Semi-supervised Document Clustering with Simultaneous Text Representation and Categorization

Yanhua Chen, Lijun Wang, and Ming Dong

Machine Vision and Pattern Recognition Lab
Department of Computer Science, Wayne State University
Detroit, MI 48202, USA
{chenyanh,ljwang,mdong}@wayne.edu

Abstract. In order to derive high quality information from text, the field of text mining has advanced swiftly from simple document clustering to co-clustering with words and categories. However, document co-clustering without any prior knowledge or background information is a challenging problem. In this paper, we propose a Semi-Supervised Non-negative Matrix Factorization (SS-NMF) framework for document co-clustering. Our method computes new *word-document* and *document-category* matrices by incorporating user provided constraints through simultaneous distance metric learning and modality selection. Using an iterative algorithm, we perform tri-factorization of the new matrices to infer the document, category and word clusters. Theoretically, we show the convergence and correctness of SS-NMF co-clustering and the advantages of SS-NMF co-clustering over existing approaches. Through extensive experiments conducted on publicly available data sets, we demonstrate the superior performance of SS-NMF for document co-clustering.

Keywords: Semi-supervised co-clustering, Non-negative matrix factorization.

1 Introduction

Document clustering is the task of automatically organizing text documents into meaningful clusters (groups) such that documents in the same cluster are similar, and are dissimilar from the ones in other clusters. It is one of the most important tasks in text mining and has received extensive attention in the data mining community. A number of different techniques [1,2,3,4] were proposed in the literature for clustering documents.

With the rapid development of the Internet and computational technologies in the past decade, the field of text mining has advanced swiftly from simple document clustering to more demanding tasks such as the production of granular taxonomies, sentiment analysis, and document summarization, in the hope of deriving higher quality information from text. These new applications in text mining typically involve multiple interrelated types of objects (e.g., categories, documents and words). Consequently, co-clustering was proposed in the literature [5,6]. In the heart of word-document co-clustering, the similarity between documents is defined by their word representations while the similarity between words is defined by their appearances in documents. In other words, document similarity and word similarity are defined in a reinforcing manner. In such a way, document and word can be grouped at the same time, leading to

W. Buntine et al. (Eds.): ECML PKDD 2009, Part I, LNAI 5781, pp. 211–226, 2009.

simultaneous document clustering and text representation. Similarly, high-order co-clustering uses the information contained in categories, documents, and words together, and is able to discover a hidden global structure in the heterogeneous text data [7,8]. This global structure, integrating document clustering with simultaneous text representation and categorization, provides us a better understanding of the roles and interactions of words, documents and categories in text analysis, which is highly valuable in many applications, and not achievable when clustering each data type independently.

However, current co-clustering methods are mostly developed based on the spectral graph model, and thus inapplicable to large text data sets. Moreover, they are completely unsupervised. Accurately co-clustering documents without domain dependent background information is still a challenging task. In this paper, we propose a Semi-Supervised NMF (SS-NMF) based framework to incorporate prior knowledge into document co-clustering. Under the proposed SS-NMF co-clustering methodology, a user is able to provide constraints on a few documents specifying whether they "must" (*must-link*) or "cannot" (*cannot-link*) be clustered together. Our goal is to improve the quality of document co-clustering by learning a distance metric based on these constraints. Using an iterative algorithm, we perform tri-factorizations of the new *word-document* and *document-category* matrices, obtained with the learnt distance metric, to infer the document clusters while simultaneously deriving the text representation (word clusters) and categorization (category clusters). The major contribution of this work is summarized as follows,

1. We propose a novel algorithm for document co-clustering based on NMF. Computationally, NMF co-clustering is more efficient and flexible than spectral methods, and can provide more meaningful clustering results.

2. To the best of our knowledge, this is the first work on semi-supervised data co-clustering providing significance of each modality. Through distance metric learning and modality selection, prior knowledge is integrated into document co-clustering, making *must-link* documents as tight as possible and *cannot-link* documents as loose as possible.

3. From a theoretical perspective, our approach is mathematically rigorous. The convergence and correctness are proved. In addition, we show that our work provides a general framework for data co-clustering. Existing approaches such as the well-established spectral co-clustering algorithms can be considered as special cases of our method.

The rest of the paper is organized as follows. We review related work in Section 2. The proposed SS-NMF co-clustering algorithm is derived in Section 3. Our theoretical analysis on the correctness and convergence of the algorithm and on the advantages over spectral co-clustering approaches are presented in Section 4. Experimental results appear in Section 5. Finally, we conclude in Section 6.

2 Related Work

In this section, we briefly review related work in co-clustering (documents, words, and categories) and semi-supervised clustering.

In general, co-clustering approaches can be divided into two representative categories: information theory-based models and graph theoretic methods. In the former category, Dhillon et al. [9] presented a pairwise co-clustering algorithm to maximize the mutual information between the clustered random variables subject to the constraints on the number of row and column clusters. Later, Gao et al.[10] extended this method for high-order co-clustering. However, there is no sound objective function and theoretical proof on the effectiveness and correctness of these algorithms. On the other hand, graph theoretic approaches have a well-defined objective function. Spectral learning, such as Bipartite Spectral Graph Partitioning (BSGP) [5], was proposed and applied to co-cluster documents and words. With the similar philosophy, Gao et al. proposed Consistent Bipartite Graph Co-partitioning for high-order co-clustering to do hierarchical taxonomy preparation [7]. Recently, Rege et al. proposed to directly minimize the isoperimetric ratio of the weighted bipartite or high-order graph[11,12]. Experimental results on word-document and word-document-category co-clustering show that their approaches outperform the spectral methods in terms of the quality, speed, and stability. More recently, Long et al. [8] proposed Spectral Relational Clustering (SRC), in which they formulated heterogenous co-clustering as collective factorization on related matrices and derived a spectral algorithm to cluster multi-type interrelated data objects simultaneously. SRC provides more flexibility by lifting the requirement of one-to-one association in graph-based co-clustering. However, as a spectral method, it requires solving an eigen-problem, which computationally is not efficient to deal with large text data sets.

Semi-supervised clustering uses class labels or pairwise constraints on examples to aid unsupervised clustering. Two sources of supervised information are usually available to a semi-supervised clustering method: class labels or some pairwise constraints (*must-link* or *cannot-link*) as *a prior*. Existing methods for semi-supervised clustering based on source information generally fall into two categories: *semi-supervised clustering with labels* and *semi-supervised clustering with constraints* methods. In constraint-based approaches, the clustering algorithm itself is modified so that the available labels or constraints are used to bias the search for an appropriate clustering of the data [13]. In distance-based approaches, an existing clustering algorithm that uses a distance measure is employed; however, the distance measure is first trained to satisfy the labels or constraints in the supervised data [14]. Recent research in semi-supervised clustering tends to combine the constraint-based with distance-based approaches. Noticeable efforts on semi-supervised clustering algorithm include: Semi-Supervised Kernel K-means [15], Semi-Supervised Spectral Normalize Cuts [16] and SS-NMF [17,18]. In [19], it is shown that SS-NMF provides a unified framework for semi-supervised clustering. Many existing algorithms can be considered as special cases of SS-NMF. However, until now all semi-supervised methods are only applicable to homogeneous data clustering.

Even though the research on document co-clustering and semi-supervised clustering has attracted substantial attention in the past years, there has been no mathematically rigorous approach for semi-supervised data co-clustering. In the following, we will derive a theoretically sound algorithm based on SS-NMF, and apply it for document co-clustering.

3 SS-NMF Co-clustering

In this section, we first propose a SS-NMF model for general data co-clustering. Then, we narrow down to document clustering with simultaneous text representation and categorization, and discuss 1) how to incorporate prior knowledge through distance metric learning and modality selection, and 2) how to efficiently infer document, word and category clusters simultaneously using matrix factorization.

