# Knowledge Transfer on Hybrid Graph

**Zheng Wang, Yangqiu Song and Changshui Zhang**
State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Automation, Tsinghua University, Beijing 100084, P. R. China
{wangzheng04,yqsong}@gmail.com, zcs@mail.tsinghua.edu.cn

## Abstract

In machine learning problems, labeled data are often in short supply. One of the feasible solution for this problem is transfer learning. It can make use of the labeled data from other domain to discriminate those unlabeled data in the target domain. In this paper, we propose a transfer learning framework based on similarity matrix approximation to tackle such problems. Two practical algorithms are proposed, which are the label propagation and the similarity propagation. In these methods, we build a hybrid graph based on all available data. Then the information is transferred cross domains through alternatively constructing the similarity matrix for different part of the graph. Among all related methods, similarity propagation approach can make maximum use of all available similarity information across domains. This leads to more efficient transfer and better learning result. The experiment on real world text mining applications demonstrates the promise and effectiveness of our algorithms.

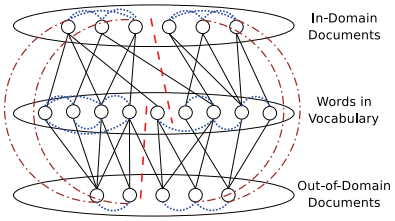
## 1 Introduction

Transfer learning is a powerful ability of human to apply knowledge and skills learned in previous tasks to novel tasks [Ormrod, 2004]. In modern machine learning and data mining fields, it has been widely investigated to try to simulate this human learning mechanism to achieve artificial intelligence [Pan and Yang, 2007]. Recently, the explosive growth in data warehouse and internet usage has made large amount of unsorted information potentially available for data mining problems. Labeling them is a very expensive and time consuming work. Transfer learning gives us a possible solution to optimally discriminate this type of data. It can borrow some supervised information from other similar tasks to enhance the discrimination of the learning machine for those unlabeled data. For instance, the difference between documents that describe *stock* and *book* may help to distinguish the documents that describe *investment* and *research*. Though they are different topics with different word distributions, *stock* and *investment* both talk about economic stuff, while *book* and *research* are topics that relate to education and academia. This is a typical transfer learning problem that discriminates the totally unlabeled data that are under a different distribution from the labeled data.

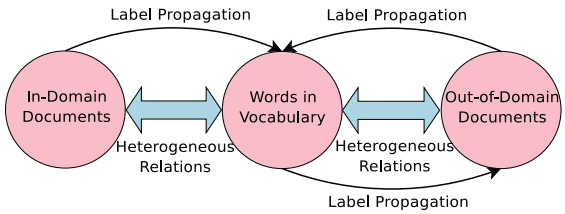
In this paper, we focus on above transfer learning approach cross different domains mainly based on text mining problems. In this problem, the labeled data are from a domain $\mathcal{D}^{(i)}$ and the unlabeled data are from another domain $\mathcal{D}^{(o)}$. $\mathcal{D}^{(i)}$ is called *in-domain* and $\mathcal{D}^{(o)}$ *out-of-domain*. In addition, it is assumed that in-domain $\mathcal{D}^{(i)}$ and out-of-domain $\mathcal{D}^{(o)}$ are related to make the domain-transfer learning feasible. The objective is to discriminate the data from out-of-domain $\mathcal{D}^{(o)}$ as accurately as possible with the help of the data from in-domain $\mathcal{D}^{(i)}$. Several previous works have been done for this problem, which are Co-Clustering based Classification (CoCC) [Dai *et al.*, 2007] and Cross-Domain Spectral Classification (CDSC) [Ling *et al.*, 2008]. CoCC is based on information-theoretic co-clustering [Dhillon *et al.*, 2003]. It regularizes the out-of-domain discrimination problem with the in-domain word partition. The labels of in-domain documents first propagate to the words and then the clustering of the words propagate to the out-of-domain documents. The procedure iterates until convergence. CDSC is based on spectral learning [Kamvar *et al.*, 2003]. It seeks an optimal partition of the data, which preserves the supervised segmentation information for the in-domain data and splits the out-of-domain data as separately as possible in terms of the cut size. The experiments in their papers have shown great improvements to conventional supervised and semi-supervised methods. However, there remain problems.

- The information-theoretic co-clustering heavily relies on the initialization. The initialization of document clusters are relatively easy. However, the word clusters may not be well-established. Even human experts could not well cluster such large amount of words themselves.
- Both methods have many heuristic parameters which the result relies on, especially the trade-off parameter for in-domain and out-of-domain information. To choose satisfied parameters is difficult for practical applications.
- Each of them only uses a part of the cross domain similarity. CoCC focuses on the conjunct word clustering, and CDSC focuses on the document similarity. In these situations, neither of them makes full use of the relationship among the relational data.

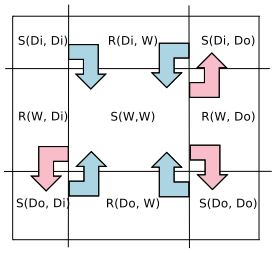
Focusing on above problems, we proposed a novel transfer




(a) Hybrid graph.




(b) Label propagation, similar with [Dai *et al.* 2007].




