Constrained Text Coclustering with Supervised and Unsupervised Constraints

Yangqiu Song, *Member*, *IEEE*, Shimei Pan, Shixia Liu, *Member*, *IEEE*,
Furu Wei, Michelle X. Zhou, *Senior Member*, *IEEE*, and Weihong Qian

Abstract—In this paper, we propose a novel constrained coclustering method to achieve two goals. First, we combine information-theoretic coclustering and constrained clustering to improve clustering performance. Second, we adopt both supervised and unsupervised constraints to demonstrate the effectiveness of our algorithm. The unsupervised constraints are automatically derived from existing knowledge sources, thus saving the effort and cost of using manually labeled constraints. To achieve our first goal, we develop a two-sided hidden Markov random field (HMRF) model to represent both document and word constraints. We then use an alternating expectation maximization (EM) algorithm to optimize the model. We also propose two novel methods to automatically construct and incorporate document and word constraints to support unsupervised constrained clustering: 1) automatically construct document constraints based on overlapping named entities (NE) extracted by an NE extractor; 2) automatically construct word constraints based on their semantic distance inferred from WordNet. The results of our evaluation over two benchmark data sets demonstrate the superiority of our approaches against a number of existing approaches.

Index Terms—Constrained clustering, coclustering, unsupervised constraints, text clustering

1 Introduction

Clustering is a popular technique for automatically organizing or summarizing a large collection of text; there have been many approaches to clustering [1]. As described below, for the purpose of our work, we are particularly interested in two of them: *coclustering* and *constrained clustering*.

Unlike traditional clustering methods that focus on 1D clustering, coclustering examines both document and word relationship at the same time. Previous studies have shown that coclustering is more effective than 1D clustering in many applications [2], [3], [4], [5].

In addition to coclustering approaches, researchers have also developed constrained clustering methods to enhance document clustering [6], [7]. However, since purely unsupervised document clustering is often difficult, most constrained clustering approaches are semi-supervised, requiring the use of manually labeled constraints.

To further enhance clustering performance, there has also been some effort on combining coclustering and constrained clustering [8], [9], [10]. However, there are two main deficiencies in the existing methods. First, they all optimize a sum squared residues-based objective function, which has been shown to be not as effective as KL-divergence [11]. Kullback-Leibler divergence (KL-divergence) on text is defined on two multinomial distributions and has proven to be very effective in coclustering text [11]. Second, they all use semi-supervised learning that requires ground-truth or human annotated labels to construct constraints. In practice, however, ground-truth labels are difficult to obtain, and human annotations are time consuming and costly. As a result, it is important to investigate methods that can automatically derive constraints based on existing knowledge sources. Next, we describe how we extend the work in [12] to address the above issues.

When clustering textual data, one of the most important distance measures is document similarity. Since document similarity is often determined by word similarity, the semantic relationships between words may affect document clustering results. For example, sharing common named entities (NE) among documents can be a cue for clustering these documents together. Moreover, the relationships among vocabularies such as synonyms, antonyms, hypernyms, and hyponyms, may also affect the computation of document similarity. Consequently, introducing additional knowledge on documents and words may facilitate document clustering. To incorporate word and document constraints, we propose an approach called constrained information-theoretic coclustering (CITCC). It integrates constraints into the information theoretic coclustering (ITCC) framework [4], where KL-divergence is adopted to better model textual data. The constraints are modeled with two-sided hidden Markov random field (HMRF) regularizations. We develop an alternating expectation maximization (EM) algorithm to optimize the model. As a result, CITCC can simultaneously cluster two sets of discrete random variables such as words and documents under the constraints extracted from both sides.

- *Y. Song, S. Liu, and F. Wei are with Microsoft Research Asia, 5/F, Beijing Sigma Center 49, Zhichun Road, Haidian District, Beijing 100193, China. E-mail: {yangqiu.song, shliu, fuwei}@microsoft.com.*
- *S. Pan is with IBM Research - T. J. Watson Center, Hawthorne, NY 10532. E-mail: shimei@us.ibm.com.*
- *M.X. Zhou is with IBM Research - Almaden Center, San Jose, CA 95120. E-mail: mzhou@us.ibm.com.*
- *W. Qian is with IBM Research - China, Beijing 100101, China. E-mail: qianwh@cn.ibm.com.*

Manuscript received 2 Apr. 2011; revised 22 Nov. 2012; accepted 22 Feb. 2012; published online 2 Mar. 2012.
Recommended for acceptance by P. Ipeirotis.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2011-04-0171.
Digital Object Identifier no. 10.1109/TKDE.2012.45.

In summary, the main contributions of this paper are twofold.

- We proposed a new constrained coclustering algorithm CITCC: 1) It performed better than the existing coclustering algorithms because it allows the system to incorporate additional constraints to guide the clustering towards the ground-truth; 2) it performed better than the existing 1D constrained clustering methods since it can take advantage of the co-occurrences of documents and words; 3) it performed better than the existing constrained coclustering approaches on text data since it optimizes a KL-divergence based objective function versus a euclidean distance-based function that is commonly used by other systems.
- We proposed two novel methods to automatically construct and incorporate constraints into CITCC to help improve document clustering performance. Since both the constraints are automatically constructed by the system, it performs purely unsupervised document clustering. More specifically: 1) we automatically construct document constraints based on the overlapping named entities extracted by an NE extractor; 2) we automatically construct word constraints based on their semantic distance inferred from WordNet. We have also conducted comprehensive evaluations over two benchmark data sets and the evaluation results demonstrated the superiority of our algorithm.

In the rest of the paper, following a review of the existing works, we describe the details of the proposed CITCC approach and the two new algorithms to automatically construct word and document constraints from existing knowledge sources. We then describe how we evaluate the effectiveness of the proposed methods using both supervised and unsupervised constraints. Finally, we conclude the paper with a summary and our plans for future work.

2 Related Work

Existing works that are most relevant to ours fall into three categories: coclustering, semi-supervised clustering, and constrained coclustering with unsupervised constraints. In this section, we briefly summarize the works in each category.

2.1 Coclustering

Most coclustering algorithms deal with dyadic data, e.g., the document and word co-occurrence frequencies. The dyadic data can be modeled as a bipartite graph, and then spectral graph theory is adopted to solve the partition problem [3]. The co-occurrence frequencies can also be encoded in co-occurrence matrices and then matrix factorizations are utilized to solve the clustering problem [5], [13]. The document and word co-occurrence can also be formulated as a two-sided generative model using a Bayesian interpretation [14], [15]. Moreover, Dhillon et al. [4] modeled the coclustering algorithm as an information-theoretic partition, which is mathematically equivalent to the empirical joint probability distribution of two sets of discrete random variables. Later, Banerjee et al. [11] extended this method to a general Bregman coclustering and matrix factorization framework.

2.2 Semi-Supervised Clustering

There are two types of semi-supervised clustering methods: semi-supervised clustering with labeled seeding points [16], [17], [18] and semi-supervised clustering with labeled constraints [19], [20], [21], [22], [23], [7]. Constraint-based clustering methods often use pairwise constraints such as “must-links” and “cannot-links” to enhance unsupervised clustering algorithms. Although these constraints are also called “side-information,” most of them are built on human provided labels and the clustering methods are thus considered as semi-supervised learning [6].

While the above semi-supervised methods are applicable to 1D clustering, we are more interested in coclustering. For text data, coclustering can not only show the relationship between document and word clusters, but also leverage the knowledge transferred between the two sides [24]. There are some initial efforts on extending the existing coclustering methods to semi-supervised coclustering and constrained coclustering [8], [9], [10], [25]. Most of these methods are based on matrix factorizations that optimize a sum squared residues-based objective function. It has been reported that among the existing coclustering methods, the ITCC algorithm that uses KL-divergence is empirically more effective in analyzing sparse and high-dimensional text data than those methods that use euclidean distance [11]. As a result, we focused this work on extending the ITCC framework to incorporate both document and word constraints.

2.3 Unsupervised Constrained Clustering

Recently, some research has been conducted to handle constraints automatically derived based on either human-provided meta data or existing knowledge sources (e.g., the ontology in Wikipedia, or the social tagging on images.) More specifically, Li et al. demonstrated that the ACM keyword taxonomy can help cluster scientific papers using a nonnegative matrix factorization (NMF) approach [26]. They suggested that the knowledge of scientific conference categories can be transferred from the word side to the document side. Moreover, Li et al. showed that sentiment words can help semi-supervised sentiment classification using NMF [27]. Yang et al. proposed a new algorithm to handle noisy constraints that are derived from the links between citations [28]. More recently, Shi et al. [25] proposed a constrained spectral coclustering approach which can also incorporate unsupervised word constraints. The method first conducts a coclustering algorithms on a fine-labeled corpus. Then it constructs the word constraints based on the word categories learned from the axillary corpus.

Unlike these approaches, we add must-links for documents when two documents have many overlapped NEs. While for word constraints, we add must-links if the two words are close to each other semantically, which is measured by a WordNet-based semantic similarity.

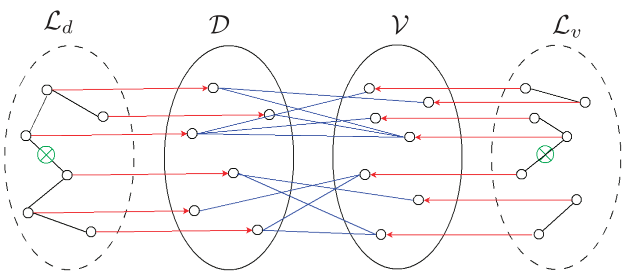


Fig. 1. Illustration of the HMRF^2-ITCC model.