3.1 Model Formulation

Nonnegative Matrix Factorization (NMF) is a group of algorithms in multivariate analysis and and linear algebra where a matrix $\mathbf{X}$ is factorized into two nonnegative matrices, $\mathbf{F}$ and $\mathbf{G}$. It is initially proposed for "parts-of-whole" decomposition [20], and later extended to a general framework for data clustering [21]. It can model widely varying data distributions and do both hard and soft clustering. Let $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n) \in \mathbb{R}^{d \times n}$ be the data matrix of nonnegative elements. NMF factorizes $\mathbf{X}$ into two non-negative matrices,

$$\mathbf{X} \approx \mathbf{F}\mathbf{G}^T, \tag{1}$$

where $\mathbf{F} \in \mathbb{R}^{d \times k}$ is cluster centroid, $\mathbf{G} \in \mathbb{R}^{n \times k}$ is cluster indicator, and k is the number of clusters. The factorizations are typically obtained by the least square minimization.

Given a Heterogenous Relational Data (HRD) set, $\mathcal{X}_1 = \{x_{11}, ..., x_{1n_1}\}$,..., $\mathcal{X}_c = \{x_{c1}, ..., x_{cn_c}\}$,..., $\mathcal{X}_l = \{x_{l1}, ..., x_{ln_l}\}$, each representing one data type, our goal is to simultaneously cluster $\mathcal{X}_1$ into k_1 disjoint clusters, ..., and $\mathcal{X}_l$ into k_l disjoint clusters. To derive a solution to the co-clustering problem under matrix factorization framework, we first model HRD as a set of related matrices, i.e., a relation matrix $\mathbf{R}^{(pq)} \in \mathbb{R}^{n_p \times n_q}$ is used to represent the relations between $\mathcal{X}_p$ and $\mathcal{X}_q$ $(1 \leq p, q \leq l)$. Then, we can formulate the task of co-clustering as a optimization problem with nonnegative tri-factorization of $\mathbf{R}^{(pq)}$,

$$J = \min_{\mathbf{G}^{(p)} \geq 0, \mathbf{G}^{(q)} \geq 0, \mathbf{S}^{(pq)} \geq 0} \sum_{1 \leq p,q \leq l} \|\mathbf{R}^{(pq)} - \mathbf{G}^{(p)}\mathbf{S}^{(pq)}\mathbf{G}^{(q)}\|^2 \tag{2}$$

where $\mathbf{G}^{(p)} \in \mathbb{R}^{n_p \times k_p}$ and $\mathbf{G}^{(q)} \in \mathbb{R}^{k_q \times n_q}$ are the cluster indicator matrices, and $\mathbf{S} \in \mathbb{R}^{k_p \times k_q}$ is the cluster association matrix which gives the relation among the clusters of different data types.

In semi-supervised document co-clustering, supervision is typically provided as two sets of pairwise constraints derived from given labels on the documents: *must-link* constraints $M = \{(\mathbf{x}_i, \mathbf{x}_j)\}$ and *cannot-link* constraints $C = \{(\mathbf{x}_i, \mathbf{x}_j)\}$, where $(\mathbf{x}_i, \mathbf{x}_j) \in M$ implies that $\mathbf{x}_i$ and $\mathbf{x}_j$ are labeled as belonging to the same cluster, while $(\mathbf{x}_i, \mathbf{x}_j) \in C$ implies that $\mathbf{x}_i$ and $\mathbf{x}_j$ are labeled as belonging to different clusters. Figure 1 shows the triplet data (e.g., categories, documents and words), which is a basic element of general HRD. If we can successfully co-cluster such triplet data, the corresponding technique can be easily extended to structures involving more data types. In Figure 1, the relations between words and documents, and documents and categories are denoted by a *word-document* matrix $\mathbf{R}^{(12)}$ and a *document-category* matrix $\mathbf{R}^{(23)}$, respectively. The edges marked with M indicate the *must-link* constraints *M*, while the edges marked with C denote *cannot-link* constraints *C*. The dotted line shows the optimal clustering result. Note that in the following discussions, we will focus on the triplet co-clustering, or more specifically, word-document-category co-clustering. However, the derived algorithm is in general applicable to structures with more than three data types.

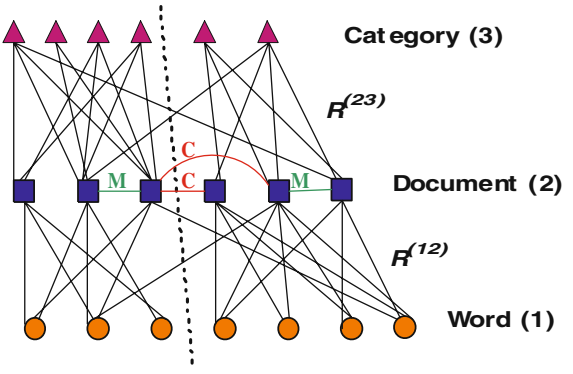


Fig. 1. Word-Document-Category co-clustering with must-link and cannot-link constraints

3.2 SS-NMF for Triplet Data

We now present the SS-NMF based triplet co-clustering algorithm. For simultaneous text representation and categorization, documents have to be clustered together with both words and categories. Let $\mathbf{R}^{(12)}$ and $\mathbf{R}^{(23)}$ denote the *word-document* and *document-category* matrix, respectively. The goal of SS-NMF triplet co-clustering is to iteratively cluster rows and columns of $\mathbf{R}^{(12)}$, and rows and columns of $\mathbf{R}^{(23)}$, subject to the *M* and *C* constraints on the documents. The first step in triplet co-clustering is to obtain the new matrix $\widetilde{\mathbf{R}}$ between different data types. In other words, we need to learn a distance metric $\mathbf{L}$ for each relation based on *must-link* and *cannot-link* constraints such that the clustering result on the central type (e.g., documents) is globally optimized. Specifically, a distance metric $\mathbf{L}^{(pq)}$ (where $(pq) \in \{(12),(23)\}$) over each relation of the form $d(\mathbf{x}_i^{(pq)}, \mathbf{x}_j^{(pq)}) = \sqrt{(\mathbf{x}_i^{(pq)} - \mathbf{x}_j^{(pq)})^T \mathbf{L}^{(pq)} (\mathbf{x}_i^{(pq)} - \mathbf{x}_j^{(pq)})}$ will be learnt, such that $(\mathbf{x}_i^{(pq)}, \mathbf{x}_j^{(pq)}) \in M$ are moved closer to each other while $(\mathbf{x}_i^{(pq)}, \mathbf{x}_j^{(pq)}) \in C$ are moved further away. That is, we solve the following optimization problem,

$$\max \quad g(\mathbf{L}^{(pq)}) = \frac{\sum_{(\mathbf{x}_i^{(pq)}, \mathbf{x}_j^{(pq)}) \in C} ||\mathbf{x}_i^{(pq)}, \mathbf{x}_j^{(pq)}||_{\mathbf{L}^{(pq)}}}{\sum_{(\mathbf{x}_i^{(pq)}, \mathbf{x}_j^{(pq)}) \in M} ||\mathbf{x}_i^{(pq)}, \mathbf{x}_j^{(pq)}||_{\mathbf{L}^{(pq)}}} \tag{3}$$

where $\|.\|$ is the Frobenius matrix norm. This maximization problem is equivalent to the generalized Semi-Supervised Linear Discriminate Analysis (SS-LDA) problem as follows,

$$J = \min \frac{trace(\mathbf{L}^{(pq)} \mathbf{W}_M{}^{(pq)})}{trace(\mathbf{L}^{(pq)} \mathbf{B}_C^{(pq)})} \tag{4}$$

