc of International Conference on Machine Learning*, 2003.
- [3] S. Basu, A. Banerjee, and R. J. Mooney. Semi-supervised clustering by seeding. In *proc. of International Conference on Machine Learning*, 2002.
- [4] S. Basu, M. Bilenko, and R. J. Mooney. A probabilistic framework for semi-supervised clustering. In *proc. of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004.
- [5] A. Blum and T. M. Mitchell. Combining labeled and unlabeled data with co-training. In *proc of Workshop on Computational Learning Theory*, 1998.
- [6] D. Cai, X. He, and J. Han. Document clustering using locality preserving indexing. *IEEE Transactions on Knowledge and Data Engineering*, 17(12):1624–1637, 2005.
- [7] W. Croft. Clustering large files of documents using the single-link method. *Journal of the American Society of Information Science*, 28:341–344, 1977.
- [8] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. Scatter/gather: a cluster-based approach to browsing large document collections. In *proc. of ACM SIGIR Conference on Research and Development in Information Retrieval*, 1992.
- [9] C. Ding and X. He. Linearized cluster assignment via spectral ordering. In *proc. of International Conference on Machine Learning*, 2004.
- [10] C. Ding, X. He, and H. D. Simon. On the equivalence of non-negative matrix factorization and spectral clustering. In *proc. of SIAM International Conference on Data Mining*, 2005.
- [11] C. H. Q. Ding, X. He, H. Zha, M. Gu, and H. D. Simon. A min-max cut algorithm for graph partitioning and data clustering. In *proc. of International Conference on Machine Learning*, 2001.
- [12] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, Inc., 2000.
- [13] S. Godbole, A. Harpale, S. Sarawagi, and S. Chakrabarti. Document classification through interactive supervision of document and term labels. In *proc. of PKDD*, 2004.
- [14] M. Goldszmidt and M. Sahami. A probabilistic approach to full-text document clustering. Technical Report ITAD-433-MS-98-044, SRI International, 1998.
- [15] E.-H. Han and G. Karypis. Centroid-based document classification: Analysis and experimental results. In *proc. of PKDD*, 2000.
- [16] W. Hersh, C. Buckley, T. Leone, and D. Hickam. Ohsumed: An interactive retrieval evaluation and new large test collection for research. In *proc. of ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994.
- [17] A. Hotho, S. Staab, , and G. Stumme. Text clustering based on background knowledge. Technical Report 425, University of Karlsruhe, Institute AIFB, 2003.
- [18] Y. Huang and T. M. Mitchell. Text clustering with extended user feedback. In *proc. of ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006.



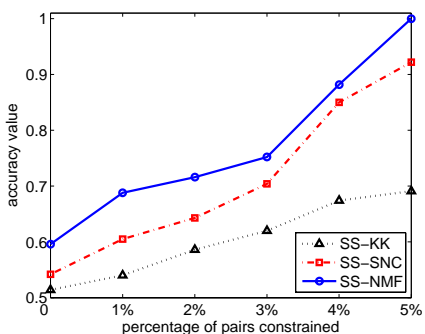
(a)



(b)



(c)



(d)

**Figure 4.** Comparison of document clustering accuracy between SS-KK, SS-SNC, and SS-NMF for different percentages of document pairs constrained (a) *Fbis2* (b) *Fbis3* (c) *Fbis4* and (d) *Fbis5* data sets.

- [19] X. Ji and W. Xu. Document clustering with prior knowledge. In *proc. of ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006.
- [20] T. Joachims. Transductive inference for text classification using support vector machines. In *proc. of International Conference on Machine Learning*, 1999.
- [21] R. Jones, A. McCallum, K. Nigam, and E. Riloff. Bootstrapping for text learning tasks. In *IJCAI-99 Workshop on Text Mining: Foundations, Techniques and Applications*, 1999.
- [22] B. Kulis, S. Basu, I. Dhillon, and R. Mooney. Semi-supervised graph clustering: a kernel approach. In *proc. of International Conference on Machine Learning*, 2005.
- [23] D. Lee and H. Seung. Algorithms for non-negative matrix factorization. In *proc. of Annual Conference on Neural Information Processing Systems*, 2001.
- [24] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.
- [25] D. D. Lewis. Reuters-21578 text categorization test collection distribution 1.0. <http://www.research.att/lewis>, 1999.
- [26] T. Li and C. Ding. The relationships among various nonnegative matrix factorization methods for clustering. In *proc. of IEEE International Conference on Data Mining*, 2006.
- [27] B. Liu, X. Li, W. S. Lee, and P. S. Yu. Text classification by labeling words. In *proc. of AAAI Conference on Artificial Intelligence*, 2004.
- [28] X. Liu, Y. Gong, W. Xu, and S. Zhu. Document clustering with cluster refinement and model selection capabilities. In *proc. of ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002.
- [29] B. Long, Z. Zhang, and P. S. Yu. Co-clustering by block value decomposition. In *proc. of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2005.
- [30] K. Nigam, A. K. McCallum, S. Thrun, and T. M. Mitchell. Learning to classify text from labeled and unlabeled documents. In *proc. of AAAI Conference on Artificial Intelligence*, 1998.
- [31] H. Raghavan, O. Madani, and R. Jones. Interactive feature selection. In *proc. of International Joint Conference on Artificial Intelligence*, 2005.
- [32] G. Salton and M. J. McGill. *Introduction to Modern Retrieval*. McGraw Hill, 1983.
- [33] TREC. Text retrieval conference, <http://trec.nist.gov>.
- [34] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl. Constrained k-means clustering with background knowledge. In *proc. of International Conference on Machine Learning*, 2001.
- [35] P. Willett. Recent trends in hierarchic document clustering: a critical review. *Inf. Process. Manage.*, 24(5):577–597, 1988.
- [36] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *proc. of ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003.
- [37] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *proc. of International Conference on Machine Learning*, 1997.
- [38] I. Yoo, X. Hu, and I.-Y. Song. Integration of semantic-based bipartite graph representation and mutual refinement strategy for biomedical literature clustering. In *proc. of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.