3 The CITCC Method

In this section, we first describe how we formulate the constrained coclustering problem as a two-sided HMRF regularized ITCC (HMRF^2-ITCC) model. Then we present how to use an alternating EM algorithm to optimize the model.

3.1 Problem Formulation

We denote the document set and word set as $\mathcal{D} = \{d_1, d_2, \dots, d_M\}$ and $\mathcal{V} = \{v_1, v_2, \dots, v_V\}$. Then the joint probability of $p(d_m, v_i)$ can be computed based on the co-occurrence count of d_m and v_i. For hard clustering problems, as illustrated by Dhillon et al. [4], a function

$$q(d_m, v_i) = p(\hat{d}_{k_d}, \hat{v}_{k_v}) p(d_m|\hat{d}_{k_d}) p(v_i|\hat{v}_{k_v}), \tag{1}$$

where $\hat{d}_{k_d}$ and $\hat{v}_{k_v}$ are cluster indicators, k_d and k_v are the cluster indices, is used to approximate $p(d_m, v_i)$ by minimizing the Kullback-Leibler (KL) divergence

$$\begin{aligned} & D_{KL}(p(\mathcal{D},\mathcal{V})\|q(\mathcal{D},\mathcal{V})) \\ &\quad = D_{KL}(p(\mathcal{D},\mathcal{V},\hat{\mathcal{D}},\hat{\mathcal{V}})\|q(\mathcal{D},\mathcal{V},\hat{\mathcal{D}},\hat{\mathcal{V}})) \\ &\quad = \sum_{k_d}^{K_d} \sum_{d_m: l_{d_m}=k_d} p(d_m) D_{KL}(p(\mathcal{V}|d_m)\|p(\mathcal{V}|\hat{d}_{k_d})) \\ &\quad = \sum_{k_v}^{K_v} \sum_{v_i: l_{v_i}=k_v} p(v_i) D_{KL}(p(\mathcal{D}|v_i)\|p(\mathcal{D}|\hat{v}_{k_v})), \end{aligned} \tag{2}$$

where $\hat{\mathcal{D}}$ and $\hat{\mathcal{V}}$ are the cluster sets, $p(\mathcal{V}|\hat{d}_{k_d})$ denotes a multinomial distribution based on the probabilities $(p(v_1|\hat{d}_{k_d}), \dots, p(v_V|\hat{d}_{k_d}))^T$, $p(v_i|\hat{d}_{k_d})) = p(v_i|\hat{v}_{k_v}) p(\hat{v}_{k_v}|\hat{d}_{k_d})$ and $p(v_i|\hat{v}_{k_v}) = p(v_i)/p(l_{v_i} = \hat{v}_{k_v})$. Symmetrically, we can define the probability for words: $p(\mathcal{D}|\hat{v}_{k_v})$ denotes a multinomial distribution based on the probabilities $(p(d_1|\hat{v}_{k_v}), \dots, p(d_V|\hat{v}_{k_v}))^T$, $p(d_i|\hat{v}_{k_v})) = p(d_i|\hat{d}_{k_d}) p(\hat{d}_{k_d}|\hat{v}_{k_v})$ and $p(d_i|\hat{d}_{k_d}) = p(d_i)/p(l_{d_i} = \hat{d}_{k_d})$.

As shown in Fig. 1, we introduce two latent label sets $\mathcal{L}_d = \{l_{d_1}, l_{d_2}, \dots, l_{d_M}\}$ for documents and $\mathcal{L}_v = \{l_{v_1}, l_{v_2}, \dots, l_{v_V}\}$ for words. Then the original ITCC can be mathematically formulated as the log-likelihood of a conditional probability in the exponential family

$$\begin{aligned} & p(\mathcal{D},\mathcal{V}|\mathcal{L}_d,\mathcal{L}_v) \\ &\quad = \exp\big(-D_{KL}(p(\mathcal{D},\mathcal{V},\hat{\mathcal{D}},\hat{\mathcal{V}})\|q(\mathcal{D},\mathcal{V},\hat{\mathcal{D}},\hat{\mathcal{V}}))\big) b_{\phi_{KL}}(\cdot), \end{aligned} \tag{3}$$

where $b_{\phi_{KL}}(\cdot)$ is a normalization constant determined by its divergency type [11].

For the constrained clustering problem, we use HMRF to formulate the prior information for both document and word latent labels. As illustrated in Fig. 1, the "must-links" and "cannot-links" for both documents and words are encoded in the HMRFs. In the following, we focus on deriving the constraints for $\mathcal{L}_d$. It is easy to generalize the derivation to $\mathcal{L}_v$.

First, for latent label l_{d_m}, the must-link set is denoted as $\mathcal{M}_{d_m}$, and the cannot-link set as $\mathcal{C}_{d_m}$. The neighbor set of l_{d_m} is denoted as $\mathcal{N}_{d_m} = \{\mathcal{M}_{d_m}, \mathcal{C}_{d_m}\}$. Then the latent labels l_{d_m} $(m = 1, \dots, M)$ construct a neighborhood graph and the random field defined on this graph is a Markov random field, following the Markov property: $p(l_{d_m}|\mathcal{L}_d - \{l_{d_m}\}) = p(l_{d_m}|l_{d_m} \in \mathcal{N}_{d_m})$. As a result, the configuration of the latent label set can be expressed as a Gibbs distribution. Following the *generalized Potts* energy function and its extension [22], we have

$$p(\mathcal{L}_d) = \frac{1}{Z_d} \exp\left(-\sum_{d_{m_1}}^{M} \sum_{d_{m_2} \in \mathcal{N}_{d_{m_1}}} V(d_{m_1}, d_{m_2})\right). \tag{4}$$

For must-links, the energy function is defined as

$$\begin{aligned} & V(d_{m_1}, d_{m_2} \in \mathcal{M}_{d_{m_1}}) \\ &\quad = a_{m_1,m_2} D_{KL}(p(\mathcal{V}|d_{m_1})\|p(\mathcal{V}|d_{m_2})) \cdot I_{l_{d_{m_1}} \neq l_{d_{m_2}}}, \end{aligned} \tag{5}$$

and for cannot-links, the energy function is formulated as

$$\begin{aligned} & V(d_{m_1}, d_{m_2} \in \mathcal{C}_{d_{m_1}}) \\ &\quad = \bar{a}_{m_1,m_2} (D_{max} - D_{KL}(p(\mathcal{V}|d_{m_1})\|p(\mathcal{V}|d_{m_2}))) \cdot I_{l_{d_{m_1}} = l_{d_{m_2}}}, \end{aligned} \tag{6}$$

where $p(\mathcal{V}|d_{m_1})$ denotes a multinomial distribution based on the probabilities $(p(v_1|d_{m_1}), \dots, p(v_V|d_{m_1}))^T$, D_{max} is the maximum value for all the $D_{KL}(p(\mathcal{V}|d_{m_1})\|p(\mathcal{V}|d_{m_2}))$, a_{m_1,m_2} and $\bar{a}_{m_1,m_2}$ are tradeoff parameters to be set empirically, and $I_{true} = 1$, $I_{false} = 0$.

Consequently, the constrained coclustering problem can be formulated as an MAP estimation for label configurations

$$p(\mathcal{L}_d, \mathcal{L}_v|\mathcal{D},\mathcal{V}) \propto p(\mathcal{D},\mathcal{V}|\mathcal{L}_d,\mathcal{L}_v) p(\mathcal{L}_d) p(\mathcal{L}_v). \tag{7}$$

As we have two HMRF priors for $\mathcal{L}_d$ and $\mathcal{L}_v$, we call this two-sided HMRF regularization. Mathematically, the objective function can be rewritten as

$$\begin{aligned} \{\mathcal{L}_d, \mathcal{L}_v\} = \arg\min & \; D_{KL}\big(p(\mathcal{D},\mathcal{V},\hat{\mathcal{D}},\hat{\mathcal{V}})\|q(\mathcal{D},\mathcal{V},\hat{\mathcal{D}},\hat{\mathcal{V}})\big) \\ & + \sum_{d_{m_1}}^{M} \sum_{d_{m_2} \in \mathcal{M}_{d_{m_1}}} V(d_{m_1}, d_{m_2} \in \mathcal{M}_{d_{m_1}}) \\ & + \sum_{d_{m_1}}^{M} \sum_{d_{m_2} \in \mathcal{C}_{d_{m_1}}} V(d_{m_1}, d_{m_2} \in \mathcal{C}_{d_{m_1}}) \\ & + \sum_{v_{i_1}}^{V} \sum_{v_{i_2} \in \mathcal{M}_{v_{i_1}}} V(v_{i_1}, v_{i_2} \in \mathcal{M}_{v_{i_1}}) \\ & + \sum_{v_{i_1}}^{V} \sum_{v_{i_2} \in \mathcal{C}_{v_{i_1}}} V(v_{i_1}, v_{i_2} \in \mathcal{C}_{v_{i_1}}), \end{aligned} \tag{8}$$

where $\mathcal{M}_{v_{i_1}}$ and $\mathcal{C}_{v_{i_1}}$ are the must-link and cannot-link sets for latent label $l_{v_{i_1}}$ of word v_{i_1}.