where $\mathbf{W}_M$ is within-distance matrix from must-link constraints, $\mathbf{B}_C$ is between-distance matrix from cannot-link constraints, and can be solved accordingly [14]. Moreover, triplet co-clustering has an additional layer of complexity. Because categories and words can play a different role in the grouping of documents, we have to consider the issue of modality selection. To this end, we introduce a factor, $\mathbf{a} = [\alpha^{(12)}, \alpha^{(23)}]$, to denote the relative importance of "word" and "category". Note that the modality selection and distance metric learning are strongly dependent. This suggests that these two

objectives must be achieved simultaneously. In Algorithm 1, we propose an iterative algorithm to learn the optimal distance metrics $\mathbf{L}^{(12)}, \mathbf{L}^{(23)}$ and modality importance factor $\mathbf{a}$ for the given constraints. Based on the learnt distance metrics $\mathbf{L}^{(12)}$ and $\mathbf{L}^{(23)}$, we compute two new relational data matrices, $\widetilde{\mathbf{R}}^{(12)}$ and $\widetilde{\mathbf{R}}^{(23)}$. To achieve triplet co-clustering, we need to perform non-negative tri-factorization of new relational matrix as follows,

$$J = \min_{\substack{\mathbf{G}^{(1)}\geq 0,\mathbf{G}^{(2)}\geq 0,\mathbf{G}^{(3)}\geq 0 \\ \mathbf{S}^{(12)}\geq 0,\mathbf{S}^{(23)}\geq 0}} (\|\widetilde{\mathbf{R}}^{(12)} - \mathbf{G}^{(1)}\mathbf{S}^{(12)}(\mathbf{G}^{(2)})\|^2 + \|\widetilde{\mathbf{R}}^{(23)} - \mathbf{G}^{(2)^T}\mathbf{S}^{(23)}(\mathbf{G}^{(3)})\|^2) \tag{5}$$

Our main idea is to iteratively update the cluster structures for each data type in Equation (5). The details are given in Algorithm 2.

Algorithm 1. Simultaneous Distance Metric Learning and Modality Selection

INPUT: Original relational matrices $\mathbf{R}^{(12)}$,$\mathbf{R}^{(23)}$, central type $\mathcal{X}_2$ with must-link constraint *M*, and cannot-link constraint *C*
OUTPUT: Optimal distance metric $\mathbf{L}^{(12)}, \mathbf{L}^{(23)}$ and modality importance factor $\mathbf{a}$
METHOD:

1. Construct target relation $\widetilde{\mathbf{M}}$ based on constraints *M* and *C*, where each element $\widetilde{m}_{ij}$ is 1 if $(\mathbf{x}_i, \mathbf{x}_j) \in M$, and 0 if $(\mathbf{x}_i, \mathbf{x}_j) \in C$
2. Obtain the initial distance metrics $\mathbf{L}^{(12)}$ and $\mathbf{L}^{(23)}$ by SS-LDA with constraints *M* and *C*
3. Set the number of iterations t=0
   (a) Learn new relational matrices $\widetilde{\mathbf{R}}^{(12)}$ and $\widetilde{\mathbf{R}}^{(23)}$
   (b) Formulate matrices $\mathbf{M}^{(12)} = (\widetilde{\mathbf{R}}'^{(12)})^T\widetilde{\mathbf{R}}'^{(12)}$ and $\mathbf{M}^{(23)} = \widetilde{\mathbf{R}}'^{(23)}(\widetilde{\mathbf{R}}'^{(23)})^T$, where $\widetilde{\mathbf{R}}'^{(12)}$ and $\widetilde{\mathbf{R}}'^{(23)}$ contain only samples of $\mathcal{X}_2$ with constraints
   (c) Optimize the following function to obtain modality importance factor $\mathbf{a}$

$$\mathbf{a}^{opt} = arg\min_{\alpha} \|\widetilde{\mathbf{M}} - \alpha^{(12)}\mathbf{M}^{(12)} + \alpha^{(23)}\mathbf{M}^{(23)}\|^2$$

   (d) Learn the new distance metrics $\mathbf{L}^{(12)}$ and $\mathbf{L}^{(23)}$ for $\alpha^{(12)}\widetilde{\mathbf{R}}^{(12)}$ and $\alpha^{(23)}\widetilde{\mathbf{R}}^{(23)}$ by SS-LDA
4. If $\mathbf{a}_{t+1} - \mathbf{a}_t > \varepsilon$, set $t = t + 1$ and go to steps (a)-(d); otherwise, stop and output the optimal distance metrics $\mathbf{L}^{(12)}, \mathbf{L}^{(23)}$ and modality importance factor $\mathbf{a}$

4 Theoretical Analysis

4.1 Algorithm Convergence and Correctness

We now prove the theoretical convergence and correctness of SS-NMF co-clustering algorithm. Motivated by [6], we render the proof based on optimization theory, auxiliary function and several matrix inequalities.

Correctness. First, we prove the correctness of the algorithm, which can be stated as,

Proposition 1. *If the solution converges based on the updating rules in Equations (6)-(10), the solution satisfies the KKT optimality condition.*

Algorithm 2. SS-NMF for Triplet Co-Clustering

INPUT: Original relational matrices $\mathbf{R}^{(12)}$ and $\mathbf{R}^{(23)}$, new distance metrics $\mathbf{L}^{(12)}$ and $\mathbf{L}^{(23)}$
OUTPUT: Cluster indicator matrices $\mathbf{G}^{(1)}$, $\mathbf{G}^{(2)}$, and $\mathbf{G}^{(3)}$, cluster association matrices $\mathbf{S}^{(12)}$ and $\mathbf{S}^{(23)}$
METHOD:

1. Obtain new relational matrices through projection: $\widetilde{\mathbf{R}}^{(12)} = \sqrt{\mathbf{L}^{(12)}}\mathbf{R}^{(12)}$ and $\widetilde{\mathbf{R}}^{(23)} = \sqrt{\mathbf{L}^{(23)}}\mathbf{R}^{(23)}$
2. Initialize $\mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \mathbf{G}^{(3)}, \mathbf{S}^{(12)}, \mathbf{S}^{(23)}$ with non-negative values.
3. Iterate for each *i* and *h* until *convergence*

(a) Cluster indicator matrices:

$$\mathbf{G}^{(1)}_{ih} \leftarrow \mathbf{G}^{(1)}_{ih} \frac{(\widetilde{\mathbf{R}}^{(12)}\mathbf{G}^{(2)^T}\mathbf{S}^{(12)^T})_{ih}}{(\mathbf{G}^{(1)}\mathbf{S}^{(12)}\mathbf{G}^{(2)}\mathbf{G}^{(2)^T}\mathbf{S}^{(12)^T})_{ih}} \tag{6}$$

$$\mathbf{G}^{(2)}_{ih} \leftarrow \mathbf{G}^{(2)}_{ih} \frac{(\mathbf{S}^{(12)^T}\mathbf{G}^{(1)^T}\widetilde{\mathbf{R}}^{(12)}) + (\widetilde{\mathbf{R}}^{(23)}\mathbf{G}^{(3)^T}\mathbf{S}^{(23)^T})^T}{(\mathbf{S}^{(12)^T}\mathbf{G}^{(1)^T}\mathbf{G}^{(1)}\mathbf{S}^{(12)}\mathbf{G}^{(2)}) + (\mathbf{G}^{(2)^T}\mathbf{S}^{(23)}\mathbf{G}^{(3)}\mathbf{G}^{(3)^T}\mathbf{S}^{(23)^T})^T} \tag{7}$$

$$\mathbf{G}^{(3)}_{ih} \leftarrow \mathbf{G}^{(3)}_{ih} \frac{(\mathbf{S}^{(23)^T}\mathbf{G}^{(2)}\widetilde{\mathbf{R}}^{(23)})_{ih}}{(\mathbf{S}^{(23)^T}\mathbf{G}^{(2)}\mathbf{G}^{(2)^T}\mathbf{S}^{(23)}\mathbf{G}^{(3)})_{ih}} \tag{8}$$