(c) Similarity propagation.

Figure 1: Illustration of hybrid graph.

learning framework on a hybrid graph. We set both documents and words as vertexes in one graph. The transfer learning procedure is conducted across different parts of the graph. First, we present an algorithm *label propagation*. It employs spectral analysis of the modified similarity between homogeneous relations cross domains to propagated the discriminative information. Moreover, a further algorithm, *similarity propagation*, is proposed. It can effectively alleviate above problems. In this algorithm, we pursue the transferred information among both document similarity and word clustering. It uses all available routes to transfer and get better and more stable results.

The rest of the paper is organized as follows. In section 2, we formulate the problem and present our algorithms. Section 3 reviews the related works and analyzes the involved problem. Section 4 shows our experiments. Finally we conclude the paper and discuss some future work in section 5.

## 2 Knowledge Transfer on Hybrid Graph

In this section, we present our transfer learning method on hybrid graph, and show the two algorithms label propagation and similarity propagation.

### 2.1 Problem Formulation

Let $\mathcal{D}^{(i)} = \{d_1^{(i)}, d_2^{(i)}, ..., d_{N_i}^{(i)}\}$ be the in-domain data set, with $N_i$ labeled documents, $\mathcal{D}^{(o)} = \{d_1^{(o)}, d_2^{(o)}, ..., d_{N_o}^{(o)}\}$ be the out-of-domain data set, with $N_o$ unlabeled documents. Each document is represented as a vector of word frequencies. Both document sets share the same word vocabulary $\mathcal{W} = \{w_1, w_2, ..., w_M\}$, where $M$ is the vocabulary size.

Given the document set $\mathcal{D} = \{\mathcal{D}^{(i)}, \mathcal{D}^{(o)}\}$ and the word set $\mathcal{W}$, we build a hybrid graph with each document and word as its vertex, and the edge between each pair of vertexes denotes their similarity. It is similar with the bipartite graph in spectral co-clustering [Dhillon, 2001; Zha *et al.*, 2001]. However in bipartite graph, it assumes the relations among documents are zeros.

In this paper, we construct the hybrid graph that involve all kinds of relationships, which is shown in Fig. 1 (a). The co-occurrence matrix between documents and words is still used to represent their similarities, which is $\mathbf{R}^{(di,w)} \in \mathbb{R}^{N_i \times M}$ for in-domain data and $\mathbf{R}^{(do,w)} \in \mathbb{R}^{N_o \times M}$ for out-of-domain data. The similarities between documents are denoted as $\mathbf{S}^{(di,di)} \in \mathbb{R}^{N_i \times N_i}$, $\mathbf{S}^{(di,do)} \in \mathbb{R}^{Ni \times N_o}$, $\mathbf{S}^{(do,di)} = \mathbf{S}^{(di,do)^T} \in \mathbb{R}^{N_o \times N_i}$, $\mathbf{S}^{(do,do)} \in \mathbb{R}^{N_o \times N_o}$ and $\mathbf{S}^{(w,w)} \in \mathbb{R}^{M \times M}$. The whole similarity matrix can be formulated as:

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}^{(di,di)} & \mathbf{R}^{(di,w)} & \mathbf{S}^{(di,do)} \\ \mathbf{R}^{(di,w)^T} & \mathbf{S}^{(w,w)} & \mathbf{R}^{(do,w)^T} \\ \mathbf{S}^{(di,do)^T} & \mathbf{R}^{(do,w)} & \mathbf{S}^{(do,do)} \end{bmatrix} \tag{1}$$

which is shown in Fig. 1 (c). We divide the similarity relationship into two types, defined as follows.

**Definition 1. (Homogeneous Relations)** *The relations between documents and documents are denoted as homogeneous relations, described by* $\mathbf{S}^{(di,di)}$, $\mathbf{S}^{(di,do)}$, $\mathbf{S}^{(w,w)}$ *and* $\mathbf{S}^{(do,do)}$.

**Definition 2. (Heterogeneous Relations)** *The relations between documents and words are denoted as heterogeneous relations, described by* $\mathbf{R}^{(di,w)}$ *and* $\mathbf{R}^{(do,w)}$.

Without loss of generality, we assume the class or clustering numbers of in-domain documents, out-of-domain documents and the words are $K_{di}$, $K_{do}$ and $K_w$ respectively. Then we define their indicator matrices as:

$$\mathbf{C}_{pq}^{(x)} = \begin{cases} \frac{1}{\sqrt{\pi_q^{(x)}}} & \text{if item } p \in \pi_q^{(x)} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $x$ can be $di$, $do$ or $w$, $\pi_q^{(x)}$ is the number of objects in the $q$th cluster in the corresponding set. And we have $(\mathbf{C}^{(x)})^T\mathbf{C}^{(x)} = \mathbf{I}_{K_x}$, where $\mathbf{I}_{K_x} \in \mathbb{R}^{K_x \times K_x}$ is the identity matrix.

