Constrained Co-clustering for Textual Documents

Yangqiu Song [†] **Shimei Pan** [‡] **Shixia Liu** [†] **Furu Wei** [†] **Michelle X. Zhou** [§] **Weihong Qian** [†]
[†]{yqsong,liusx,weifuru,qianwh}@cn.ibm.com; [‡]shimei@us.ibm.com; [§]mzhou@us.ibm.com
[†]IBM Research – China, Beijing, China
[‡]IBM Research – T. J. Watson Center, Hawthorne, NY, USA
[§]IBM Research – Almaden Center, San Jose, CA, USA

Abstract

In this paper, we present a constrained co-clustering approach for clustering textual documents. Our approach combines the benefits of information-theoretic co-clustering and constrained clustering. We use a two-sided hidden Markov random field (HMRF) to model both the document and word constraints. We also develop an alternating expectation maximization (EM) algorithm to optimize the constrained co-clustering model. We have conducted two sets of experiments on a benchmark data set: (1) using human-provided category labels to derive document and word constraints for semi-supervised document clustering, and (2) using automatically extracted named entities to derive document constraints for unsupervised document clustering. Compared to several representative constrained clustering and co-clustering approaches, our approach is shown to be more effective for high-dimensional, sparse text data.

Introduction

Clustering is a popular machine learning method commonly used in exploratory text analysis. Numerous clustering methods have been proposed previously and many of them focus on one-dimensional clustering (Jain, Murty, and Flynn 1999). In practice, it is often desirable to co-cluster documents and words simultaneously by exploiting the co-occurrence among them. It has been shown that co-clustering is more effective than clustering along a single dimension in many applications (Cheng and Church 2000; Dhillon 2001; Dhillon, Mallela, and Modha 2003; Cho et al. 2004).

Typical clustering algorithms are unsupervised. It is also preferable for a clustering or co-clustering algorithm to be able to take prior information about clusters, such as human-specified category labels, into consideration. Constrained clustering was proposed as a solution to the problem. It leverages additional constraints derived from human-annotated instances such as pre-defined semantic categories of documents. However, most of the existing works on constrained clustering have focused on one-dimensional clustering (Basu, Davidson, and Wagstaff 2008).

To combine the benefits of both co-clustering and constrained clustering, in this paper, we propose an approach called constrained information-theoretic co-clustering (CITCC). It incorporates constraints into the information theoretic co-clustering (ITCC) framework (Dhillon, Mallela, and Modha 2003) using a two-sided hidden Markov random field (HMRF) regularization. We also develop an alternating expectation maximization (EM) algorithm to optimize the model. Consequently, CITCC can simultaneously cluster two sets of discrete random variables such as words and documents under the constraints from both variables.

Following a review of existing works, we describe the details of the proposed CITCC approach. In addition, to evaluate the effectiveness of the proposed method, we use a benchmark data set to test its performance in two different experimental settings: (1) as a semi-supervised approach with human-provided category labels, (2) as an unsupervised method that incorporates additional application-specific constraints such as the named entity overlapping constraints, for document clustering.

Related Work

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

There are two types of semi-supervised clustering methods: semi-supervised clustering with labeled seeding points (Basu, Banerjee, and Mooney 2002; Nigam et al. 2000) and semi-supervised clustering with labeled constraints (Wagstaff et al. 2001; Xing et al. 2002; Bilenko, Basu, and Mooney 2004; Basu, Bilenko, and Mooney 2004; Lu and Leen 2007). Constraint-based clustering methods often use pairwise constraints such as "must-links" and "cannot-links" to enhance unsupervised clustering algorithms. These constraints are also called "side-information". Similar to constraint-based clustering methods, we also use pairwise must-links and cannot-links to encode prior knowledge. While all the above semi-supervised methods are applicable to one-dimensional clustering, we focus on extending these techniques to co-clustering.

Most co-clustering 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 spectral graph theory is used to solve the partition problem (Dhillon

2001). The co-occurrence frequencies can also be encoded in co-occurrence matrices and then matrix factorizations are used to solve the clustering problem (Cho et al. 2004; Ding et al. 2006). The document and word co-occurrence can also be formulated as a two-sided generative model using a Bayesian interpretation (Shan and Banerjee 2008; Wang, Domeniconi, and Laskey 2009). Moreover, Dhillon et al. (Dhillon, Mallela, and Modha 2003) modeled the co-clustering algorithm as an information-theoretic partition of the empirical joint probability distribution of two sets of discrete random variables. Later, Banerjee et. al. (Banerjee et al. 2007) extended this method to a general Bregman co-clustering and matrix factorization framework.