(b) Cluster association matrices:

$$\mathbf{S}^{(12)}_{ih} \leftarrow \mathbf{S}^{(12)}_{ih} \frac{(\mathbf{G}^{(1)^T}\widetilde{\mathbf{R}}^{(12)}\mathbf{G}^{(2)^T})_{ih}}{(\mathbf{G}^{(1)^T}\mathbf{G}^{(1)}\mathbf{S}^{(12)}\mathbf{G}^{(2)}\mathbf{G}^{(2)^T})_{ih}} \tag{9}$$

$$\mathbf{S}^{(23)}_{ih} \leftarrow \mathbf{S}^{(23)}_{ih} \frac{(\mathbf{G}^{(2)}\widetilde{\mathbf{R}}^{(23)}\mathbf{G}^{(3)^T})_{ih}}{(\mathbf{G}^{(2)}\mathbf{G}^{(2)^T}\mathbf{S}^{(23)^T}\mathbf{G}^{(3)}\mathbf{G}^{(3)^T})_{ih}} \tag{10}$$

Proof. Following the standard theory of constrained optimization, we introduce the Lagrangian multipliers λ_1, λ_2, λ_3 , λ_4, λ_5 and λ_6 to minimize the Lagrangian function,

$$\begin{aligned} &L(\mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \mathbf{G}^{(3)}, \mathbf{S}^{(12)}, \mathbf{S}^{(23)}, \lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6) \\ &= \|\widetilde{\mathbf{R}}^{(12)} - \mathbf{G}^{(1)}\mathbf{S}^{(12)}\mathbf{G}^{(2)}\|^2 + \|\widetilde{\mathbf{R}}^{(23)} - \mathbf{G}^{(2)^T}\mathbf{S}^{(23)}\mathbf{G}^{(3)}\|^2 \\ &\quad - \mathrm{Tr}(\lambda_1\mathbf{C}^{(1)^T}) - \mathrm{Tr}(\lambda_2\mathbf{S}^{(12)^T}) - \mathrm{Tr}(\lambda_3\mathbf{C}^{(2)^T}) \\ &\quad - \mathrm{Tr}(\lambda_4\mathbf{C}^{(2)}) - \mathrm{Tr}(\lambda_5\mathbf{S}^{(23)^T}) - \mathrm{Tr}(\lambda_6\mathbf{C}^{(3)^T}) \end{aligned} \tag{11}$$

Based on the KKT complementarity conditions $\frac{\partial L}{\partial \mathbf{G}^{(1)}} = 0$, $\frac{\partial L}{\partial \mathbf{S}^{(12)}} = 0$, $\frac{\partial L}{\partial \mathbf{G}^{(2)}} = 0$, $\frac{\partial L}{\partial \mathbf{S}^{(23)}} = 0$ and $\frac{\partial L}{\partial \mathbf{G}^{(3)}} = 0$, we obtain the following five equations,

$$\begin{aligned} &2\widetilde{\mathbf{R}}^{(12)}\mathbf{G}^{(2)^T}\mathbf{S}^{(12)^T} - 2\mathbf{G}^{(1)}\mathbf{S}^{(12)}\mathbf{G}^{(2)}\mathbf{G}^{(2)^T}\mathbf{S}^{(12)^T} + \lambda_1 = 0 \\ &2\mathbf{G}^{(1)^T}\widetilde{\mathbf{R}}^{(12)}\mathbf{G}^{(2)^T} - 2\mathbf{G}^{(1)^T}\mathbf{G}^{(1)}\mathbf{S}^{(12)}\mathbf{G}^{(2)}\mathbf{G}^{(2)^T} + \lambda_2 = 0 \\ &2\mathbf{S}^{(12)^T}\mathbf{G}^{(1)^T}\widetilde{\mathbf{R}}^{(12)} - 2\mathbf{S}^{(12)^T}\mathbf{G}^{(1)^T}\mathbf{G}^{(1)}\mathbf{S}^{(12)}\mathbf{G}^{(2)} + \lambda_3 + \\ &(2\widetilde{\mathbf{R}}^{(23)}\mathbf{G}^{(3)^T}\mathbf{S}^{(23)^T} - 2\mathbf{G}^{(2)^T}\mathbf{S}^{(23)}\mathbf{G}^{(3)}\mathbf{G}^{(3)^T}\mathbf{S}^{(23)^T})^T + \lambda_4 = 0 \\ &2\mathbf{G}^{(2)}\widetilde{\mathbf{R}}^{(23)}\mathbf{G}^{(3)^T} - 2\mathbf{G}^{(2)}\mathbf{G}^{(2)^T}\mathbf{S}^{(23)}\mathbf{G}^{(3)}\mathbf{G}^{(3)^T} + \lambda_5 = 0 \\ &2\mathbf{S}^{(23)^T}\mathbf{G}^{(2)}\widetilde{\mathbf{R}}^{(23)} - 2\mathbf{S}^{(23)^T}\mathbf{G}^{(2)}\mathbf{G}^{(2)^T}\mathbf{S}^{(23)}\mathbf{G}^{(3)} + \lambda_6 = 0 \end{aligned}$$

We apply the Hadamard multiplication on both sides of above five equations by $\mathbf{G}^{(1)}$, $\mathbf{S}^{(12)}$, $\mathbf{G}^{(2)}$, $\mathbf{S}^{(23)}$, and $\mathbf{G}^{(3)}$, respectively. Using KKT conditions of

$$\lambda_1 \odot \mathbf{G}^{(1)} = 0 \quad \lambda_2 \odot \mathbf{S}^{(12)} = 0 \quad \lambda_3 \odot \mathbf{G}^{(2)} = 0$$
$$\lambda_4 \odot \mathbf{G}^{(2)} = 0 \quad \lambda_5 \odot \mathbf{S}^{(23)} = 0 \quad \lambda_6 \odot \mathbf{G}^{(3)} = 0$$

where $\odot$ denotes the Hadamard product of two matrices and letting $\lambda_3 = \lambda_4$, we can prove that if $\mathbf{G}^{(1)}$, $\mathbf{S}^{(12)}$, $\mathbf{G}^{(2)}$, $\mathbf{S}^{(23)}$, and $\mathbf{G}^{(3)}$ are a local minimizer of the objective function in Equation (11), the following five equations are satisfied,

$$((\widetilde{\mathbf{R}}^{(12)}\mathbf{G}^{(2)^T}\mathbf{S}^{(12)^T}) - (\mathbf{G}^{(1)}\mathbf{S}^{(12)}\mathbf{G}^{(2)}\mathbf{G}^{(2)^T}\mathbf{S}^{(12)^T})) \odot \mathbf{G}^{(1)} = 0$$
$$((\mathbf{G}^{(1)^T}\widetilde{\mathbf{R}}^{(12)}\mathbf{G}^{(2)^T}) - (\mathbf{G}^{(1)^T}\mathbf{G}^{(1)}\mathbf{S}^{(12)}\mathbf{G}^{(2)}\mathbf{G}^{(2)^T})) \odot \mathbf{S}^{(12)} = 0$$
$$((\mathbf{S}^{(12)^T}\mathbf{G}^{(1)^T}\widetilde{\mathbf{R}}^{(12)} + (\widetilde{\mathbf{R}}^{(23)}\mathbf{G}^{(3)^T}\mathbf{S}^{(23)^T})^T) - (\mathbf{S}^{(12)^T}\mathbf{G}^{(1)^T}\mathbf{G}^{(1)}$$
$$\mathbf{S}^{(12)}\mathbf{G}^{(2)} + (\mathbf{G}^{(2)^T}\mathbf{S}^{(23)}\mathbf{G}^{(3)}\mathbf{G}^{(3)^T}\mathbf{S}^{(23)^T})^T)) \odot \mathbf{G}^{(2)} = 0$$
$$((\mathbf{G}^{(2)}\widetilde{\mathbf{R}}^{(23)}\mathbf{G}^{(3)^T}) - (\mathbf{G}^{(2)}\mathbf{G}^{(2)^T}\mathbf{S}^{(23)}\mathbf{G}^{(3)}\mathbf{G}^{(3)^T})) \odot \mathbf{S}^{(23)} = 0$$
$$((\mathbf{S}^{(23)^T}\mathbf{G}^{(2)}\widetilde{\mathbf{R}}^{(23)}) - (\mathbf{S}^{(23)^T}\mathbf{G}^{(2)}\mathbf{G}^{(2)^T}\mathbf{S}^{(23)}\mathbf{G}^{(3)})) \odot \mathbf{G}^{(3)} = 0$$