3.2 Alternating EM

Since globally optimizing the latent labels as well as the approximating function $q(d_m, v_i)$ is intractable, we substitute (2) into the objective function (8) to alternate the optimization process. To achieve this, we first fix $\mathcal{L}_v$ and minimize the objective in (8) w.r.t $\mathcal{L}_d$. Then we fix $\mathcal{L}_d$ and minimize the objective in (8) w.r.t $\mathcal{L}_v$. The process continues until convergence is achieved.

When we fix $\mathcal{L}_v$, the objective function for $\mathcal{L}_d$ is rewritten as

$$\begin{aligned}\mathcal{L}_d = \arg\min \sum_{k_d}^{K_d} \sum_{d_m : l_{d_m} = k_d} & p(d_m) D_{KL}(p(\mathcal{V}|d_m) \| p(\mathcal{V}|\hat{d}_{k_d})) \\ + \sum_{d_{m_1}}^{M} \sum_{d_{m_2} \in \mathcal{M}_{d_{m_1}}} & V(d_{m_1}, d_{m_2} \in \mathcal{M}_{d_{m_1}}) \\ + \sum_{d_{m_1}}^{M} \sum_{d_{m_2} \in \mathcal{C}_{d_{m_1}}} & V(d_{m_1}, d_{m_2} \in \mathcal{C}_{d_{m_1}}),\end{aligned} \tag{9}$$

However, optimizing this objective function is still computationally intractable. It becomes clear that we need a more feasible approximation. Here, we use a general EM algorithm to find an estimation [22]. There are two steps in the EM algorithm: the E-step and the M-step.

In the E-Step, we update the cluster labels based on the fixed model function $q(d_m, v_i)$ from the last iteration. More exactly, we use the iterated conditional mode (ICM) algorithm [22] to find the cluster labels. ICM greedily solves the objective function by updating one latent variable at a time, and keeping all the other latent variables fixed. In our implementation, we find the label l_{d_m} by

$$\begin{aligned}l_{d_m} = \arg\min_{l_{d_m} = k_d} & D_{KL}(p(\mathcal{V}|d_m) \| p(\mathcal{V}|\hat{d}_{k_d})) \\ + \sum_{\substack{d_{m'} \in \mathcal{M}_{d_m} \\ I_{l_{d_m} \neq l_{d_{m'}}}}} & a_{m,m'} D_{KL}(p(\mathcal{V}|d_m) \| p(\mathcal{V}|d_{m'})) \\ + \sum_{\substack{d_{m'} \in \mathcal{C}_{d_m}; \\ I_{l_{d_m} = l_{d_{m'}}}}} & \bar{a}_{m,m'} (D_{max} - D_{KL}(p(\mathcal{V}|d_m) \| p(\mathcal{V}|d_{m'}))).\end{aligned} \tag{10}$$

In the M-Step, we update the model function $q(d_m, v_i)$ by fixing $\mathcal{L}_d$ and $\mathcal{L}_v$. Since the latent labels are fixed, the update of q is not affected by the must-links and cannot-links. Thus, we can modify them as

$$q(\hat{d}_{k_d}, \hat{v}_{k_v}) = \sum_{l_{d_m} = k_d} \sum_{l_{v_i} = k_v} p(d_m, v_i), \tag{11}$$

$$q(d_m|\hat{d}_{k_d}) = \frac{q(d_m)}{q(l_{d_m} = k_d)} \quad [q(d_m|\hat{d}_{k_d}) = 0 \text{ if } l_{d_m} \neq k_d], \tag{12}$$

$$q(v_i|\hat{v}_{k_v}) = \frac{q(v_i)}{q(l_{v_i} = k_v)} \quad [q(v_i|\hat{v}_{k_v}) = 0 \text{ if } l_{v_i} \neq k_v], \tag{13}$$

where $q(d_m) = \sum_{v_i} p(d_m, v_i)$, $q(v_i) = \sum_{d_m} p(d_m, v_i)$, $q(\hat{d}_{k_d}) = \sum_{k_v} p(\hat{d}_{k_d}, \hat{v}_{k_v})$ and $q(\hat{v}_{k_v}) = \sum_{k_d} p(\hat{d}_{k_d}, \hat{v}_{k_v})$. More detailed derivations of M-step can be found in [4]. Algorithm 1 summarizes the main steps in the procedure. The convergence property is described in the Lemma.

Algorithm 1 Alternating EM for HMRF^2-ITCC model.

```
Input: Document and word sets $\mathcal{D}$ and $\mathcal{V}$; cluster numbers $K_d$ and $K_v$; pairwise constraints $\mathcal{M}$ and $\mathcal{C}$.
Initialize document and word cluster labels using Kmeans.
Initialize $q^{(0)}(\hat{d}_{k_d}, \hat{v}_{k_v})$, $q^{(0)}(d_m|\hat{d}_{k_d})$ and $q^{(0)}(v_i|\hat{v}_{k_v})$.
while $t <$ maxIter and $\delta >$ max$\delta$ do
    Document E-Step: compute document clusters using the ICM algorithm to minimize
        $\mathcal{L}_d^{(t+1)} = \arg\min \sum_{k_d}^{K_d} \sum_{d_m : l_{d_m} = k_d} p(d_m) D_{KL}(p(\mathcal{V}|d_m) || p(\mathcal{V}|\hat{d}_{k_d})) + \sum_{d_{m_1}}^{M} \sum_{d_{m_2} \in \mathcal{M}_{d_{m_1}}} V(d_{m_1}, d_{m_2} \in \mathcal{M}_{d_{m_1}}) + \sum_{d_{m_1}}^{M} \sum_{d_{m_2} \in \mathcal{C}_{d_{m_1}}} V(d_{m_1}, d_{m_2} \in \mathcal{C}_{d_{m_1}})$.
    Document M-Step: update parameters
        $q^{(t+1)}(\hat{d}_{k_d}, \hat{v}_{k_v})$, $q^{(t+1)}(d_m|\hat{d}_{k_d})$ and $q^{(t+1)}(v_i|\hat{v}_{k_v})$.
    and compute $q^{(t+1)}(d_m|\hat{v}_{k_v})$.
    Word E-Step: compute document clusters using the ICM algorithm to minimize
        $\mathcal{L}_v^{(t+2)} = \arg\min \sum_{k_v}^{K_v} \sum_{v_i : l_{v_i} = k_v} p(v_i) D_{KL}(p(\mathcal{D}|v_i) || p(\mathcal{D}|\hat{v}_{k_v})) + \sum_{v_{i_1}}^{V} \sum_{v_{i_2} \in \mathcal{M}_{v_{i_1}}} V(v_{i_1}, v_{i_2} \in \mathcal{M}_{v_{i_1}}) + \sum_{v_{i_1}}^{V} \sum_{v_{i_2} \in \mathcal{C}_{v_{i_1}}} V(v_{i_1}, v_{i_2} \in \mathcal{C}_{v_{i_1}})$.
    Word M-Step: update parameters
        $q^{(t+2)}(\hat{d}_{k_d}, \hat{v}_{k_v})$, $q^{(t+2)}(d_m|\hat{d}_{k_d})$ and $q^{(t+2)}(v_i|\hat{v}_{k_v})$.
    and compute $q^{(t+2)}(v_i|\hat{d}_{k_d})$.
    Compute cost $cost^{(t+2)}$ using Eq. (8) and compute $\delta = (cost^{(t+2)} - cost^{(t)})/cost^{(t)}$.
end while
```

Lemma. *The objective function (8) in the* HMRF^2*-ITCC model monotonically decreases to a local optimum.*

This lemma is easy to prove since the ICM algorithm decreases the nonnegative objective function (8) to a local optimum given a fixed q function. Then the update of q is monotonically decreasing as guaranteed by the theorem proven in [4].

3.3 Implementation Details

In this section, we discuss some issues in developing the constrained coclustering algorithm, which affects not only its performance (e.g., scalability) but also the final clustering results. In the following, we call our constrained ITCC approach CITCC, while using HMRF^2-ITCC for the internal model.

3.3.1 Initialization

Initialization is one of the most important issues in clustering since it directly affects the clustering quality. In this work, we utilize Kmeans to initialize the document and word clusters. Before we apply this method, however, we need to initialize Kmeans first. To make the Kmeans algorithm more stable for document and word clustering, we employ a farthest-first traversal method [29]. It aims to find K data points that are maximally separated from each other. In our implementation, at the beginning of initialization, we randomly select a data point as the first cluster center. Then, to identify a new center, we choose a data point that has not been selected previously using the following procedure. We first compare the distances between a candidate data point and all the previously selected centers, and record the minimal distance between this point and the centers. Then the candidate point with the largest minimum distance is selected as the new center. Finally, K centers are selected to initialize the cluster centers of Kmeans.

3.3.2 Data Structure for Matrix Operations

In our implementation, the original document and word co-occurrences, as well as the intermediate parameters such as $q^{(t+1)}(\hat{d}_{k_d}, \hat{v}_{k_v})$, $q^{(t+1)}(d_m|\hat{d}_{k_d})$, and $q^{(t+1)}(v_i|\hat{v}_{k_v})$ are all stored in matrices. The matrices adopt a row-style. Specifically, for a dense matrix, we use an array to store the row elements; for a sparse matrix, after comparing different hash table implementations in Java, we choose the COLT[1] hash to store the row elements. When implementing Kmeans, we store the norm of each row beforehand since the computation of the norm of each point is one of the biggest overheads in computing the euclidean distance or cosine similarity. When implementing NMF-based clustering methods, we carefully encode matrix multiplication since it affects the system performance the most. When implementing the ITCC-based methods, the computational sequence of KL-divergence is designed to make the computation faster. For example, when we compute the distances of

$$D_{KL}(p(\mathcal{V}|d_m)\|p(\mathcal{V}|\hat{d}_{k_d})) = \sum_i p(v_i|d_m)\log\frac{p(v_i|d_m)}{p(v_i|\hat{d}_{k_d})}, \quad (14)$$

in the for-loops, we follow the sequence of $\hat{d}_{k_d}$, d_m and then v_i since the matrix is stored in a row-style and the computation is much faster when we traverse the matrix row by row.