Recently, there are some initial efforts on extending the existing co-clustering methods to constrained co-clustering (Pensa and Boulicaut 2008; Wang, Li, and Zhang 2008; Chen, Wang, and Dong 2009). Most of these methods are based on matrix factorizations that optimize a sum squared residues-based objective function. Since it has been reported that, among the existing co-clustering methods, the ITCC algorithm that uses I-divergence is empirically more effective in analyzing sparse and high-dimensional text data than those methods that use Euclidean distance (Banerjee et al. 2007), we focused our work on extending the ITCC framework to incorporate constraints.

The CITCC Algorithm

In this section, we first describe how we formulate the constrained co-clustering 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.

Problem Formulation

Denote the document set and word set as $\mathcal{D} = \{d_1, d_2, \ldots, d_M\}$ and $\mathcal{V} = \{v_1, v_2, \ldots, v_V\}$. 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, shown by Dhillon, Mallela, and Modha (2003), a function

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

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:

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

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 $(p(v_1|\hat{d}_{k_d}), \ldots, p(v_V|\hat{d}_{k_d}))^T$, $p(v_i|\hat{d}_{k_d})) = p(v_i|\hat{v}_{k_v})p(\hat{v}_{k_v}|\hat{d}_{k_d})$ and $p(v_i|\hat{v}_{k_v}) = p(v_i)/p(l_{v_i} = \hat{v}_{k_v})$ due to hard clustering labels. Symmetrically we can define the probability for words: $p(\mathcal{D}|\hat{v}_{k_v})$ denotes a multinomial distribution based on the probabilities $(p(d_1|\hat{v}_{k_v}), \ldots, p(d_V|\hat{v}_{k_v}))^T$, $p(d_i|\hat{v}_{k_v})) = p(d_i|\hat{d}_{k_d})p(\hat{d}_{k_d}|\hat{v}_{k_v})$ and $p(d_i|\hat{d}_{k_d}) = p(d_i)/p(l_{d_i} = \hat{d}_{k_d})$.

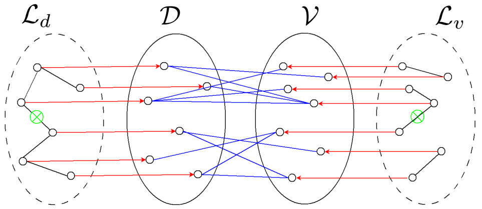


Figure 1: Illustration of the HMRF^2-ITCC model.

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

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

where $b_{\phi_{KL}}(\cdot)$ is a normalization constant determined by its divergency type (Banerjee et al. 2007).

For 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} ($m = 1, \ldots, M$) 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 (Basu, Bilenko, and Mooney 2004), we have

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

For must-links, the energy function is

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

and for cannot-links, the energy function is

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

where $p(\mathcal{V}|d_{m_1})$ denotes a multinomial distribution based on the probabilities $(p(v_1|d_{m_1}), \ldots, p(v_V|d_{m_1}))^T$, 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 $I_{true} = 1$ and $I_{false} = 0$.

Then, the constrained co-clustering problem can be formulated as a MAP estimation for label configurations:

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

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

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

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}.

Alternating EM

Globally optimizing the latent labels as well as the approximating function $q(d_m, v_i)$ is intractable. By substituting Eq. (2) into objective function (8), we can alternate the optimization process. First, the algorithm fixes $\mathcal{L}_v$ and minimizes the objective in (8) w.r.t $\mathcal{L}_d$. Then it fixes $\mathcal{L}_d$ and minimizes 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:

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

Optimizing this objective function is still computationally intractable. Here, we use a general EM algorithm to find an estimation (Basu, Bilenko, and Mooney 2004). 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 (Basu, Bilenko, and Mooney 2004) 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. Here, we find the label l_{d_m} by

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

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 update them as

$$q(\hat{d}_{k_d}, \hat{v}_{k_v}) = \sum_{l_{d_m} = k_d} \sum_{l_{v_i} = k_v} p(d_m, v_i) \tag{10}$$