Using the indicator matrix, we calculate the corresponding similarity matrix within the homogeneous data sets as:

$$\mathbf{S}^{(x,x)} = \mathbf{C}^{(x)} diag(\pi^{(x)})(\mathbf{C}^{(x)})^T. \tag{3}$$

$diag(\pi^{(di)}) \in \mathbb{R}^{K_x \times K_x}$ is the diagonal matrix, where the elements are $\pi_q^{(di)}$, $q = 1, ..., K_x$. The in-domain documents have the groundtruth labels. Thus we denote the class indicator matrix as $\bar{\mathbf{C}}^{(di)}$, and the corresponding similarity matrix as $\bar{\mathbf{S}}^{(di,di)}$.

Moreover, to balance the relations and similarities, we normalize them as

$$\begin{aligned} \bar{\mathbf{S}}^{(di,di)} &\leftarrow (\bar{\mathbf{D}}^{(di,di)})^{-\frac{1}{2}} \bar{\mathbf{S}}^{(di,di)} (\bar{\mathbf{D}}^{(di,di)})^{-\frac{1}{2}} \\ \mathbf{S}^{(do,do)} &\leftarrow (\mathbf{D}^{(do,do)})^{-\frac{1}{2}} \mathbf{S}^{(do,do)} (\mathbf{D}^{(do,do)})^{-\frac{1}{2}} \\ \mathbf{S}^{(w,w)} &\leftarrow (\mathbf{D}^{(w,w)})^{-\frac{1}{2}} \mathbf{S}^{(w,w)} (\mathbf{D}^{(w,w)})^{-\frac{1}{2}} \\ \mathbf{R}^{(di,w)} &\leftarrow (\mathbf{D}^{(di,w)})^{-\frac{1}{2}} \mathbf{R}^{(di,w)} (\mathbf{D}^{(w,di)})^{-\frac{1}{2}} \\ \mathbf{R}^{(do,w)} &\leftarrow (\mathbf{D}^{(do,w)})^{-\frac{1}{2}} \mathbf{R}^{(do,w)} (\mathbf{D}^{(w,do)})^{-\frac{1}{2}} \end{aligned} \tag{4}$$

where $\bar{\mathbf{D}}_{pp}^{(di,di)} = \sum_q \bar{\mathbf{S}}_{pq}^{(di,di)}$, $\mathbf{D}_{pp}^{(do,do)} = \sum_q \mathbf{S}_{pq}^{(do,do)}$, $\mathbf{D}_{pp}^{(w,w)} = \sum_q \mathbf{S}_{pq}^{(w,w)}$, $\mathbf{D}_{pp}^{(di,w)} = \sum_q \mathbf{R}_{pq}^{(di,w)}$, $\mathbf{D}_{qq}^{(w,di)} = \sum_p \mathbf{R}_{pq}^{(di,w)}$, $\mathbf{D}_{pp}^{(do,w)} = \sum_q \mathbf{R}_{pq}^{(do,w)}$ and $\mathbf{D}_{qq}^{(w,do)} = \sum_p \mathbf{R}_{pq}^{(do,w)}$. This is inspired by the normalized cut [Shi and Malik, 2000] and spectral co-clustering [Dhillon, 2001]. After the normalization, the similarity matrix of in-domain documents can be re-written as $\bar{\mathbf{S}}^{(di,di)} = \bar{\mathbf{C}}^{(di)}(\bar{\mathbf{C}}^{(di)})^T$ and we can easily verify that the Moore-Penrose pseudoinverse is $(\bar{\mathbf{S}}^{(di,di)})^\dagger = \bar{\mathbf{S}}^{(di,di)}$.

Based on the description above, our problem is to find the unknown indicator matrix $\mathbf{C}^{(do)}$ for the out-of-domain data. In the next two subsections, we will present two novel algorithms to deal with this problem.

## 2.2 Label Propagation

The main idea of label propagation is to transfer the supervised information from in-domain to out-of-domain through the similar word structure they share. We adopt the spectral relational clustering method [Long *et al.*, 2006] for both in-domain and out-of-domain data simultaneously. The objective function is:

$$\begin{aligned} & J(\mathbf{C}^{(do)}, \mathbf{C}^{(w)}, \mathbf{H}^{(di,w)}, \mathbf{H}^{(do,w)}) \\ &= \sum ||\mathbf{R}^{(di,w)} - \bar{\mathbf{C}}^{(di)}\mathbf{H}^{(di,w)}(\mathbf{C}^{(w)})^T||_F^2 \\ &+ \sum ||\mathbf{R}^{(do,w)} - \mathbf{C}^{(do)}\mathbf{H}^{(do,w)}(\mathbf{C}^{(w)})^T||_F^2 \end{aligned} \quad (5)$$

where $\mathbf{H}^{(di,w)} \in \mathbb{R}^{K_{di} \times K_w}$ and $\mathbf{H}^{(do,w)} \in \mathbb{R}^{K_{do} \times K_w}$ denote the co-occurrence relationship between clusters of documents and words. This objective aims to find the best partition of both out-of-domain documents and word vocabulary, under the restriction from the in-domain relational structure.