3.3.3 Constraints and ICM Inference

The constraints are stored in a set of hash tables. We choose this data structure to make the constraints symmetric for pairs of data points. It also makes it easier to extend the appointed constraints with neighborhood inference [22]. In this paper, however, we only use the original constraint set and do not infer new constraints, since the NE-based constraints and the WordNet-based constraints are noisy, which may not satisfy the consistency assumption [22]. Moreover, for the CITCC approach, in the ICM iteration, we cache all the KL-divergences $D_{KL}(p(\mathcal{V}|d_m)\|p(\mathcal{V}|\hat{d}_{k_d}))$ and $D_{KL}(p(\mathcal{D}|v_i)\|p(\mathcal{D}|\hat{v}_{k_v}))$, to avoid repeatedly computing them during the ICM algorithm.

The following remark points out the overall computational complexity of the CITCC algorithm under our implementation.

Remark 2. The time complexity of the alternating EM algorithm for the HMRF^2-ITCC model is $O((n_{nz} + (n_c * iter_{ICM})) \cdot (K_d + K_v)) \cdot iter_{AEM}$, where n_{nz} is the total number of nonzero elements in the document-word co-occurrence matrix, n_c is the constraint number, $iter_{ICM}$ is the ICM iteration number in the E-Step, K_d and K_v are the cluster numbers, and $iter_{AEM}$ is the iteration number of the alternating EM algorithm.

It has been shown that ICM, a greedy approximate inference method, is faster (by at least an order of magnitude) than other global approximate inference methods, e.g., loopy belief propagation and LP relaxation [30]. Moreover, when the number of constraints increases, ICM performs no worse than the global inference methods [30]. In this work, since the number of NE-based document constraints and the number of WordNet-based word constraints are quite large, we adopt the ICM algorithm to handle the large number of constraints.

4 Unsupervised Constraints

In this section, we show how to generate additional semantic constraints for clustering. Specifically, we introduce named-entity-based document constraints and WordNet relatedness-based word constraints using the following approaches.

4.1 Document Constraints

In practice, document constraints constructed based on human annotations are difficult to obtain. To cope with this problem, in this work, we propose new methods to derive "good but imperfect" constraints using information automatically extracted from either the content of a document (e.g., NE constraints) or existing knowledge sources (e.g., Wordnet constraints). For example, if two documents share the same people names such as "Barack Obama," "Sarah Palin," and "John McCain," then both documents are probably about US politics, thus both are likely to be in the same document cluster. Similarly, if two documents share the same organization names such as "AIG," "Lehman Brothers," and "Merrill Lynch," then both of them may be belong to the same document cluster about the financial markets. Consequently, the document must-link constraints can be constructed from the correlated named entities such as *person*, *location*, and *organization*. Specifically, if there are overlapping NEs in two documents and the number of overlapping NEs is larger than a predefined threshold, we may add a must-link to these documents.

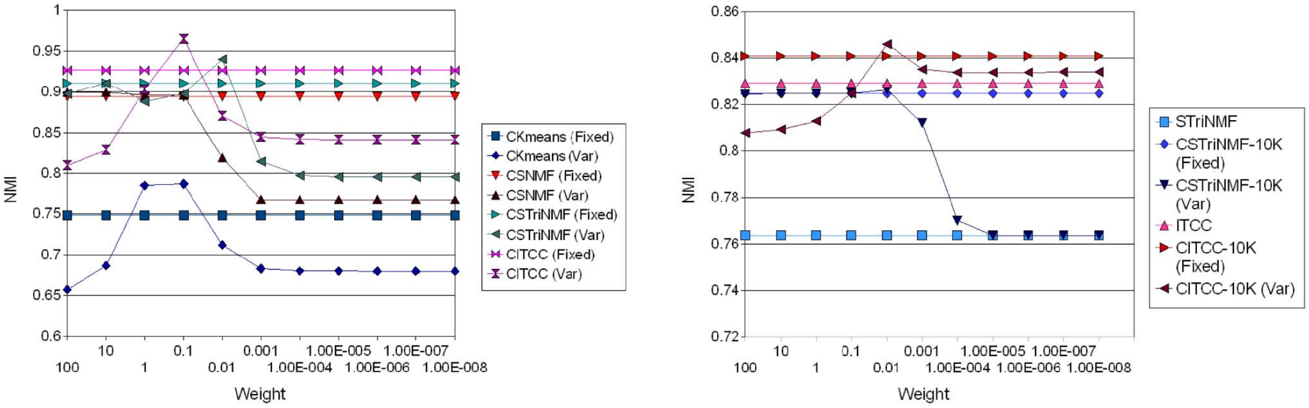
4.2 Word Constraints

Besides named-entity-based document constraints, it is possible to incorporate additional lexical constraints derived from existing knowledge sources to further improve clustering results. In our experiment, we leverage the information in WordNet, an online lexical database [31], to construct word constraints. Specifically, the semantic distance of two words can be computed based on their relationships in WordNet. Since we can construct word must-links based on semantic distances, for example, we can add a word must-link if the distance between two words is less than a threshold, additional lexical information can be seamlessly incorporated into the clustering algorithm to derive better word clusters. Moreover, since word knowledge can be transferred to the document side during coclustering, with additional word constraints, it is possible to further improve document clustering as well.

5 Experiments

To evaluate the effectiveness of the proposed CITCC approach, we ran our experiments using the 20-newsgroups

[1] http://acs.lbl.gov/software/colt/.



(a) Effects of document weights.

(b) Effects of word weights.

Fig. 2. Test with different parameters (alt.atheism and comp.graphics).

data set[2] and the SRAA data set.[3] The 20-newsgroups data set is a collection of approximately 20,000 newsgroups documents, partitioned evenly across 20 different newsgroups. The SRAA data set is a collection of 73,218 UseNet articles from four discussion groups: simulated autoracing, simulated aviation, real autos, and real aviation. These data sets are often used as benchmarks for classification as well as semi-supervised learning.

To evaluate the performance of CITCC against various clustering algorithms, we employed a widely used normalized mutual information (NMI)-based measure [32]. The NMI between two random variables X and Y is defined as

$$NMI(X,Y) = \frac{I(X;Y)}{\sqrt{H(X)H(Y)}},$$

where $I(X,Y)$ is the mutual information between X and Y. The entropies $H(X)$ and $H(Y)$ are used for normalizing the mutual information to be in the range of $[0,1]$. In practice, we estimate the NMI score [33] using the following formulation:

$$NMI = \frac{\sum_{s=1}^{K}\sum_{t=1}^{K} n_{s,t}\log\left(\frac{nn_{s,t}}{n_s\cdot n_t}\right)}{\sqrt{\left(\sum_s n_s \log\frac{n_s}{n}\right)\left(\sum_t n_t \log\frac{n_t}{n}\right)}}, \tag{15}$$

where n is the number of data samples, n_s and n_t denote the amount of the data in class s and cluster t, $n_{s,t}$ denotes the amount of data in both class s and cluster t. The NMI score is 1 if the clustering results match the category labeling perfectly and 0 if the clusters were obtained from a random partition. In general, the larger the scores are, the better the clustering results are.

5.1 20-Newsgroups Data Set

In this section, first we present some results on semi-supervised document clustering in which human annotated categories were used to derive document and word constraints. We want to demonstrate the performance of the algorithm in ideal situations in which constraints were constructed from human-provided clean data. Then we present a few experiments to examine the performance of the algorithm in unsupervised document clustering in which the automatically derived noisy word and document constraints were used.

5.1.1 Semi-Supervised Document Clustering

We first tested CITCC in a two-class document clustering setting where documents from two newsgroups (alt.atheism and comp.graphics) were used. There were 1,985 documents after removing the documents with less than five words. The vocabulary size was 11,149 after removing the words that appear in less than two documents. Each document was represented as a term frequency (TF) vector. In this experiment, we compared the performance of CITCC with that of several representative clustering algorithms such as Kmeans, constrained Kmeans (CKmeans) [22], Semi-NMF (SNMF) [34], constrained SNMF (CSNMF) [9], Tri-factorization of Semi-NMF (STriNMF) [9], constrained STriNMF (CSTriNMF) [9], and ITCC [4]. Among all the methods we tested, CKmeans, CSNMF, CSTriNMF, and CITCC are constrained clustering algorithms; STriNMF, CSTriNMF, ITCC, and CITCC are coclustering methods; and CSTriNMF and CITCC are constrained coclustering methods. For document constraints, we added a must-link between two documents if they shared the same category label. We also added a cannot-link if two documents come from different newsgroups. For the word constraints, after stop word removal, we counted the term frequencies of words in each newsgroup, and then chose the top 1,000 words in each group to randomly generate word pairs to add the word must-links. We did not use any word cannot-links in our experiments. In the following experiments, the document cluster number was set to 2, the ground-truth number.