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 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 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**
```

$$q(d_m|\hat{d}_{k_d}) = \frac{q(d_m)}{q(l_{d_m} = k_d)} \quad [q(d_m|\hat{d}_{k_d}) = 0 \text{ if } l_{d_m} \neq k_d] \tag{11}$$

$$q(v_i|\hat{v}_{k_v}) = \frac{q(v_i)}{q(l_{v_i} = k_v)} \quad [q(v_i|\hat{v}_{k_v}) = 0 \text{ if } l_{v_i} \neq k_v] \tag{12}$$

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 (Dhillon, Mallela, and Modha 2003). Algorithm 1 summarizes the main steps in the procedure. The convergence property is described in the Lemma and the time complexity is described in the Remark.

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

This lemma is easy to prove since the ICM algorithm decreases the non-negative 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 (Dhillon, Mallela, and Modha 2003).

Remark: The time complexity of the alternating EM algorithm for 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 nonzero number of document-word co-occurrences, 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.

Experiments

To evaluate the effectiveness of the proposed CITCC approach, we ran our experiments using the 20-newsgroups data set.[1] It is a collection of approximately 20,000 newsgroup documents, partitioned evenly across 20 different newsgroups. It is a benchmark data set with a ground truth category label for each document. We conducted our experiments in two different settings: (1) as a semi-supervised co-clustering algorithm that can incorporate human-annotated categories to improve document clustering performance, and (2) as an unsupervised document clustering algorithm that can incorporate constraints constructed from automatically extracted named entities. To evaluate the performance of CITCC against various clustering algorithms, we employed a widely-used normalized mutual information (NMI)-based measure (Strehl and Ghosh 2002). The NMI score is 1 if the clustering results perfectly match the category labeling and 0 if the clusters were obtained from random partition. In general, the larger the scores are, the better the clustering results are.

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 1985 documents after removing documents with less than five words. The vocabulary size was 11149 after removing words that appear in less than two documents. Each document was represented as a TF (term frequency) vector. In this experiment, we compared the performance of CITCC with that of several representative clustering algorithms such as Kmeans, constrained Kmeans (CKmeans) (Basu, Bilenko, and Mooney 2004), Semi-NMF (SNMF) (Ding, Li, and Jordan 2010), constrained SNMF (CSNMF) (Wang, Li, and Zhang 2008), Tri-factorization of Semi-NMF (STriNMF) (Wang, Li, and Zhang 2008), constrained STriNMF (CSTriNMF) (Wang, Li, and Zhang 2008) and ITCC (Dhillon, Mallela, and Modha 2003). Among all the methods, CKmeans, CSNMF, CSTriNMF and CITCC were constrained clustering algorithms; STriNMF, CSTriNMF, ITCC and CITCC were co-clustering methods; and CSTriNMF and CITCC were constrained co-clustering methods. In the experiment, the document cluster number was set to 2, the ground-truth number. For all the co-clustering algorithms tested, the word cluster number was empirically set to 4.

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 came 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 1000 words in each group to generate the word must-links. We did not use any word cannot-links in our experiments. We also varied the number of document and word constraints in each experiment by randomly selecting a fixed number of constraints from all the possible must-links and cannot-links to investigate their impact on clustering performance. The tradeoff parameters a_{m_1,m_2} and $\bar{a}_{m_1,m_2}$ for document constraints in Eqs. (5) and (6) 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.

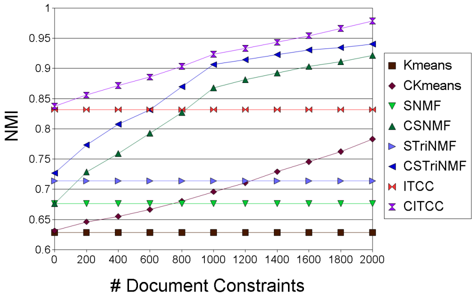


Figure 2: Semi-supervised document clustering.

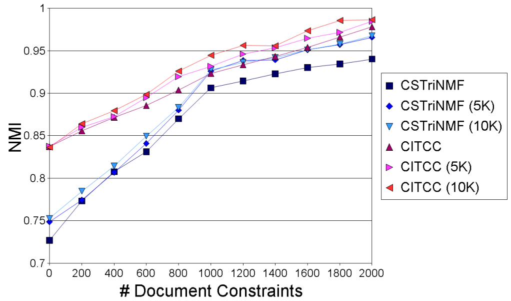


Figure 3: Effects of word constraints.