Taking in indicator matrixes constraints, the final task is

$$\begin{aligned} & \min J. \\ \text{s.t.} \ & (\mathbf{C}^{(w)})^T\mathbf{C}^{(w)} = \mathbf{I}_{K_i} \\ & (\mathbf{C}^{(do)})^T\mathbf{C}^{(do)} = \mathbf{I}_{K_{do}} \end{aligned} \quad (6)$$

Based on the analysis in [Long *et al.*, 2006], we can deduce that the optimal association matrixes are $\mathbf{H}^{(di,w)} = (\bar{\mathbf{C}}^{(di)})^T\mathbf{R}^{(di,w)}\mathbf{C}^{(w)}$ and $\mathbf{H}^{(do,w)} = (\mathbf{C}^{(do)})^T\mathbf{R}^{(do,w)}\mathbf{C}^{(w)}$ for our objective. Take these results into the objective function Eq. (5), we have:

$$\begin{aligned} J &= tr((\mathbf{R}^{(di,w)})^T\mathbf{R}^{(di,w)}) \\ &- tr((\mathbf{C}^{(w)})^T(\mathbf{R}^{(di,w)})^T\bar{\mathbf{C}}^{(di)}(\bar{\mathbf{C}}^{(di)})^T\mathbf{R}^{(di,w)}\mathbf{C}^{(w)}) \\ &+ tr((\mathbf{R}^{(do,w)})^T\mathbf{R}^{(do,w)}) \\ &- tr((\mathbf{C}^{(w)})^T(\mathbf{R}^{(do,w)})^T\mathbf{C}^{(do)}(\mathbf{C}^{(do)})^T\mathbf{R}^{(do,w)}\mathbf{C}^{(w)}). \end{aligned} \quad (7)$$

where $tr(\cdot)$ is the trace of a matrix. Note that

$$tr((\mathbf{R}^{(do,w)})^T\mathbf{R}^{(do,w)}) = tr((\mathbf{R}^{(do,w)})(\mathbf{R}^{(do,w)})^T)$$

and

$$\begin{aligned} & tr((\mathbf{C}^{(w)})^T(\mathbf{R}^{(do,w)})^T\mathbf{C}^{(do)}(\mathbf{C}^{(do)})^T\mathbf{R}^{(do,w)}\mathbf{C}^{(w)}) \\ = \ & tr((\mathbf{C}^{(do)})^T(\mathbf{R}^{(do,w)})\mathbf{C}^{(w)}(\mathbf{C}^{(w)})^T(\mathbf{R}^{(do,w)})^T\mathbf{C}^{(do)}). \end{aligned}$$

We can see that the objective is convex w.r.t. both $\mathbf{C}^{(do)}$ and $\mathbf{C}^{(w)}$. As a result, we use *alternating optimization* technique to find the optimal solution. The following *Ky-Fan* theorem guarantees a closed-form solution for each alternative step.

Table 1: Transfer Learning by Label Propagation

**Input:** The labeled in-domain data set $\mathcal{D}^{(i)}$, unlabeled out-of-domain data set $\mathcal{D}^{(o)}$ and the word feature set $\mathcal{W}$.

**Initialize:** The heterogeneous relations $\mathbf{R}^{(di,w)}$ and $\mathbf{R}^{(do,w)}$. The indicator matrices $\bar{\mathbf{C}}^{(di)}$ and $\mathbf{C}^{(do)}$.

**Repeat:**

1: Update $\mathbf{C}^{(w)}$ by the leading $K_w$ eigenvectors of

$$\begin{aligned} & (\mathbf{R}^{(di,w)})^T\bar{\mathbf{C}}^{(di)}(\bar{\mathbf{C}}^{(di)})^T\mathbf{R}^{(di,w)} \\ + \ & (\mathbf{R}^{(do,w)})^T\mathbf{C}^{(do)}(\mathbf{C}^{(do)})^T\mathbf{R}^{(do,w)} \end{aligned} \quad (8)$$

2: Discretize $\mathbf{C}^{(w)}$ as in [Yu and Shi, 2003].

3: Update $\mathbf{C}^{(bo)}$ by the leading $K_o$ eigenvectors of

$$(\mathbf{R}^{(do,w)})\mathbf{C}^{(w)}(\mathbf{C}^{(w)})^T(\mathbf{R}^{(do,w)})^T \quad (9)$$

4: Discretize $\mathbf{C}^{(bo)}$ as in [Yu and Shi, 2003].

**Until:** Convergence.

**Output:** The label indicator matrices $\mathbf{C}^{(do)}$ and $\mathbf{C}^{(w)}$.

**Theorem 3. (Ky-Fan Theorem)** *[Bhatia, 1997] Let* $\mathbf{S}$ *be a symmetric matrix with eigenvalues* $\lambda_1 \geqslant \lambda_2 \geqslant ... \geqslant \lambda_K$ *and corresponding eigenvectors* $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_K]$. *Then* $\sum_{i=1}^K \lambda_i = \max_{V^TV=I_K} \mathbf{V^TSV}$. *Moreover, the optimal* $\mathbf{V}$ *is given by* $\mathbf{UT}$ *where* $\mathbf{T}$ *is an arbitrary orthogonal matrix.*■

Based on this theorem, we can calculate $\mathbf{C}^{(do)}$ and $\mathbf{C}^{(w)}$ alternatively until converged. This procedure is shown in Fig. 1 (b). The"labels" of in-domain and out-of-domain documents first propagate to the words, then the "labels" of words propagate back to the out-of-domain documents. The algorithm is summarized in Table 1. Since the objective function (5) is convex for each variable, it has local optimum.