We first tested how different model parameters affect the document clustering results. The tradeoff parameters a_{m_1,m_2} and $\bar{a}_{m_1,m_2}$ for constraints in (5) and (6) were set to numbers between $1E-8$ and 100. We also present the results with a fixed parameter value for each algorithm. For CITCC, the tradeoff parameters for documents a_{m_1,m_2} and $\bar{a}_{m_1,m_2}$ were empirically set to $1/\sqrt{M}$, where M is the document number. Similarly, the tradeoff parameters for word constraints were set to be $1/\sqrt{V}$, where V is the word number. In addition, the tradeoff parameters in Kmeans was set to $1/\sqrt{M}$. The tradeoff parameters in SNMF and STriNMF were set to 0.5. Fig. 2 compares the clustering performance with different

[2] http://people.csail.mit.edu/~jrennie/20Newsgroups/.

[3] http://www.cs.umass.edu/~mccallum/code-data.html.

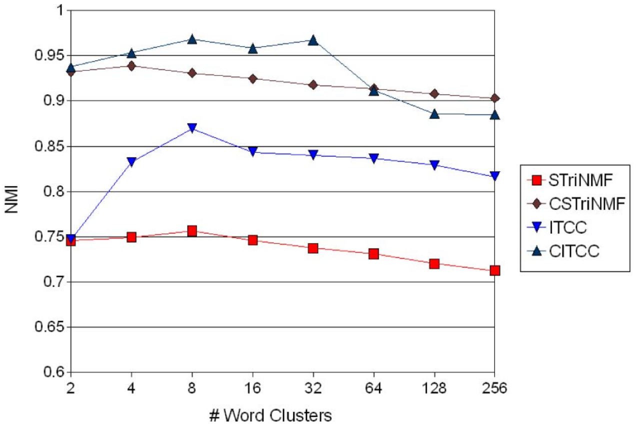


Fig. 3. Effect of word cluster numbers (alt.atheism and comp.graphics).

parameters for CITCC as well as other constrained methods. Since the clustering results with fixed parameters were comparable to the best results with varied parameters, in the following experiments, we used fixed parameters.

We also evaluated the effect of different word cluster numbers on document clustering performance. Fig. 3 shows the results of two coclustering algorithms CSTriNMF and CITCC with different word cluster numbers. It is shown that for this data set, more word clusters may not result in improved document clustering results when a sufficient number of word clusters is reached. For example, after reaching 8 for ITCC and 32 for CITCC, the NMI scores of ITCC and CITCC actually decreased when the number of word clusters further increased. In the rest of the experiments, we fixed the word cluster number to be twice the document cluster number.

We then varied the number of document and word constraints in each experiment by randomly selecting a fixed number of constraints from all possible must-links and cannot-links to investigate their impact on clustering performance. Figs. 4 shows the experiment results. Each x-axis represents the number of document constraints used in each experiment and y-axis the average NMI of five random trials. As shown in Fig. 4a, among all the methods we tested, CITCC consistently performed the best. It outperformed the nonconstrained coclustering algorithm ITCC significantly. Its clustering performance was also better than all the 1D clustering algorithms, regardless of the number of document constraints used. Moreover, it was more effective than a known constrained coclustering algorithm CSTriNMF. The number of document constraints seems to have a significant impact on the performance. The more document constraints we added, the better the clustering results were.

In addition, to evaluate the effect of the number of word constraints on the constrained coclustering performance, we evaluated three versions of the CITCC and CSTriNMF algorithms 1) CITCC and CSTriNMF: with only document constraints and no word constraints, 2) CITCC (5K) and CSTriNMF (5K): with document constraints plus 5,000 word constraints, and 3) CITCC (10K) and CSTriNMF (10K): with document constraints plus 10,000 word constraints. As shown in Fig. 4b, in general, more word constraints resulted in better clustering performance. The impact of the word constraints, however, was not as strong as that of the document constraints.

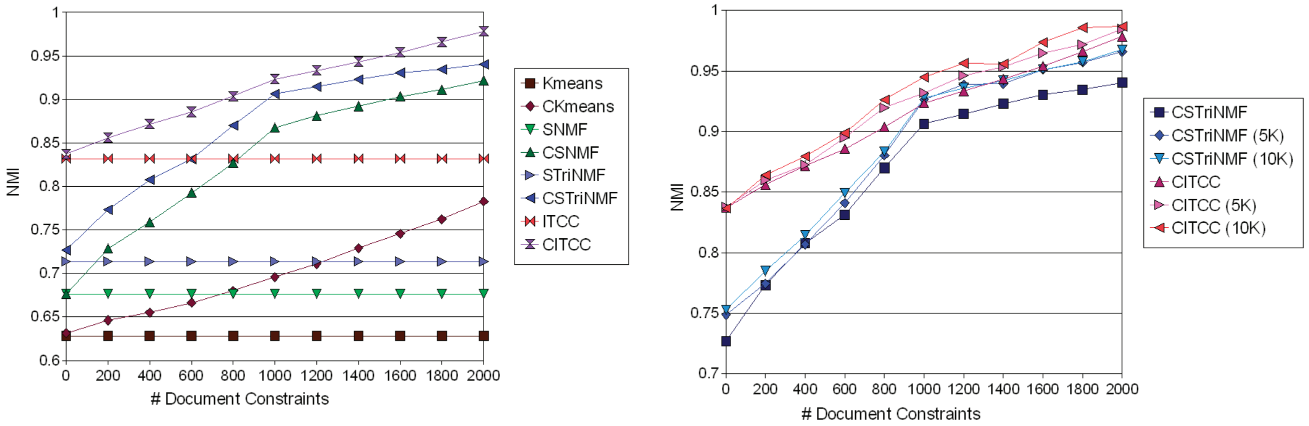
5.1.2 Unsupervised Document Constraints

In this experiment, the unsupervised document must-link constraints were automatically derived from NEs such as *person*, *location*, and *organization*. We used a state-of-the-art NE recognizer[4] to find NEs in these documents. If there were some overlapping NEs in two documents and the number of overlapping NEs was larger than a threshold, then we added a must-link between these documents.

Before we present the evaluation results, we first examine the quality of the NE constraints. Table 1 shows the related statistics. Here,"*#NEs (mean(std))*" represents the average number and standard deviation of overlapping NEs in two documents that had a must-link; "*#Must-links*" is the total number of the added must-links based on overlapping NEs; and "*Correct Percentage*" indicates the percentage of all the correct must-links that were added, the percentage of correct ones (the associated documents belong to the same newsgroup). "*Similarity (mean(std))*" indicates the average cosine similarity and the standard deviation among the documents with must-links. As shown in Table 1, increasing the number of overlapping NEs required to add a must-link decreased the number of total must-links added, increased the accuracy of the derived must-links, as well as increased the document similarities with must-links. Moreover, after the minimum number of required overlapping NEs reached 2, the quality of the derived must-links was quite high (95.6 percent). After that, the accuracy improvement became less significant, while the total number of must-links added continued to decrease significantly.

To demonstrate how different methods utilized the additional NE constraints, we first tested them under the two-class setting (alt.atheism versus comp.graphics). Specifically, we compared the performance of the constrained version of each algorithm with that of the nonconstrained version. The comparison results shown in Table 2 are the means and the standard deviations of the NMI scores across 30 random runs. The column *no constraint* represents the performance of the nonconstrained version of each method (i.e., Kmeans, SNMF, STriNMF, and ITCC). As shown in Table 2, among all the methods we tested, CITCC achieved the best performance (0.843). Under the non-parametric Mann-Whitney U test,[5] CITCC performed significantly better than ITCC. Moreover, for each method, the constrained version was able to take advantage of the additional NE constraints to improve its clustering performance over its nonconstrained version. In addition, if a must-link was added when at least one overlapping NE was detected, the performance of the constrained version was worse than that of the nonconstrained version. This seems to suggest that if we define the must-link constraints loosely (e.g., only at least 1 overlapping NE is required to add a must-link), the additional NE constraints were too noisy for a constrained clustering system to achieve good performance. Furthermore, the automatically derived NE constraints were not as effective as the constraints constructed from category labels provided by human. To investigate the reason for this, we computed the average similarity of documents with must-links. As shown in

[4] http://nlp.stanford.edu/software/CRF-NER.shtml.
[5] http://en.wikipedia.org/wiki/Mann-Whitney_U.



(a) Effects of document constraints.

(b) Effects of word constraints.

Fig. 4. Semi-supervised document clustering (alt.atheism and comp.graphics).

TABLE 1
Statistics of NEs Constraints under a Two-Class Setting (alt.atheism versus comp.graphics)

#NEs	≥1	≥2	≥3	≥4	≥5	≥6
#NEs (mean(std))	1.33(3.11)	2.99(7.35)	4.59(11.71)	5.96(15.86)	10.57(28.57)	14.81(37.41)
#Must-links	35,156	5,938	2,271	1,219	363	206
Correct Percentage	89.5%	95.6%	97.4%	98.4%	96.4%	97.1%
Similarity (mean(std))	0.151(0.568)	0.266(0.323)	0.332(0.222)	0.366(0.168)	0.469(0.107)	0.496(0.084)

The average similarity and standard deviation of all documents in the two newsgroups is 0.033(2.212).

Table 1, the average document similarity increased as more overlapping NEs were required. This implies that the information encoded in the NE constraints may be mostly redundant to that encoded in the document similarity metric. In contrast, human-provided category labels may be less redundant and thus provide additional information (e.g., topic information) to guide the clustering towards the ground truth.