Based on the above five equations, we derive the proposed updating rules of Equations (6)-(10). If the updating rules converge, the solution satisfies the KKT optimality condition. Proof is completed.

Convergence. Next, we prove the convergence of the algorithm. In Proposition 2, we show that the objective function decreases monotonically under the five updating rules of Equations (6)-(10) . This can be done by making use of an auxiliary function similar to that used in [21,22].

Proposition 2. *If any four of five matrices $\boldsymbol{G}^{(1)}$, $\boldsymbol{S}^{(12)}$, $\boldsymbol{G}^{(2)}$, $\boldsymbol{S}^{(23)}$, and $\boldsymbol{G}^{(3)}$ are fixed, $J = \|\widetilde{\boldsymbol{R}}^{(12)} - \boldsymbol{G}^{(1)}\boldsymbol{S}^{(12)}\boldsymbol{G}^{(2)}\|^2 + \|\widetilde{\boldsymbol{R}}^{(23)} - \boldsymbol{G}^{(2)^T}\boldsymbol{S}^{(23)}\boldsymbol{G}^{(3)}\|^2$ decreases monotonically under the updating rules of Equations (6)-(10).*

Proof. Due to the space constraints, we give the proof of convergence for one updating rule (e.g., the rule in Equation (6)) and skip the others. However, the proof of all five updating rules is similar to each other. The mathematical derivation below can be applied to other rules as well.

So, we need to show: If $\mathbf{S}^{(12)}, \mathbf{G}^{(2)}, \mathbf{S}^{(23)}$, and $\mathbf{G}^{(3)}$ are fixed matrices, then $J(\mathbf{G}^{(1)}) = \|\widetilde{\mathbf{R}}^{(12)} - \mathbf{G}^{(1)}\mathbf{S}^{(12)}\mathbf{G}^{(2)}\|^2 + \|\widetilde{\mathbf{R}}^{(23)} - \mathbf{G}^{(2)^T}\mathbf{S}^{(23)}\mathbf{G}^{(3)}\|^2$ decreases monotonically under the updating rule of Equation (6).

First, a function $F(\mathbf{G}^{(1)^{(t+1)}}, \mathbf{G}^{(1)^{(t)}})$ is called an auxiliary function of $J(\mathbf{G}^{(1)^{(t+1)}})$ if it satisfies the following two conditions: $F(\mathbf{G}^{(1)^{(t+1)}}, \mathbf{G}^{(1)^{(t)}}) \geq J(\mathbf{G}^{(1)^{(t+1)}})$ and $F(\mathbf{G}^{(1)^{(t+1)}}, \mathbf{G}^{(1)^{(t)}}) = J(\mathbf{G}^{(1)^{(t+1)}})$ for any $\mathbf{G}^{(1)^{(t+1)}}, \mathbf{G}^{(1)^{(t)}}$.

We define $\mathbf{G}^{(1)^{(t+1)}} = arg\min F(\mathbf{G}^{(1)^{(t+1)}}, \mathbf{G}^{(1)^{(t)}})$. By constructing an appropriate auxiliary function, we can prove the following equation,

$$J(\mathbf{G}^{(1)^{(t)}}) = F(\mathbf{G}^{(1)^{(t)}}, \mathbf{G}^{(1)^{(t)}}) \geq F(\mathbf{G}^{(1)^{(t+1)}}, \mathbf{G}^{(1)^{(t)}}) \geq J(\mathbf{G}^{(1)^{(t+1)}})$$

Thus, $J(\mathbf{G}^{(1)^{(t)}})$ is monotonic decreasing (non-increasing). The proof is completed.

4.2 Advantages of SS-NMF

We now show that NMF provides a general framework for data co-clustering by establishing the relationship between NMF and other well-known spectral high-order co-clustering algorithms, i.e., Spectral Relational Clustering (SRC) [8]. In fact, this algorithm can be considered as a special case of NMF co-clustering.

SRC is proposed in [8] for high-order document co-clustering. It iteratively embeds each type of data into low dimensional spaces and benefits from the interactions in the hidden structure of different data types. The underlying objective function is,

$$\min_{\mathbf{G}^{(1)T}\mathbf{G}^{(1)}=\mathbf{I},\mathbf{G}^{(2)T}\mathbf{G}^{(2)}=\mathbf{I},\mathbf{G}^{(3)T}\mathbf{G}^{(3)}=\mathbf{I}} (\|\mathbf{R}^{(12)} - \mathbf{G}^{(1)}\mathbf{S}^{(12)}\mathbf{G}^{(2)}\|^2 + \|\mathbf{R}^{(23)} - \mathbf{G}^{(2)T}\mathbf{S}^{(23)}\mathbf{G}^{(3)}\|^2)$$