## 2.3 Similarity Propagation

In the label propagation algorithm, it only uses the heterogeneous relations to conduct the transfer procedure, which is similar with [Dai *et al.*, 2007]. Although we can add a tradeoff parameter to control the strength of the information transfer, like [Dai *et al.*, 2007], the homogeneous relations are still ignored. In this situation, the optimal transfer can hardly be achieved. In this subsection, we propose a more efficient algorithm, similarity propagation, to make full use of the available information.

In the similarity propagation algorithm, we do not merely stare at the discrimination indication for the documents and words, but aim at revealing all types of similarities in the hybrid graph as shown in Fig. 1 (c). In this setting, the heterogeneous relations of the data and the in-domain homogeneous relations (given by the document labels) are provided. The task is to recover all other homogeneous relations as better as possible. During this reconstruction process, the Nyström method is needed for matrix approximation.

**Theorem 4. (Nyström approximation)** *[Williams and*

*Seeger, 2001] Let positive semi-definite matrix* $\mathbf{S}$ *be blocked as*

$$\mathbf{S} = \begin{bmatrix} \mathbf{O} & \mathbf{P} \\ \mathbf{P}^T & \mathbf{Q} \end{bmatrix}$$

*If* $\mathbf{Q}$ *is missing, the the first* $K$ *eigenvalues of* $\mathbf{S}$ *and the corresponding eigenvectors can be approximated as*

$$\lambda_{\mathbf{S},i} = \frac{N_{\mathbf{S}}}{N_{\mathbf{P}}}\lambda_{\mathbf{P},i} \quad and \quad \mathbf{u}_{\mathbf{S},i} = \sqrt{\frac{N_{\mathbf{P}}}{N_{\mathbf{S}}}}\frac{1}{\lambda_{\mathbf{P},i}}\begin{bmatrix} \mathbf{O} \\ \mathbf{P}^T \end{bmatrix}\mathbf{u}_{\mathbf{S},i}$$

*where* $\lambda_{\mathbf{P},i}$ *and* $\mathbf{u}_{\mathbf{S},i}$ *are the leading non-zero eigenvalues and corresponding eigenvectors of* $\mathbf{P}$*. And* $\mathbf{S}$ *can be approximated as* $\sum \mathbf{u}_{\mathbf{S},i}\lambda_{\mathbf{S},i}{\mathbf{u}_{\mathbf{S},i}}^T$*. Moreover, the quality of the approximation can be quantified by the norm of the Schur complement*

$$||\mathbf{Q} - \mathbf{P}^T\mathbf{O}^\dagger\mathbf{P}||_F \tag{10}$$

*where* $||\cdot||_F$ *represents the Frobenius norm of a matrix,* $\mathbf{O}^\dagger$ *is the Moore-Penrose pseudoinverse.* ■

Based on theorems 3 and 4, we can make use of eigenvalue decomposition approach to approximate $\mathbf{S}$ defined in Eq. (1) alternatively. Reset the blocked $\mathbf{S}$ as

$$\mathbf{S} = \left[\begin{array}{cc|c} \mathbf{S}^{(di,di)} & \mathbf{S}^{(di,do)} & \mathbf{R}^{(di,w)} \\ {\mathbf{S}^{(di,do)}}^T & \mathbf{S}^{(do,do)} & \mathbf{R}^{(do,w)} \\ \hline {\mathbf{R}^{(di,w)}}^T & {\mathbf{R}^{(do,w)}}^T & \mathbf{S}^{(w,w)} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{O} & \mathbf{P} \\ \hline \mathbf{P}^T & \mathbf{Q} \end{array}\right]$$

We implement the approximation alternatively by two steps: first compute $\hat{\mathbf{Q}} = \mathbf{P}^T\mathbf{O}^\dagger\mathbf{P}$; second compute $\hat{\mathbf{O}} = \mathbf{P}\mathbf{Q}^\dagger\mathbf{P}^T$ and reset $\mathbf{S}^{(di,di)} = \bar{\mathbf{S}}^{(di,di)}$[1]. This procedure can be seen as a propagation of the similarities of full homogeneous relations through heterogeneous similarity, which is shown in Fig. 1 (c). The propagation stops when the recovered unknown similarity blocks are no longer changing. The flow chart is given in Table 2[2].

Moreover, in the following corollary, we show that label propagation method can also be formulated as an alternative Nyström approximation for the similarity matrix.