We also tested the algorithms under all 190 two-class clustering conditions for all 20 newsgroups. The number of overlapping NEs was set to be at least 3. The resulting number of must-link constraints and the correct percentage values are presented in Fig. 5. This data shows that, overall, the derived NE constraints were quite accurate. In most cases, over 95 percent of the derived NE constraints were correct. In addition, as shown in Figs. 6a and 6d, for all 190 cases, the NE constraints can help improve the clustering performance for both CKmeans and CITCC rather consistently. Moreover, for CITCC, the dots are concentrated at the upper right corner, thus indicating consistently high performance for both ITCC and CITCC. For the results in Figs. 6b and 6c, however, the usefulness of NE constraints for CSNMF and CSTriNMF are less consistent. Many times the additional constraints actually hurt the performance. We speculate that this may be due to two factors. First, as shown in Table 2, the clustering results were quite sensitive to the number of overlapping NEs used in constructing the must-links, especially for CSNMF and CSTriNMF. Since we set the least number of overlapping NEs required to add a must-link to be the same for all the systems and across all the 190 test conditions, the results for CSNMF and CSTriNMF may not always be optimal. Second, we used the same tradeoff parameters in all experiments, which may not be optimal for CSNMF and CSTriNMF.

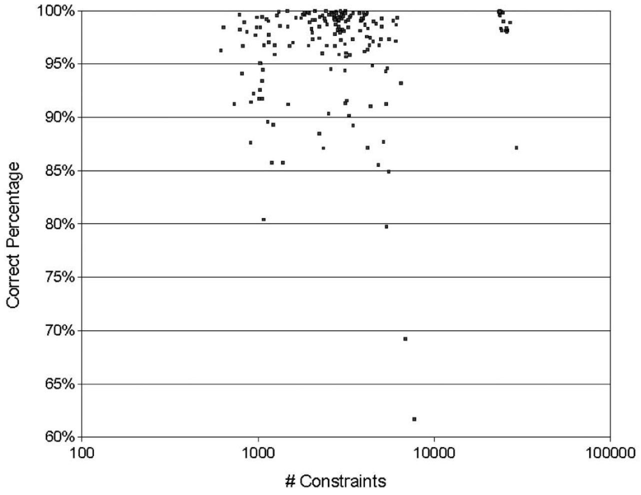
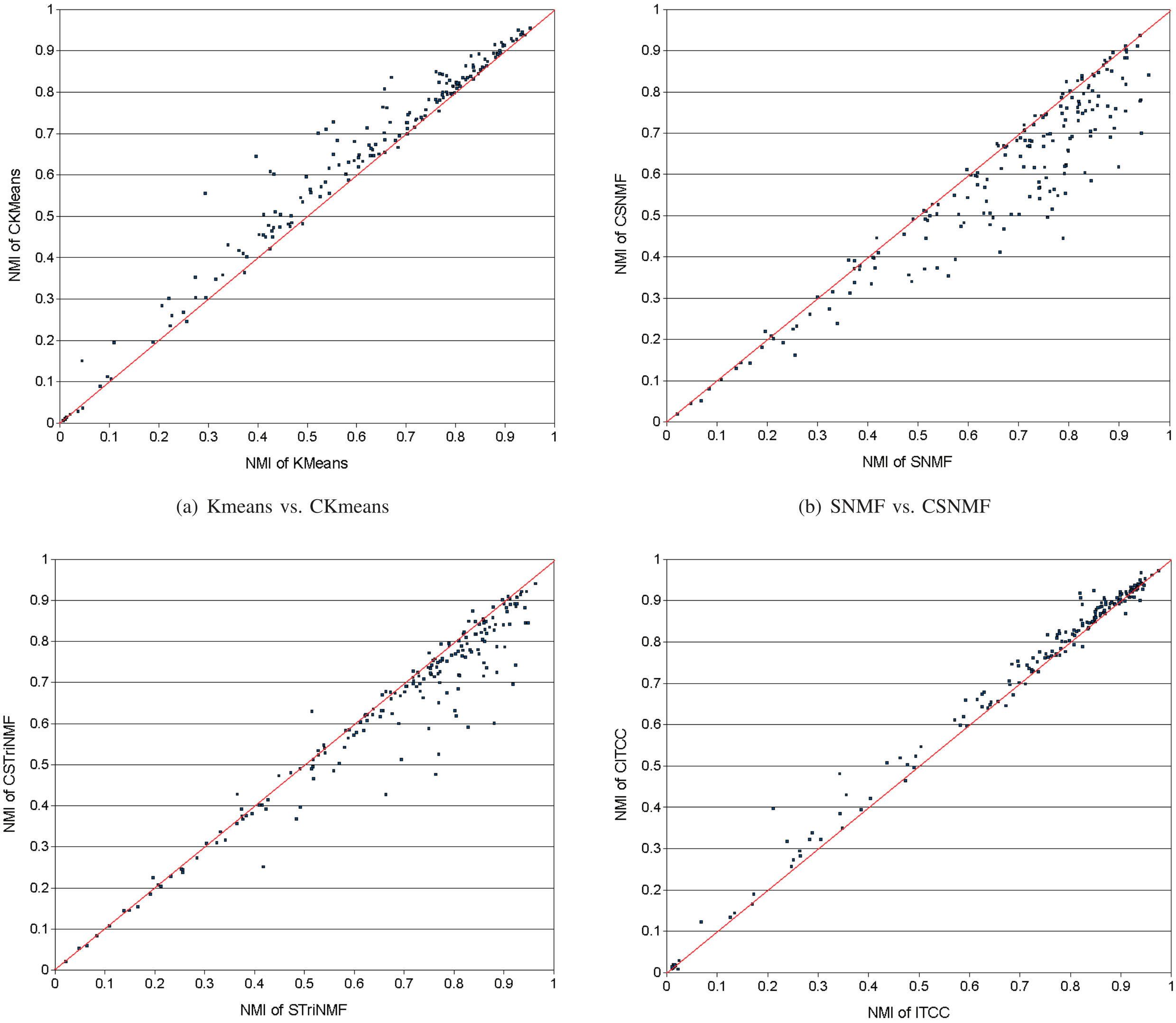


Fig. 5. # Constraints versus correct percentage (all 190 two-class problems in 20-newsgroups data).

5.1.3 Unsupervised Word Constraints

In this experiment, the unsupervised word must-link constraints were derived based on WordNet[6] [31], an online lexical resource widely used in the natural language processing (NLP) and text mining community. WordNet groups English words, primarily nouns and verbs, into sets of synonyms called synsets; it provides short, general definitions, and records the various semantic relations between these synonym sets, such as

[6] http://wordnet.princeton.edu.



(a) Kmeans vs. CKmeans

(b) SNMF vs. CSNMF

(c) STriNMF vs. CSTriNMF

(d) ITCC vs. CITCC

Fig. 6. NE constraints results on 190 two-class clustering problems in the 20-newsgroups data set.

TABLE 2
Comparison of NE Constraints for Different Algorithms under a Two-Class Setting (alt.atheism versus comp.graphics)

#NEs	no constraint	$\geq$1	$\geq$2	$\geq$3	$\geq$4	$\geq$5	$\geq$6
CKmeans	0.648(0.012)	0.651(0.007)(+)	**0.697(0.011)**(+)	0.674(0.008)(+)	0.666(0.008)(+)	0.659(0.008)(+)	0.656(0.009)(+)
CSNMF	0.714(0.011)	0.448(0.005)(-)	0.476(0.004)(-)	0.509(0.008)(-)	0.529(0.004)(-)	0.645(0.002)(-)	**0.757(0.002)**(+)
CSTriNMF	**0.750(0.028)**	0.543(0.175)(-)	0.550(0.113)(-)	0.669(0.142)(-)	0.742(0.191)	0.678(0.145)(-)	0.512(0.024)(-)
CITCC	0.809(0.021)	0.353(0.089)(-)	**0.843(0.014)**(+)	0.827(0.015)(+)	0.818(0.014)(+)	0.819(0.014)(+)	0.818(0.016)(+)

"$+/-$" represent the statistical significance of the difference between the constrained version and unconstrained version of a method based on Mann-Whitney U test. "$+$" means the constrained version is significant better than unconstrained version with $p < 0.05$ and "-" means the unconstrained version is significant better than constrained version with $p < 0.05$.

hypernyms, hyponyms, and meronyms. Both nouns and verbs are organized into hierarchies, defined by hypernym relations. The semantic relatedness between words can be measured based on the word hierarchies in the Wordnet. For example, the shortest path between two words can be used to measure the semantic distance between them. As a result, we can utilize the semantic relatedness between words to derive word must-link constraints. For simplification, in this experiment, we only considered nouns.

In the experiments, to obtain the word must-links, we selected semantically related words based on their WordNet distance. The semantic distance between two words is computed as follows:[7]

- Locate the common parent cp of words w_1 and w_2. If one exists, check each sense of each lemma; if one is not found, return 1.0.

[7] http://www.rednoise.org/rita/wordnet.

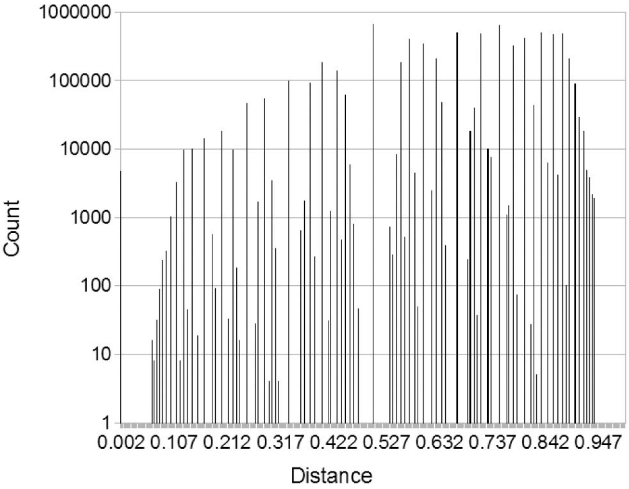


Fig. 7. Word constraints with different WordNet distances.