Figs. 2 and 3 show 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. 2, among all the methods we tested, CITCC consistently performed the best. It outperformed the non-constrained co-clustering algorithm ITCC significantly. It was also better than all the one-dimensional clustering algorithms, regardless of the number of document constraints used. Moreover, it was more effective than a known constrained co-clustering algorithm CSTriNMF. The number of document constraints seems to have big 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 co-clustering performance, we tested 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. 3,

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

Table 1: Statistics of NEs constraints under a two-class setting (alt.atheism vs. comp.graphics). The average similarity and standard deviation of all the documents in the two newsgroups is 0.033(2.212).

#NEs	≥1	≥2	≥3	≥4	≥5	≥6	≥11
#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)	58.19(83.35)
#Must-links	35156	5938	2271	1219	363	206	32
Correct Percentage	89.5%	95.6%	97.4%	98.4%	96.4%	97.1%	100%
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)	0.722(0.029)

Table 2: Comparison of different algorithms under a two-class setting (alt.atheism vs. comp.graphics).

#NEs	≥1	≥2	≥3	≥4	≥5	≥6	≥11	no constraint
CKmeans	0.631(0.018)	**0.666(0.018)**	0.655(0.023)	0.636(0.022)	0.629(0.022)	0.629(0.022)	0.628(0.022)	0.623(0.031)
CSNMF	0.476(0.014)	0.646(0.003)	0.648(0.009)	0.659(0.004)	**0.722(0.004)**	0.696(0.005)	0.664(0.002)	0.669(0.009)
CSTriNMF	0.586(0.061)	0.674(0.004)	0.687(0.004)	0.685(0.004)	0.705(0.020)	**0.759(0.027)**	0.595(0.018)	0.712(0.006)
CITCC	0.582(0.021)	0.843(0.026)	**0.844(0.027)**	0.842(0.026)	0.842(0.026)	0.843(0.027)	0.844(0.028)	0.842(0.029)

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.

Unsupervised Document Clustering with Additional Constraints

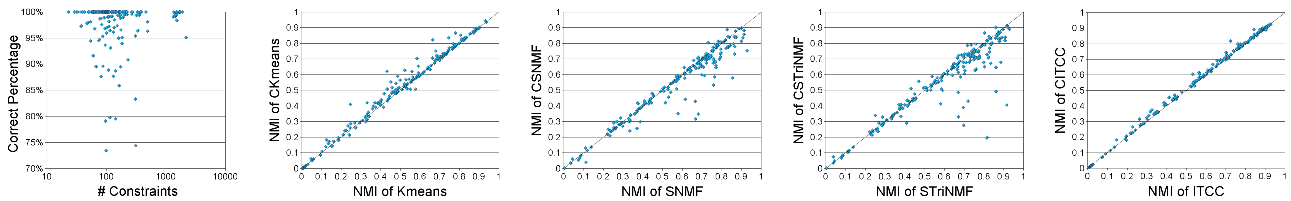
In practice, it is often difficult and costly to obtain sufficient human-annotated training examples for semi-supervised clustering. Consequently, we investigate whether constraints automatically derived from text can improve clustering performance. In this experiment, the constraints were constructed from named entities (NE) such as *person*, *location* and *organization*, derived by an NE recognizer[2]. If there are some overlapping NEs in two documents and the number of overlapping NEs is larger than a threshold, then we add a must-link to these documents.

Before we present the evaluation results, we first examine the quality of the NE constraints. Table 1 shows 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 must-links added based on overlapping NEs; and “*Correct Percentage*” indicates that of all the must-links 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 link. Moreover, after the minimum number of required overlapping NEs reached 2, the quality of the must-links derived was quite high (95.6%). 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 vs. comp.graphics). Specifically, we compared the performance of the constrained version of each algorithm with that of the non-constrained version. All the numbers shown in Table 2 are the means and the standard deviations of the NMI scores across five random runs. The column *no constraint* shows the performance of the non-constrained 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 (0.844). Moreover, for each method, the constrained version was able to take advantage of the additional NE constraints to improve its clustering performance over its non-constrained 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 non-constrained 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 based on human-provided category labels. To explain this, we computed the average similarity of documents with must-links. As shown in Table 1, the average document similarity increased as more overlapping NEs were required. This implies that NE constraints may provide redundant information as provided by the document similarity metric. In contrast, human-provided category labels may provide additional information (e.g., topic information) to guide the clustering towards the ground truth.

We also tested the algorithms under all the 190 two-class clustering conditions for all the 20 newsgroups. For all the two-class problems, we sampled 50% of the documents for testing. The number of overlapping NEs was set to be at least 5. The resulting number of must-link constraints (Cons.) and the correct percentage (Cor. Perc.) values are shown in Fig. 4 (a). These data show that, overall, the derived NE constraints were quite accurate. In most cases, over 95% of the derived NE constraints were correct. In addition, as shown in Figs. 4 (b) and (e), for all the 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 cor-