**Corollary 5.** *Label propagation is also corresponding to a two-step Nyström approximation procedure.*

*Proof:* In this situation, the similarity matrix $\mathbf{S}$ is reset as

$$\mathbf{S} = \left[\begin{array}{cc|c} \mathbf{S}^{(di,di)} & \mathbf{0} & \mathbf{R}^{(di,w)} \\ \mathbf{0} & \mathbf{S}^{(do,do)} & \mathbf{R}^{(do,w)} \\ \hline {\mathbf{R}^{(di,w)}}^T & {\mathbf{R}^{(do,w)}}^T & \mathbf{S}^{(w,w)} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{O} & \mathbf{P} \\ \hline \mathbf{P}^T & \mathbf{Q} \end{array}\right].$$

The Nyström approximations lead to $\mathbf{S}^{(w,w)} = {\mathbf{R}^{(di,w)}}^T(\bar{\mathbf{S}}^{(di,di)})^\dagger\mathbf{R}^{(di,w)} + {\mathbf{R}^{(do,w)}}^T(\mathbf{S}^{(do,do)})^\dagger\mathbf{R}^{(do,w)}$ and $\mathbf{S}^{(do,do)} = \mathbf{R}^{(do,w)}{\mathbf{S}^{(w,w)}}^\dagger(\mathbf{R}^{(do,w)})^T$. Using the similarity definition in Eq. (4) we have $(\bar{\mathbf{S}}^{(di,di)})^\dagger = \bar{\mathbf{S}}^{(di,di)}$, $(\mathbf{S}^{(do,do)})^\dagger = \mathbf{S}^{(do,do)} = \mathbf{C}^{(do)}(\mathbf{C}^{(do)})^T$ and $(\mathbf{S}^{(w,w)})^\dagger = \mathbf{S}^{(w,w)} = \mathbf{C}^{(w)}(\mathbf{C}^{(w)})^T$. Take these back into the above equations, the corollary is proven. ■

As it is shown in corollary 5, in label propagation algorithm, it assumes that $\mathbf{R}^{(di,do)} = \mathbf{0}$ and the similarities in $\mathbf{S}^{(do,do)}$ are only propagated from $\mathbf{S}^{(w,w)}$. Similarly analysis can be done to show that CoCC loses $\mathbf{R}^{(di,do)}$, and CDSC loses $\mathbf{R}^{(w,w)}$. This is the reason why these methods are insufficient. We will approve this empirically in section 4.

Table 2: Transfer Learning by Similarity Propagation

**Input:** The labeled in-domain data set $\mathcal{D}^{(i)}$, unlabeled out-of-domain data set $\mathcal{D}^{(o)}$ and the word feature set $\mathcal{W}$.

**Initialize:** The heterogeneous relations $\mathbf{R}^{(di,w)}$ and $\mathbf{R}^{(do,w)}$. The homogeneous similarities $\mathbf{S}^{(di,di)} = \bar{\mathbf{S}}^{(di,di)}$, $\mathbf{S}^{(do,do)}$ and $\mathbf{S}^{(di,do)}$. Set $\mathbf{O}$, $\mathbf{P}$ and $\mathbf{Q}$ as the forms in Eq. (11).

**Repeat:**

1: Update $\mathbf{Q} = \mathbf{P}^T\mathbf{O}^\dagger\mathbf{P}$ .

2: Update $\mathbf{O} = \mathbf{P}\mathbf{Q}^\dagger\mathbf{P}^T$ .

3: Reset $\mathbf{S}^{(di,di)} = \bar{\mathbf{S}}^{(di,di)}$.

**Until:** Convergence.

**Output:** The label indicator matrices $\mathbf{C}^{(do)}$ and $\mathbf{C}^{(w)}$ are computed by using normalized cut algorithm with homogeneous similarities $\mathbf{S}^{(do,do)}$ and $\mathbf{S}^{(w,w)}$.

## 3 Related Works

The most related works to our approach are transfer learning, semi-supervised learning and spectral clustering. We will look through these works and present our understanding for the transfer learning problem in this paper.

### 3.1 Learning with Labeled and Unlabeled Data

In our problem, there are both labeled and unlabeled data. The most popular technique used to learn with labeled and unlabeled data is semi-supervised learning [Zhu, 2005]. Many successful methods have been proposed under this topic, such as tranductive SVM [Joachims, 1999] and manifold regularization [Belkin *et al.*, 2006]. However, the conventional semi-supervised learning methods have a strict assumption that the labeled and unlabeled data should be sampled from the same distribution, which is always not the case in real applications [Pan and Yang, 2007].

Transfer learning is thus adopted to solve such a complex problem. Besides, it can be used in many other cases, such as learning with different training and test data distributions [Sugiyama *et al.*, 2007] and learning multiple tasks [Caruana, 1997]. The problem in this paper is just a particular setting for transfer learning, which transfers the labeled information to a totally unsupervised task. We can not arbitrarily judge whether the out-of-domain task is a classification or a clustering task. As the data have hiberarchical structure. In the small scale, these data have their own class attributes which are different from the in-domain label information. In the

[1] To guarantee the validity of nyström method, $\mathbf{S}^{(di,do)}$ will be re-initialized if the positive semi-definition of S is violated.

[2] The initialization of $\mathbf{S}^{(do,do)}$ can be done using any clustering method or semi-supervised method with the help of the labeled data. And $\mathbf{S}^{(di,do)}$ can be initially calculated as the inner product of $\bar{\mathbf{C}}^{(di)}$ and $\mathbf{C}^{(do)}$. Instead of these, we can initialize $\mathbf{S}^{(w,w)}$ and start from the second step.