- Calculate the length of the shortest path from either word to cp: $sp(w_1, w_2)$.
- Calculate the length of the path from cp to the root of ontology $len(cp, root)$.
- Calculate the distance between the two words and return

$$Dist(w_1, w_2) = \frac{sp(w_1, w_2)}{sp(w_1, w_2) + len(cp, root)}. \qquad (16)$$

If the semantic distance between two words was less than a predefined threshold, we added a must-link between them. We evaluated the effectiveness of WordNet-based must-links using the data from two newsgroups (alt.atheism and comp.graphics). There were 4,680 nouns among all the 11,149 vocabulary words. As shown in Fig. 7. the number of constraints increased exponentially when the threshold value was increased. We further tested the clustering results of CSTriNMF and CITCC by varying the distance thresholds from 0.05 to 0.5. The NMI values as well as the numbers of word constraints are presented in Table 3 with different distance thresholds. We can see that small distance values seem to improve the document clustering results since they will result in more reliable word must-links. In contrast, large threshold values often introduce noise which makes the clustering performance worse. CSTriNMF performed significantly better than STriNMF. Although CITCC in average performed better than ITCC, their difference, however, was not statistically significant. This may be because the semantic relatedness information in WordNet is very noisy. For example, the word "bank" can be a financial institute or a "river bank." In terms of semantic relatedness, the distance between "bank" as a financial institute and "money" is small while the distance between "bank" as in "river bank" and "money" is big. Without word-sense disambiguation, it is difficult to accurately compute semantic relatedness.

5.2 SRAA Data Set

In the SRAA data set, there are originally four classes, including simulated autoracing, simulated aviation, real autos, and real aviation. The data set contains 73,218 articles in total, and we sampled 5 percent of the data to derive a test collection whose size is comparable to the 20-newsgroups data set. After performing stop word and short document removal, we obtained 3,603 documents for clustering. The vocabulary size was 10,460. The document cluster number was set to 4, the ground truth number, and the word cluster number was empirically set to 8. All parameters were the same as those used in the 20-newsgroups experiments. Here, we focus on verifying the effect of the NE-based document constraints and WordNet-based word constraints for unsupervised document clustering.

5.2.1 Unsupervised Document Constraints

We tested the SRAA data set using the same parameters as those used in Section 5.1.2. Table 4 shows the statistics of the corresponding NE constraints. Similar to the results on the 20-newsgroups data, when we increased the number of overlapping NEs, the number of must-links decreased, while the precision of the derived must-links increased. It seems that the number of overlapping NEs in the SRAA data set was much less than that in the 20-newsgroups data. For example, when we set the threshold to 11, there was only 1 must-link added to the documents. This may be due to the nature of the data set. In general, news groups have more information on people, locations and organizations. The clustering results of the SRAA data set is in Table 5. As shown in the results, CITCC outperformed the all the other methods. It is also shown that the constraints were noisy when the overlapping NE threshold was set to 1.

TABLE 3
Comparison of Different Algorithms with Different WordNet Distance Thresholds for the Two-Newsgroups (alt.atheism versus comp.graphics) Data Set

Threshold	no constraint	0.05	0.1	0.15	0.2	0.25
# Constraints	4,704	5,406	29,580	44,627	118,900	175,161
CSTriNMF	0.715(0.237)	**0.764(0.254)**(+)	0.725(0.240)	0.683(0.226)(-)	0.682(0.225)(-)	0.681(0.225)(-)
CITCC	0.837(0.025)	0.845(0.025)	0.846(0.026)	**0.848(0.027)**	0.833(0.032)	0.775(0.037)(-)

Threshold	0.3	0.35	0.4	0.45	0.5
# Constraints	276,277	371,197	756,514	1,417,557	0
CSTriNMF	0.681(0.225)(-)	0.681(0.225)(-)	0.681(0.225)(-)	0.681(0.225)(-)	0.681(0.225)(-)
CITCC	0.770(0.038)(-)	0.766(0.028)(-)	0.755(0.026)(-)	0.731(0.026)(-)	0.768(0.023)(-)

"$+/-$" represent the statistical significance of the difference between the constrained version and unconstrained version of a method based on Mann-Whitney U test. "+" means the constrained version is significant better than unconstrained version with $p < 0.05$ and "$-$" means the unconstrained version is significant better than constrained version with $p < 0.05$.

TABLE 4
Statistics of NEs Constraints of the SRAA Data Set

#NEs	≥1	≥2	≥3	≥4	≥5	≥6
#NEs (mean(std))	1.10(0.33)	2.05(0.31)	3.27(0.83)	4.81(1.37)	5.83(1.55)	7.00(1.89)
#Must-links	85,311	8,211	363	54	24	10
Correct Percentage	79.3%	80.7%	93.6%	92.5%	95.8%	100%

TABLE 5
Comparison of NE Constrains for Different Algorithms on the SRAA Data Set

#NEs	no constraint	≥1	≥2	≥3	≥4	≥5	≥6
CKmeans	0.378(0.060)	0.338(0.033)(-)	0.370(0.056)	**0.386(0.062)**	0.385(0.061)	0.385(0.061)	0.385(0.061)
CSNMF	0.341(0.042)	0.142(0.051)(-)	0.178(0.029)(-)	0.322(0.056)(-)	0.293(0.038)(-)	**0.372(0.057)**(+)	0.310(0.048)(-)
CSTriNMF	0.346(0.043)	0.190(0.037)(-)	0.169(0.038)(-)	0.212(0.102)(-)	0.327(0.099)(-)	**0.369(0.118)**(+)	**0.369(0.111)**(+)
CITCC	0.425(0.034)	0.330(0.026)(-)	0.390(0.037)(-)	**0.432(0.042)**	0.428(0.041)	0.429(0.041)	0.429(0.041)

"$+/-$" represent the statistical significance of the difference between the constrained version and unconstrained version of a method based on Mann-Whitney U test. "+" means the constrained version is significant better than unconstrained version with $p < 0.05$ and "$-$" means the unconstrained version is significant better than constrained version with $p < 0.05$.

TABLE 6
Comparison of Different Algorithms with Different WordNet Distance Thresholds for the SRAA Data Set

Threshold	no constraint	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
# Constraints	4,048	4,696	25,054	37,410	95,453	147,570	228,378	323,085	649,089	1,173,297	0
CSTriNMF	**0.311(0.032)**	0.313(0.033)	0.311(0.033)	0.310(0.033)	0.310(0.033)	0.311(0.033)	0.311(0.033)	0.311(0.033)	0.311(0.033)	0.311(0.033)	0.311(0.033)
CITCC	0.402(0.044)	**0.411(0.044)**	0.411(0.047)	0.396(0.085)	0.372(0.072)	0.376(0.090)	0.360(0.082)(-)	0.365(0.078)(-)	0.355(0.075)(-)	0.359(0.075)(-)	0.378(0.058)

"$+/-$" represent the statistical significance of the difference between the constrained version and unconstrained version of a method based on Mann-Whitney U test. "+" means the constrained version is significant better than unconstrained version with $p < 0.05$ and "$-$" means the unconstrained version is significant better than constrained version with $p < 0.05$.

5.2.2 Unsupervised Word Constraints

To test the effects of unsupervised word constraints using the SRAA data set, we focused on 4,327 nouns among the 10,460 vocabulary words. The clustering results with different threshold values are shown in Table 6. We can see that the number of constraints also increased significantly when we increased the threshold of WordNet distance. The clustering results were better when the threshold was smaller, e.g., smaller than 0.1. Similar to the results obtained from the 20-newsgroups data, when the threshold was increased, the derived constraints also became more noisy, which hurt the performance of constrained clustering.

6 Conclusion and Future Work

In this paper, we have demonstrated how to construct various document and word constraints and apply them to the constrained coclustering process. We proposed a novel constrained coclustering approach that automatically incorporates various word and document constraints into information-theoretic coclustering. Our evaluations on two benchmark data sets demonstrated the effectiveness of the proposed method for clustering textual documents. Furthermore, our algorithm consistently outperformed all the tested constrained clustering and coclustering methods under different conditions.

There are several directions for future research. Our investigation of unsupervised constraints is still preliminary. We will further investigate whether better text features that can be automatically derived by using natural language processing or information extraction tools. We are also interested in applying CITCC to other text analysis applications such as visual text summarization.

Acknowledgments

The authors would like to thank Fei Wang for insightful discussions and help on implementing the constrained NMF algorithms. They would like to thank Keith Houck and Mat Callcut for their kind help on proofreading our paper.

References

[1] A. Jain, M. Murty, and P. Flynn, "Data Clustering: A Review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264-323, 1999.

[2] Y. Cheng and G.M. Church, "Biclustering of Expression Data," *Proc. Int'l System for Molecular Biology Conf. (ISMB)*, pp. 93-103, 2000.

[3] I.S. Dhillon, "Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning," *Proc. Seventh ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD)*, pp. 269-274, 2001.

[4] I.S. Dhillon, S. Mallela, and D.S. Modha, "Information-Theoretic Co-Clustering," *Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD)*, pp. 89-98, 2003.