[2] http://nlp.stanford.edu/software/CRF-NER.shtml



(a) # Cons. and Cor. Perc. (b) Kmeans vs. CKmeans (c) SNMF vs. CSNMF (d) STriNMF vs. CSTriNMF (e) ITCC vs. CITCC

Figure 4: NE constraints results on 190 two-class clustering problems in 20-newsgroups data.

ner, thus indicating consistently high performance for both ITCC and CITCC. For the results in Figs. 4 (c) and (d), 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, since we used the same tradeoff parameters in all the experiments, the parameters may not be optimal for CSNMF and CSTriNMF.

Conclusion and Future Work

In this paper, we proposed a novel constrained co-clustering approach that automatically incorporates constraints into information-theoretic co-clustering. Our evaluations on a benchmark data set demonstrated the effectiveness of the proposed method for clustering textual documents. Our algorithm consistently outperformed all the tested constrained clustering and co-clustering methods under different conditions. There are several directions for the future research. For example, we will explore other text features that can be automatically derived by natural language processing (NLP) tools to further improve unsupervised document clustering performance. We are also interested in applying CITCC to other text analysis applications such as visual text summarization.

Acknowledgements

We would like to thank Fei Wang for insightful discussions and help on implementing the constrained NMF algorithms. We would like to thank Keith Houck for his kind help on proofreading our paper.

References

Banerjee, A.; Dhillon, I.; Ghosh, J.; Merugu, S.; and Modha, D. S. 2007. A generalized maximum entropy approach to bregman co-clustering and matrix approximation. *Journal of Machine Learning Research* 8:1919–1986.

Basu, S.; Banerjee, A.; and Mooney, R. J. 2002. Semi-supervised clustering by seeding. In *ICML*, 27–34.

Basu, S.; Bilenko, M.; and Mooney, R. J. 2004. A probabilistic framework for semi-supervised clustering. In *SIGKDD*, 59–68.

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

Bilenko, M.; Basu, S.; and Mooney, R. J. 2004. Integrating constraints and metric learning in semi-supervised clustering. In *ICML*, 81–88.

Chen, Y.; Wang, L.; and Dong, M. 2009. Non-negative matrix factorization for semi-supervised heterogeneous data co-clustering. *IEEE Trans. on Knowledge and Data Engineering*.

Cheng, Y., and Church, G. M. 2000. Biclustering of expression data. In *ISMB*, 93–103.

Cho, H.; Dhillon, I. S.; Guan, Y.; and Sra, S. 2004. Minimum sum-squared residue co-clustering of gene expression data. In *SDM*.

Dhillon, I. S.; Mallela, S.; and Modha, D. S. 2003. Information-theoretic co-clustering. In *KDD*, 89–98.

Dhillon, I. S. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. In *KDD*, 269–274.

Ding, C.; Li, T.; Peng, W.; and Park, H. 2006. Orthogonal nonnegative matrix t-factorizations for clustering. In *KDD*, 126–135.

Ding, C. H. Q.; Li, T.; and Jordan, M. I. 2010. Convex and semi-nonnegative matrix factorizations. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 32(1):45–55.

Jain, A.; Murty, M.; and Flynn, P. 1999. Data clustering: A review. *ACM Computing Surveys* 31(3):264–323.

Lu, Z., and Leen, T. K. 2007. Penalized probabilistic clustering. *Neural Computation* 19(6):1528–1567.

Nigam, K.; McCallum, A. K.; Thrun, S.; and Mitchell, T. M. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning* 39(2/3):103–134.

Pensa, R. G., and Boulicaut, J.-F. 2008. Constrained co-clustering of gene expression data. In *SDM*, 25–36.

Shan, H., and Banerjee, A. 2008. Bayesian co-clustering. In *ICDM*, 530–539.

Strehl, A., and Ghosh, J. 2002. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *Journal on Machine Learning Research* 3:583–617.

Wagstaff, K.; Cardie, C.; Rogers, S.; and Schrödl, S. 2001. Constrained k-means clustering with background knowledge. In *ICML*, 577–584.

Wang, P.; Domeniconi, C.; and Laskey, K. B. 2009. Latent dirichlet bayesian co-clustering. In *ECML/PKDD*, 522–537.

Wang, F.; Li, T.; and Zhang, C. 2008. Semi-supervised clustering via matrix factorization. In *SDM*, 1–12.

Xing, E. P.; Ng, A. Y.; Jordan, M. I.; and Russell, S. J. 2002. Distance metric learning with application to clustering with side-information. In *NIPS*, 505–512.