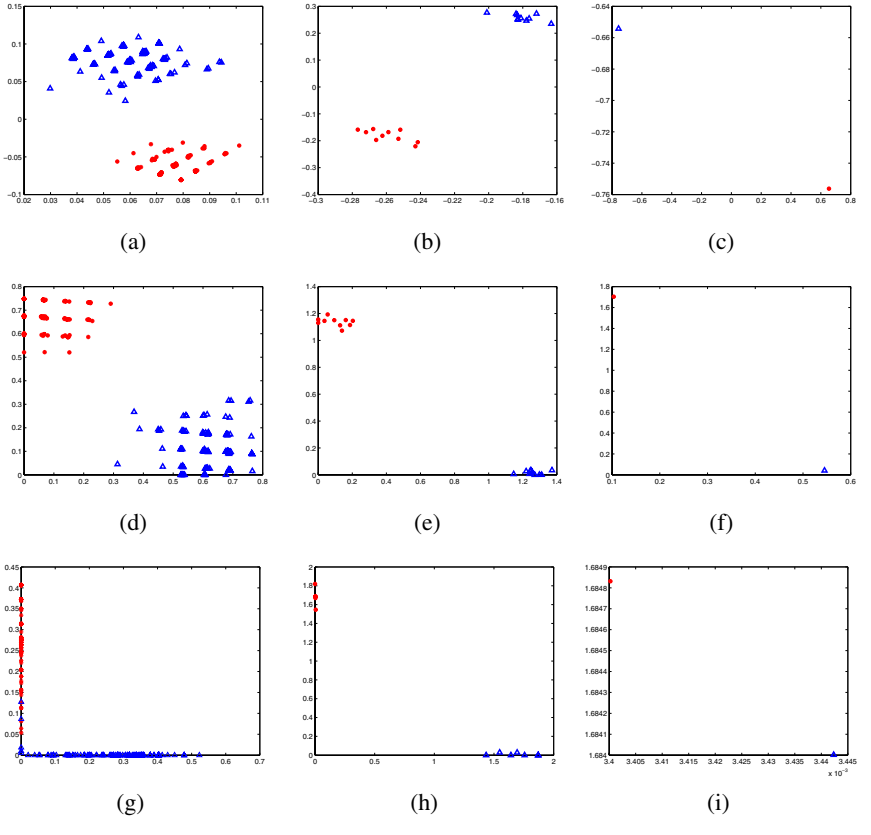


Fig. 2. (a)-(c): Clustering results by SRC in the subspace of the first two singular vectors of $\mathbf{G}^{(1)}$,$\mathbf{G}^{(2)}$ and $\mathbf{G}^{(3)}$. There is no direct relationship between the axes and the clusters. (d)-(f): Clustering results by NMF in the subspace of the two column vectors of $\mathbf{G}^{(1)}$,$\mathbf{G}^{(2)}$ and $\mathbf{G}^{(3)}$. The data points from the two clusters are distributed closely to the two axes. (g)-(i): Clustering results by SS-NMF (with 5% constraints) in the subspace of the two column vectors of $\mathbf{G}^{(1)}$,$\mathbf{G}^{(2)}$ and $\mathbf{G}^{(3)}$. The data points from the two clusters are distributed exactly along the two axes.

On the other hand, NMF-based high-order co-clustering is to minimize the following function,

$$\min_{\mathbf{G}^{(1)}\geq 0,\mathbf{G}^{(2)}\geq 0,\mathbf{G}^{(3)}\geq 0,\mathbf{S}^{(12)}\geq 0,\mathbf{S}^{(23)}\geq 0} (\|\mathbf{R}^{(12)} - \mathbf{G}^{(1)}\mathbf{S}^{(12)}\mathbf{G}^{(2)}\|^2 + \|\mathbf{R}^{(23)} - \mathbf{G}^{(2)^T}\mathbf{S}^{(23)}\mathbf{G}^{(3)}\|^2)$$

It is clear that the major difference between NMF co-clustering and SRC lies in the fact that SRC requires the cluster indictor matrices be orthogonal, while NMF co-clustering relaxes this requirements to be near-orthogonal. This relaxation can provide us more meaning clustering results.

The advantage of NMF or SS-NMF over SRC can best be illustrated using an example. In the following example, the synthetic data set has 200 words, 20 documents, and 2 categories, each having two clusters of equal size. More specifically, we have two relational matrices: $\mathbf{R}^{(12)}$ of size 200×20 and $\mathbf{R}^{(23)}$ of size 20×2, both binary matrices with 2-by-2 block structures generated by the Bernoulli distribution. $\mathbf{R}^{(12)}$ is generated based on the block structure $\begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}$ and $\mathbf{R}^{(23)}$ is based on the block structure $\begin{bmatrix} 0.8 & 0.2 \\ 0.1 & 0.9 \end{bmatrix}$.

Unlike SRC, NMF or SS-NMF maps the data into a non-negative latent semantic space which is not required to be orthogonal. Panels (a)-(c), (d)-(f) and (g)-(i) in Figure 2 show the clustering results by SRC, NMF and SS-NMF, in which two clusters are denoted by stars and triangles, respectively. For NMF or SS-NMF, we plot the data points in the subspace of two column vectors of $\mathbf{G}^{(1)}$, $\mathbf{G}^{(2)}$ and $\mathbf{G}^{(3)}$, while for SRC the subspace of the first two singular vectors is used. Note that for either NMF or SS-NMF, each data point takes a non-negative value on both axes. In the NMF subspace, each axis corresponds to a cluster, and all the data points belonging to the same cluster are nicely located closely to the axis. In the SS-NMF subspace, the data points belonging to the same cluster are almost spread along the axis. This indicates that SS-NMF can provide better clustering accuracy than unsupervised NMF because the cluster label for a data point is determined by finding the axis with which the data point has the largest projection value. On the other hand, in the SRC subspace, we observe no direct relationship between the axes (singular vectors) and the clusters.

5 Experiments and Results

In this section, we empirically demonstrated the performance of SS-NMF in co-clustering documents, words and categories by comparing it with well-established co-clustering algorithms. Through these comparisons, we showed the relative position of SS-NMF with respect to existing approaches to document co-clustering[1]. All algorithms were implemented using MATLAB 7.0.

[1] At present, there is no other existing work on semi-supervised co-clustering with constraints, so a comparison is not feasible.

5.1 Data Description and Preprocessing

We have primarily utilized the data set used in [23] [2]. Data sets $oh5$ and $oh15$ are from OHSUMED collection, a subset of MEDLINE database, which contains $233,445$ documents indexed using $14,321$ unique categories. Data set WAP is from the WebACE Project, and each document corresponds to a web page listed in the subject hierarchy of Yahoo!. Data set $re0$ is from $Reuters-21578$ text categorization collection (distribution 1.0). We also used $Newsgroup$ data which contains about 2000 articles from 20 newsgroups [24] [3]. In our experiments, we mixed up some of the data sets mentioned above. Table 1 gives the details of the data sets for word-document-category co-clustering.

Table 1. Data sets for text (word-document-category) co-clustering

Name	*Data sets*	*Data structure*	*No. of categories*	*No. of clusters*	*No. of documents*
HT1	*oh15,re0*	{*Adenosine-Diphosphate,Aluminum,Cell-Movement*}, {*cpi,money*}	*2*	*5*	*899*
HT2	*oh15,re0*	{*Blood-Coagulation-Factors,Enzyme-Activation,Staphylococcal-Infections*}, {*jobs,reserves*}	*2*	*5*	*461*
HT3	*oh15,re0*	{*Aluminum,Blood-Coagulation-Factors,Blood-Vessels*} {*housing,retail*}	*2*	*5*	*256*
HT4	*oh5,re0*	{*Aluminum,Cell-Movement,Staphylococcal-Infections*}, {*cpi,wpi*}	*2*	*5*	*391*
HT5	*WAP,re0*	{*media,film,music*}, {*cpi,jobs*}	*2*	*5*	*404*
HT6	*Newsgroup*	{*rec.sport.baseball,rec.sport.hockey*}, {*talk.politics.guns,talk.politics.mideast,talk.politics.misc*}	*2*	*5*	*500*
HT7	*Newsgroup*	{*comp.graphics,comp.os.ms-windows.misc*}, {*rec.autos,rec.motorcycles*}, {*sci.crypt,sci.electronics*}	*3*	*6*	*300*

We used term frequency to build *word-document* matrix and carry out feature selection to choose the top 1000 words by the mutual information. The *Document-category* matrix is constructed by computing the probability of each document belonging to each category. The following technique is used: (1) For each class of documents, select the top 1000 words based on mutual information. (2) For each document, if any of the top 1000 word occurs, the amount of occurrence is 1, otherwise 0. (3) The probability of one document belonging to a category is the ratio of the sum of occurrence of the top 1000 words in this document to 1000. Thus, every element of *document-category* matrix is in the range $[0,1]$. In addition, for semi-supervised clustering, we defined the percentage $(\%)$ of pairwise constraints with respect to all the possible document pairs, which is $\binom{\text{total docs}}{2}$. The document constrains are generated by randomly selecting documents from each class of the data set.

5.2 Evaluation Method

We evaluated the clustering results using the accuracy rate AC. The AC metric measures how accurately a learning method assigns labels $\hat{y}_i$ to the ground truth y_i, and is defined as,

[2] http://www.cs.umn.edu/ ∼han/data/tmdata.tar.gz

[3] http://www.cs.uiuc.edu/homes/dengcai2/Data/TextData.html

$$AC = \frac{\sum_{i=1}^{n} \delta(y_i, \hat{y_i})}{n}. \tag{12}$$

where n denotes the total number of documents/categories in the experiment, and δ is the delta function that equals one if $\hat{y_i} = y_i$, otherwise zero. Since an 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 the average of all the test runs as the final accuracy value. In our experiments, for each given cluster number k, we conducted 10 test runs and final AC value is the average of all the 10 test runs.