[5] H. Cho, I.S. Dhillon, Y. Guan, and S. Sra, "Minimum Sum-Squared Residue Co-Clustering of Gene Expression Data," *Proc. Fourth SIAM Int'l Conf. Data. Mining (SDM)*, 2004.

[6] *Semi-Supervised Learning*, O. Chapelle, B. Schölkopf, and A. Zien, eds. MIT Press, http://www.kyb.tuebingen.mpg.de/ssl-book, 2006.

[7] S. Basu, I. Davidson, and K. Wagstaff, *Constrained Clustering: Advances in Algorithms, Theory, and Applications.* Chapman & Hall/CRC, 2008.

[8] R.G. Pensa and J.-F. Boulicaut, "Constrained Co-Clustering of Gene Expression Data," *Proc. SIAM Int'l Conf. Data Mining (SDM),* pp. 25-36, 2008.

[9] F. Wang, T. Li, and C. Zhang, "Semi-Supervised Clustering via Matrix Factorization," *Proc. SIAM Int'l Conf. Data. Mining (SDM),* pp. 1-12, 2008.

[10] Y. Chen, L. Wang, and M. Dong, "Non-Negative Matrix Factorization for Semi-Supervised Heterogeneous Data Co-Clustering," *IEEE Trans. Knowledge and Data Eng.,* vol. 22, no. 10, pp. 1459-1474, Oct. 2010.

[11] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D.S. Modha, "A Generalized Maximum Entropy Approach to Bregman Co-Clustering and Matrix Approximation," *J. Machine Learning Research,* vol. 8, pp. 1919-1986, 2007.

[12] Y. Song, S. Pan, S. Liu, F. Wei, M.X. Zhou, and W. Qian, "Constrained Co-Clustering for Textual Documents," *Proc. Conf. Artificial Intelligence (AAAI),* 2010.

[13] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal Nonnegative Matrix T-Factorizations for Clustering," *Proc. 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,* pp. 126-135, 2006.

[14] H. Shan and A. Banerjee, "Bayesian Co-Clustering," *Proc. IEEE Eight Int'l Conf. Data Mining (ICDM),* pp. 530-539, 2008.

[15] P. Wang, C. Domeniconi, and K.B. Laskey, "Latent Dirichlet Bayesian Co-Clustering," *Proc. European Conf. Machine Learning and Knowledge Discovery in Databases (ECML/PKDD),* pp. 522-537, 2009.

[16] K. Nigam, A.K. McCallum, S. Thrun, and T.M. Mitchell, "Text Classification from Labeled and Unlabeled Documents using EM," *Machine Learning,* vol. 39, no. 2/3, pp. 103-134, 2000.

[17] S. Basu, A. Banerjee, and R.J. Mooney, "Semi-Supervised Clustering by Seeding," *Proc. 19th Int'l Conf. Machine Learning (ICML),* pp. 27-34, 2002.

[18] F.G. Cozman, I. Cohen, and M.C. Cirelo, "Semi-Supervised Learning of Mixture Models," *Proc. Int'l Conf. Machine Learning (ICML),* pp. 99-106, 2003.

[19] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, "Constrained K-Means Clustering with Background Knowledge," *Proc. 18th Int'l Conf. Machine Learning (ICML),* pp. 577-584, 2001.

[20] E.P. Xing, A.Y. Ng, M.I. Jordan, and S.J. Russell, "Distance Metric Learning with Application to Clustering with Side-Information," *Proc. Advances in Neural Information Processing Systems Conf.,* pp. 505-512, 2002.

[21] M. Bilenko, S. Basu, and R.J. Mooney, "Integrating Constraints and Metric Learning in Semi-Supervised Clustering," *Proc. 21st Int'l Conf. Machine Learning (ICML),* pp. 81-88, 2004.

[22] S. Basu, M. Bilenko, and R.J. Mooney, "A Probabilistic Framework for Semi-Supervised Clustering," *Proc. SIGKDD,* pp. 59-68, 2004.

[23] Z. Lu and T.K. Leen, "Penalized Probabilistic Clustering," *Neural Computation,* vol. 19, no. 6, pp. 1528-1567, 2007.

[24] W. Dai, G.-R. Xue, Q. Yang, and Y. Yu, "Co-Clustering Based Classification for Out-of-Domain Documents," *Proc. 13th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,* pp. 210-219, 2007.

[25] X. Shi, W. Fan, and P.S. Yu, "Efficient Semi-Supervised Spectral Co-Clustering with Constraints," *Proc. IEEE 10th Int'l Conf. Data Mining(ICDM),* pp. 1043-1048, 2010.

[26] T. Li, C. Ding, Y. Zhang, and B. Shao, "Knowledge Transformation from Word Space to Document Space," *Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR),* pp. 187-194, 2008.

[27] T. Li, Y. Zhang, and V. Sindhwani, "A Non-Negative Matrix Tri-Factorization Approach to Sentiment Classification with Lexical Prior Knowledge," *Proc. Joint Conf. 47th Ann. Meeting of the ACL and the Fourth Int'l Joint Conf. Natural Language Processing of the AFNLP (ACL-IJCNLP),* pp. 244-252, 2009.

[28] T. Yang, R. Jin, and A.K. Jain, "Learning from Noisy Side Information by Generalized Maximum Entropy Model," *Proc. Int'l Conf. Machine Learning (ICML),* pp. 1199-1206, 2010.

[29] D.S. Hochbaum and D.B. Shmoys, "A Best Possible Heuristic for the K-Center Problem," *Math. Operations Research,* vol. 10, no. 2, pp. 180-184, 1985.

[30] M. Bilenko and S. Basu, "A Comparison of Inference Techniques for Semi-Supervised Clustering with Hidden Markov Random Fields," *Proc. ICML Workshop Statistical Relational Learning and Its Connections to Other Fields (SRL '04),* 2004.

[31] G.A. Miller, "Wordnet: A Lexical Database for English," *Comm. ACM,* vol. 38, pp. 39-41, 1995.

[32] A. Strehl and J. Ghosh, "Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions," *J. Machine Learning Research,* vol. 3, pp. 583-617, 2002.

[33] S. Zhong and J. Ghosh, "Generative Model-Based Clustering of Documents: A Comparative Study," *Knowledge and Information Systems,* vol. 8, pp. 374-384, 2005.

[34] C.H.Q. Ding, T. Li, and M.I. Jordan, "Convex and Semi-Nonnegative Matrix Factorizations," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 32, no. 1, pp. 45-55, Jan. 2010.

Yangqiu Song received the BEng and PhD degrees from the Department of Automation, Tsinghua University, China, in July, 2003 and January, 2009, respectively. He joined Microsoft Research Asia in November, 2010 as an associate researcher. Before that, he was a researcher at IBM Research, China. His research interests include machine learning algorithms with applications to knowledge engineering, information retrieval, and visualization. He is a member of the IEEE.

Shimei Pan received the PhD degree in computer science from Columbia University in 2002. She is a research staff member at IBM T. J. Watson Research Center in New York. Her research areas include intelligent user interfaces, natural language processing, interactive visual and text analytics, spoken dialogue systems, multimodal query understanding, and multimedia presentation generation. In 2005, she was the chair of the IBM Natural Language Processing Professional Interest Community (PIC). She has published more than 30 papers in major AI, NLP, and intelligence user interfaces conferences. She also served on various US National Science Foundation (NSF) panels and conference program committees such as IUI, IJCAI, ACL, EMNLP, and ACM Recommender Systems. She was also the recipient of the IBM Outstanding Innovation Award in 2005, IBM invention Achievement Award in 2006, and IBM Research Division Award in 2008.

Shixia Liu received the BS and MS degrees in computational mathematics from Harbin Institute of Technology, and the PhD degree in computer aided design and computer graphics from Tsinghua University. She is a lead researcher in the Internet Graphics Group at Microsoft Research Asia. Her research interest mainly focuses on interactive, visual text analytics and interactive, visual graph analytics. Before she joined MSRA, she worked as a research staff member and research manager at IBM China Research Lab, where she managed the Departments of Smart Visual Analytics and User Experience. She is a member of the IEEE.

Furu Wei received the BSc and PhD degrees from the Department of Computer Science of Wuhan University, China, in 2004 and June 2009, respectively. Currently, he is an associate researcher in Natural Language Computing group at Microsoft Research Asia, Beijing, China. Before joining MSRA-NLC in November 2010, he has been a staff researcher at IBM Research, China (IBM CRL) since July 2009. His research interests include natural language processing, information retrieval, and machine learning.

Michelle X. Zhou received the PhD degree in computer science from Columbia University in 1999 and named an ACM Distinguished Scientist in 2009. She manages the User Systems and Experience Research (USER) group at the IBM Almaden Research Center. Her expertise is in the interdisciplinary areas of intelligent user interaction and analytics-driven social computing. She has published more than 70 peer-reviewed articles in top conferences and journals. Her work received the best paper award at ACM IUI 2005 and was nominated for the best paper award at CIKM 2009. She was the general conference cochair for ACM IUI 2007 and is the technical program cochair for ACM MM 2009 and IUI 2010. She serves on the editorial board of three technical journals: *ACM TOMCCAP*, *ACM TIST*, and *ACM TIIS*. She is a senior member of the IEEE.

Weihong Qian received the BS and MS degrees from Zhejiang University, majored in computer science and technology. She is a staff researcher in the IBM China Research Lab (CRL). Her research interests include interactive visual text analysis, interactive visual social network analysis, simple visualization, text analytics, embedded system, etc.