Table 3: Transfer Learning Data Sets.

| Data Set | $\mathcal{D}^{(i)}$ | $\mathcal{D}^{(o)}$ |
|---|---|---|
| comp vs rec | comp.graphics<br>comp.os.ms-windows.misc<br>rec.autos<br>rec.motorcycles | comp.sys.ibm.pc.hardware<br>comp.sys.mac.hardware<br>rec.sport.baseball<br>rec.sport.hockey |
| comp vs sci | comp.graphics<br>comp.os.ms-windows.misc<br>sci.crypt<br>sci.med | comp.sys.ibm.pc.hardware<br>comp.sys.mac.hardware<br>sci.electronics<br>sci.space |
| comp vs talk | comp.graphics<br>comp.os.ms-windows.misc<br>talk.politics.guns<br>talk.politics.mideast | comp.sys.ibm.pc.hardware<br>comp.sys.mac.hardware<br>talk.politics.misc |
| rec vs sci | rec.autos<br>rec.motorcycles<br>sci.crypt<br>sci.med | rec.sport.baseball<br>rec.sport.hockey<br>sci.electronics<br>sci.space |
| rec vs talk | rec.autos<br>rec.motorcycles<br>talk.politics.guns<br>talk.politics.mideast | rec.sport.baseball<br>rec.sport.hockey<br>talk.politics.misc |
| sci vs talk | sci.crypt<br>sci.med<br>talk.politics.guns<br>talk.politics.mideast | sci.electronics<br>sci.space<br>talk.politics.misc |

large scale, the data out-of-domain and in-domain may be subject to a same wide type. From this point of view, both conventional clustering methods and semi-supervised methods can be modified to tackle this transfer learning problem. What we borrow is the spectral clustering method.

### 3.2 Spectral Learning

When we represent the data as a graph, spectral analysis is easily adopted to investigate the structure of these data. The normalized cut [Shi and Malik, 2000] is one of the most representative work for spectral clustering. [Kamvar *et al.*, 2003] extends this type of methods to more learning situations, e.g. under pairwise link constraints or with labeled examples. Instead of clustering data using homogeneous relations, co-clustering techniques have been developed to cluster data using heterogeneous information with bipartite relationships, which are mainly used for document clustering [Dhillon, 2001]. Recently, spectral clustering has been extended to multi-type relational data and many well-known algorithms are unified into a general framework [Long *et al.*, 2006; 2007].

In next section, we will compare the mainly related methods described in this section with our algorithms for our transfer learning problem.

## 4 Experiments

In the experiments, we use the real text data to demonstrate the effectiveness of our algorithms.

**20-NewsGroups Data:** The 20-newsgroups data set collects approximately 20,000 documents across 20 different newsgroups. It is widely used to test the performance of text mining algorithms. The data is preprocessed as [Zhong and Ghosh, 2005]. And we set up the cross domain transfer learning data sets in a similar way as [Dai *et al.*, 2007; Ling *et al.*, 2008], which focus on binary problems[3]. It means $K_i = K_o = 2$. Table 3 shows the data sets in the experiments.

**Compared Methods:** Most of the related state-of-the-art methods are compared. They are support vector machine (SVM), LapSVM for manifold learning [Belkin *et al.*, 2006], transductive SVM (TSVM) [Joachims, 1999], S-Kmeans [Dhillon and Modha, 2001], normalized cut (NCut) [Shi and Malik, 2000], spectral co-clustering (S-Co-C), information theoretical co-clustering (IT-Co-C), CoCC [Dai *et al.*, 2007] and CDSC [Ling *et al.*, 2008].

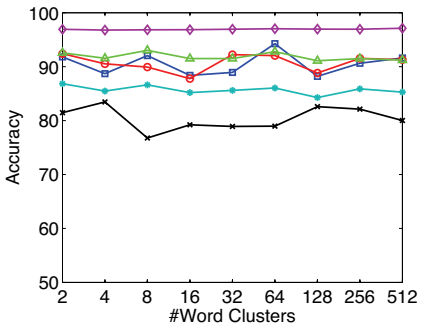



Figure 2: The accuracy curve over the number of word clusters for all six data sets in similarity propagation.

**Experimental Result:** The supervised method is trained in-domain and tested using out-of-domain data; The semi-supervised methods are applied in a transductive setting using all available data. The clustering methods are conducted directly for out-of-domain data. As the existence of supervised information, all clustering methods can assign the cluster in their result with the most relevant in-domain class as in [Dai *et al.*, 2007]. Then the learning accuracy can be used as the performance measure for all compared methods. For the parameter setting, $K_w = K_i + K_o = 4$ is fix for label propagation as the algorithm demands. In similarity propagation, the result is not sensitive to $K_w$, which is shown in Fig 2. These make our methods have few heuristic factors and easy to apply. The parameters for all other compared methods are set to a relatively best one according to the referenced papers.

Each experiment has 20 repeated runs and the average learning accuracy with standard deviation is presented in Table 4. S-Co-C-LP is our label propagation algorithm, and S-Co-C-SP is the similarity propagation. The results show our methods perform the best, which approves our analysis in section 2. Note that the normalized cut sometimes gets the similar good result, which means the unlabeled data set is well discriminated itself.