5.3 Word-Document-Category Co-clustering

We conducted experiments to co-cluster words, documents and categories, and compared the performance of SS-NMF with SRC [8] and unsupervised NMF.

Co-clustering Accuracy. Table 2 shows document clustering accuracy obtained by SRC, unsupervised NMF, and SS-NMF with 15% constraints, respectively. It is obvious that SS-NMF outperforms SRC or unsupervised NMF in all the data sets. In general, SRC performs the worst amongst the three. Its accuracy on data set HT7 with 3 categories and 6 document clusters is only 19%, while SS-NMF provides an accuracy over 63%. Also from Table 2, we observed that SS-NMF can achieve high clustering accuracy, over 80% in 5 out of 7 data sets. Note that we have at least five document clusters in each of these data sets, so comparatively the baseline accuracy is only 20%. In Figure 3, we plotted the AC value against increasing percentage of pairwise constraints for SS-NMF. It is obvious to see that SRC and unsupervised NMF are consistently outperformed by SS-NMF with varying amounts of constrains across all the data sets. In addition, when more prior knowledge is available, the performance of SS-NMF clearly gets better.

In the left panel of Table 3, we reported the accuracy of text categorization by SRC, unsupervised NMF, and SS-NMF. For all the data sets, the AC value of SS-NMF is either the best or the closely-followed second best amongst the three methods. This result shows that even though the original document-category matrix is biased in the distance metric learning towards the constraints on the documents, SS-NMF still can provide a highly competitive results on category clustering.

For co-clustering, we obtained the clusters of words simultaneously with the clusters of documents and categories. However, for text representation, there is no ground truth available to compute an AC value. Here, we selected the "top" 10 words based on mutual information for each word cluster associated with a category cluster, and listed them in the right panel of Table 3. These words can be used to represent the underlying "concept" of the corresponding category cluster.

Modality Selection. As described in Section 3.2, distance metric and modality importance are learnt iteratively in Algorithm 1. First, modality selection can provide additional information on the relative importance of various relations (e.g., "word" and "category") for the grouping the central data type (e.g., "document"). Moreover, from a technical point of view, it also acts like feature selection when computing the new relational data matrix. Table 4 lists the modality importance for the two relations: *word-document* and *document-category* in SS-NMF with 1% constraints. A higher value in

Table 2. Comparison of document clustering accuracy between SRC, unsupervised NMF and SS-NMF with 15% constraints on word-document-category co-clustering

Name	*SRC*	*NMF*	*SS-NMF*
HT1	*0.4772*	*0.5250*	*0.8509*
HT2	*0.4989*	*0.6529*	*0.8243*
HT3	*0.3359*	*0.5391*	*0.6875*
HT4	*0.4450*	*0.5601*	*0.8261*
HT5	*0.6411*	*0.6386*	*0.8267*
HT6	*0.4989*	*0.5780*	*0.8620*
HT7	*0.1900*	*0.4333*	*0.6467*

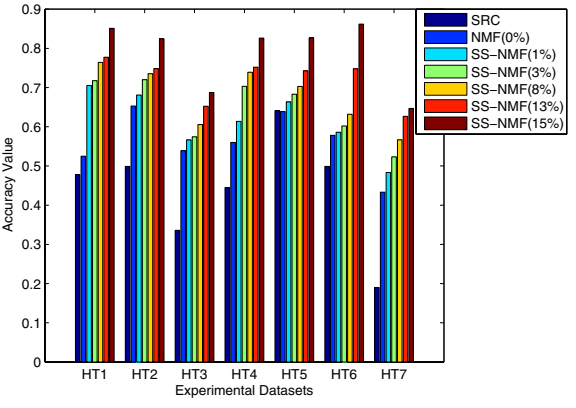


Fig. 3. Comparison of document clustering accuracy between SRC, unsupervised NMF, and SS-NMF with different amounts of constraints for word-document-category co-clustering

Table 3. Text categorization: clustering accuracy of categories and Text representation: top ten words for each category

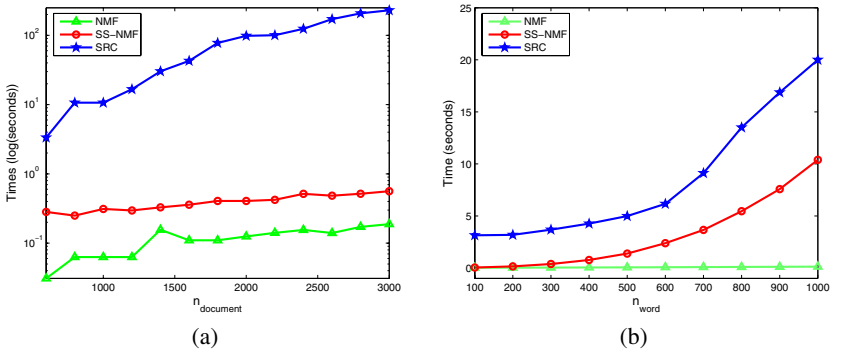
Name	*SRC*	*NMF*	*SS-NMF*	*Representative words for each category*
HT1	*0.8*	*0.8*	*0.8*	*{via,coverag,calcium,purif,modifi,increm,identif,receiv,explant,delta}* *{market,pct,bank,rate,monei,billion,dollar,mln,dlr,currenc}*
HT2	*0.8*	*0.6*	*0.8*	*{studi,activ,patient,suggest,protein,increas,result,effect,treat,infect}* *{januari,pct,februari,reserv,unemploy,billion,bank,fell,mln,rose}*
HT3	*0.4*	*0.8*	*0.6*	*{increas,patient,activ,perform,suggest,studi,effect,examin,result,factor}* *{februari,adjust,fall,sale,depart,retail,fell,season,level,month}*
HT4	*0.4*	*0.8*	*0.8*	*{cell,treatment,determin,site,bone,neutrophil,single,anim,change,differ}* *{consum,statist,index,inflat,rise,compar,base,month,increas,rose}*
HT5	*0.8*	*0.4*	*0.8*	*{pm,star,film,hollywood,set,releas,octob,director,time,million}* *{rise,price,rose,statist,unemploy,inflat,compar,consum,januari,increas}*
HT6	*0.8*	*0.6*	*0.8*	*{disregard,jai,pyramid,winner,aaron,baltimor,dean,leaf,ban,stanlei}* *{sahak,ohanus,melkonian,appressian,serazuma,armenian,serdar,escap,turkish,sdpa}*
HT7	*0.8*	*0.5*	*0.7*	*{mac,color,al,push,bit,sse,lower,size,traffic,screen}* *{licenc,egreeneast,clipper,drink,claim,biker,safeti,clean,dod,motorcycl}* *{vga,univ,pub,servic,educ,bill,robert,school,technic,game}*

Table 4. Modality importance: words v.s. categories

Name	*word-document*	*document-category*
HT1	*0.9996*	*0.3884*
HT2	*0.9999*	*0.4331*
HT3	*0.6837*	*0.9949*
HT4	*0.7607*	*0.7233*
HT5	*0.2479*	*0.9998*
HT6	*0.9999*	*0.1751*
HT7	*0.2390*	*0.9990*

the table indicates more importance. It is clear that the significance of words and categories are quite distinct in different data sets. Specifically, *word-document* relation seems to play an more important role for document clustering in data sets HT1, HT2, HT4 and HT6, while *document-category* relation is more important in the rest. This information provides us a better understanding of the underlying process that generates the document clusters.

Time Complexity. Finally, we compared the computational speed of SRC, unsupervised NMF and SS-NMF for document co-clustering. The time complexity of SRC is $\mathcal{O}(t(l\max(n_d, n_w)^3 + kn_dn_f)$, SS-NMF is $\mathcal{O}(tl(n_d^3 + kn_dn_f))$, and unsupervised NMF is $\mathcal{O}(tlkn_dn_f)$, where t is the number of iterations, l is the number of data types, $k = \max(k_d, k_f)$ is maximum number of clusters in all data types (e.g., word, document or category), n_d is the number of documents, and n_f is the maximum number of words or categories for all modalities. Normally, $n_f = n_w$ since n_w(number of words)¿ n_c (number of categories). So, given t, l and k , the actual computational speed is usually determined by n_d or n_w. Figure 4(a) illustrates the computational speed of SRC, SS-NMF and unsupervised NMF, with increasing number of documents for a fixed n_w, while Figure 4(b) shows the computational speed with increasing number of words for a fixed n_d. The experiments were performed on a machine with Dual 3GHz Intel Xeon processors and 2GB RAM.



(a) (b)

Fig. 4. Computational speed comparison for SRC , Unsupervised NMF, and SS-NMF. The time required by each of the algorithms are displayed in log(seconds) for increasing $n_{document}$ (a), and in seconds for increasing n_{word} (b).

Amongst the three, unsupervised NMF is the quickest as it uses an efficient iterative algorithm to compute the cluster indicator and cluster association matrices. SS-NMF ranks the second and its time gradually increases as the number of samples or features increases. The difference between SS-NMF and unsupervised NMF is mainly due to the additional computation required to learn the new distance metric through SS-LDA, in which we need to solve a generalized eigen-problem. We observed that in Figure 4(a), the computing time for SS-NMF is close to unsupervised NMF since both have linear complexity with n_d when n_w is fixed. On the other hand as shown in Figure 4(b), time for SS-NMF increases more quickly ($\mathcal{O}(tln_w^3)$) when n_c is fixed. In both cases, SRC is the slowest comparatively. Even though SRC is completely unsupervised, it needs to solve a computationally more expensive constrained eigen-decomposition problem and require additional k-means post-processing to infer the clusters. In short, SS-NMF approach provides the efficient way for document co-clustering.

6 Conclusions

In this paper, we presented SS-NMF co-clustering: a novel semi-supervised approach for document clustering with simultaneous text representation and categorization. In SS-NMF co-clustering model, users are able to provide supervision in terms of *must-link* and *cannot-link* pairwise constraints on the documents, which are used to derive new *word-document* and *document-category* matrices through distance metric learning and modality selection. Tri-factorization of the new matrices is then performed to obtain the grouping of documents, words and categories. We demonstrated that SS-NMF outperforms existing methods in document co-clustering on publicly available text data sets.

Acknowledgment

This research was partially funded by U. S. National Science Foundation under grants IIS-0713315 and CNS-0751045, and by the 21st Century Jobs Fund Award, State of Michigan, under grant 06-1-P1-0193.

References

1. Liu, X., Gong, Y., Xu, W., Zhu, S.: Document clustering with Cluster Refinement and Model Selection Capabilites. In: Proc. of ACM SIGIR, pp. 191–198 (2002)
2. Willett, P.: Recent Trends in Hierarchic Document Clustering: a Critical Review. Information Process Management 24(5), 577–597 (1988)
3. Ding, C., He, X., Zha, H., Simon, H.: A Min-max Cut Algorithm for Graph Partitioning and Data Clustering. In: Proc. of IEEE ICDM, pp. 107–114 (2001)
4. Xu, W., Liu, X., Gong, Y.: Document Clustering based on Non-negative Matrix Factorization. In: Proc. of ACM SIGIR, pp. 267–273 (2003)
5. Dhillon, I.S.: Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning. In: Proc. of ACM SIGKDD, pp. 269–274 (2001)

6. Long, B., Zhang, Z., Yu, P.S.: Co-clustering by Block Value Decomposition. In: Proc. of ACM SIGKDD, pp. 635–640 (2005)
7. Gao, B., Liu, T.-Y., Cheng, Q., Feng, G., Qin, T., Ma, W.-Y.: Hierarchical Taxonomy Preparation for Text Categorization using Consistent Bipartite Spectral Graph Copartitioning. IEEE Transactions on Knowledge and Data Engineering 17(9), 1263–1273 (2005)
8. Long, B., Zhang, Z., Wu, X., Yu, P.S.: Spectral Clustering for Multi-type Relational Data. In: Proc. of ICML, pp. 585–592 (2006)
9. Dhillon, I.S., Mallela, S., Modha, D.S.: Information-theoretic Co-clustering. In: Proc. of ACM SIGKDD, pp. 89–98 (2003)
10. Gao, B., Liu, T.-Y., Mao, W.-Y.: Star-structured High-order Heterogeneous Data Co-clustering based on Consistent Information Theory. In: Proc. of IEEE ICDM, pp. 880–884 (2006)
11. Rege, M., Dong, M., Fotouhoi, F.: Co-clustering Documents and Words using Bipartite Isoperimetric Graph Partitioning. In: Proc. of IEEE ICDM, pp. 532–541 (2006)
12. Rege, M., Dong, M., Hua, J.: Graph Theoretical Framework for Simultaneously Integrating Visual and Textual Features for Efficient Web Image Clustering. In: Proc. of WWW, pp. 317–326 (2008)
13. Hiu, M., Law, C., Topchy, A., Jain, A.K.: Model-based Clustering with Probabilistic Constraints. In: Proc. of SIAM ICDM, pp. 641–645 (2005)
14. Xing, E.P., Ng, A.Y., Jordan, M.I., Russell, S.: Distance Metirc Learning with Application to Clustering with Side-information. In: Proc. of NIPS, pp. 362–371 (2001)
15. Kulis, B., Basu, S., Dhillon, I.S., Mooney, R.: Semi-supervised Graph Clustering:a Kernel Approach. In: Proc. of ICML, pp. 457–464 (2005)
16. Ji, X., Xu, W.: Document Clustering with Prior Knowledge. In: Proc. of ACM SIGIR, pp. 405–412 (2006)
17. Chen, Y., Rege, M., Dong, M., Hua, J.: Incorporating User Provided Constraints into Document Clusteirng. In: Proc. of IEEE ICDM, pp. 577–582 (2007)
18. Chen, Y., Rege, M., Dong, M., Fotouhi, F.: Deriving Semantics for Image Clustering from Accumulated User Feedbacks. In: Proc. of ACM MM, pp. 313–316 (2007)
19. Chen, Y., Rege, M., Dong, M., Hua, J.: Non-negative Matrix Factorization for Semi-supervised Data Clustering. Journal of Knowledge and Information Systems 17(3), 355–379 (2008)
20. Lee, D.D., Seung, H.S.: Learning the Parts of Objects by Non-negative Matrix Factorization. Nature 401, 788–791 (1999)
21. Ding, C., Li, T., Peng, W., Park, H.: Orthogonal Nonnegative Matrix Tri-factorizations for Clustering. In: Proc. of ACM SIGKDD, pp. 126–135 (2006)
22. Lee, D.D., Seung, H.S.: Algorithms for Non-negative Matirx Factroization. In: Proc. of NIPS, pp. 362–371 (2001)
23. Han, E.-H., Karypis, G.: Centroid-based document classification: Analysis and experimental results. In: Zighed, D.A., Komorowski, J., Żytkow, J.M. (eds.) PKDD 2000. LNCS (LNAI), vol. 1910, pp. 424–431. Springer, Heidelberg (2000)
24. Lang, K.: News weeder: Learning to Filter Networks. In: Proc. of ICML, pp. 331–339 (1995)