## 5 Conclusion and Discussion

In this paper, we analysis the across domains transfer learning problem, establish a framework of transfer learning with similarity matrix on hybrid graph, and propose two practical algorithms. During the transfer process, we seek all possible approaches to transfer the useful information from the in-domain data to the out-of domain data. The experiments on text data mining show the efficiency of our method.

[3]Our algorithm is designed using arbitrary class numbers to design a general framework. However, the compared methods are binary. We will leave the multi-class case for future study.

Table 4: Learning accuracy for 20-newsgroups data ($mean \pm std\%$).

| | comp vs rec | comp vs sci | comp vs talk | rec vs sci | rec vs talk | sci vs talk |
|---|---|---|---|---|---|---|
| SVM | 85.31(9.16) | 72.15(5.09) | 95.75(1.20) | 73.44(8.58) | 82.49(8.76) | 78.41(7.06) |
| LapSVM | 86.23(7.50) | 71.78(5.05) | 95.40(1.24) | 72.60(8.09) | 83.01(7.90) | 73.81(8.85) |
| TSVM | 91.51(2.26) | 80.63(4.35) | 96.88(0.88) | 86.05(5.60) | 92.50(2.06) | 86.39(4.76) |
| S-Kmeans | 82.81(9.16) | 76.18(9.51) | 93.07(8.56) | 65.30(12.02) | 83.79(13.28) | 74.79(12.73) |
| NCut | 89.44(7.30) | 81.88(9.18) | **96.08(0.74)** | 76.15(8.34) | **93.86(2.16)** | 88.69(6.65) |
| S-Co-C | 79.21(18.39) | 68.56(14.84) | 89.96(15.91) | 80.59(11.42) | 87.26(13.79) | 82.99(13.87) |
| IT-Co-C | 86.31(12.85) | 80.65(10.79) | 91.74(9.63) | 78.70(14.67) | 82.06(10.85) | 80.85(11.54) |
| CoCC | 88.75(7.67) | 81.10(6.78) | 96.33(3.82) | 82.48(7.84) | 91.49(5.71) | 82.21(6.82) |
| CDSC | 82.55 ( 7.24) | 78.70(4.35) | 96.23(4.8) | 82.65(3.32) | 92.43 (5.49) | 83.41 (4.14) |
| S-Co-C-LP | 91.94(2.72) | **83.59(3.67)** | **96.97(0.88)** | 86.41(3.62) | 92.96(2.13) | 88.20(4.35) |
| S-Co-C-SP | **93.55(6.91)** | 80.31(11.59) | **96.99(1.04)** | **89.49(6.79)** | **94.32(2.22)** | **90.00(5.11)** |

This paper solves the problem of how to transfer and how to make optimal use of the available information. However, it assumes that the data in different domains are closely related to make the domain-transfer learning feasible. In practical problems, it is still inconvenient for users to find the similar tasks themselves. So it is more contributive to solve the what to transfer problem. This is an interesting and challenge issue for our future study.

## Acknowledgments

We gratefully acknowledge the helpful discussion and insightful advice of Chih-Jen Lin, Edward Y. Chang and Qiang Yang. We also thank Xiao Ling for his discussion about the implementation of his algorithm. This research was supported by National Science Foundation of China ( No. 60835002 and No. 60721003 ).

## References

[Belkin *et al.*, 2006] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 1(1):1–48, 2006.

[Bhatia, 1997] R. Bhatia. *Matrix analysis*. Springer-Cerlag, 1997.

[Caruana, 1997] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[Dai *et al.*, 2007] Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu. Co-clustering based classification for out-of-domain documents. In *SIGKDD*, pages 210–219, 2007.

[Dhillon and Modha, 2001] Inderjit S. Dhillon and Dharmendra S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1–2):143–175, 2001.

[Dhillon *et al.*, 2003] Inderjit S. Dhillon, Subramanyam Mallela, and Dharmendra S. Modha. Information-theoretic co-clustering. In *SIGKDD*, pages 89–98, 2003.

[Dhillon, 2001] Inderjit S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *SIGKDD*, pages 269–274, 2001.

[Joachims, 1999] Thorsten Joachims. Transductive inference for text classification using support vector machines. In *ICML*, pages 200–209, 1999.

[Kamvar *et al.*, 2003] Sepandar D. Kamvar, Dan Klein, and Christopher D. Manning. Spectral learning. In *IJCAI*, pages 561–566, 2003.

[Ling *et al.*, 2008] Xiao Ling, Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu. Spectral domain-transfer learning. In *SIGKDD*, pages 488–496, 2008.

[Long *et al.*, 2006] Bo Long, Zhongfei (Mark) Zhang, Xiaoyun Wu, and Philip S. Yu. Spectral clustering for multi-type relational data. In *ICML*, pages 585–592, 2006.

[Long *et al.*, 2007] Bo Long, Zhongfei (Mark) Zhang, and Philip S. Yu. A probabilistic framework for relational clustering. In *SIGKDD*, pages 470–479, 2007.

[Ormrod, 2004] J. E. Ormrod. *Human learning (4th Ed.)*. Pearson, 2004.

[Pan and Yang, 2007] Sinno Jialin Pan and Qiang Yang. Introduction to transfer learning. Technical report, Computer Science and Engineering, HKUST, Hong Kong, 2007.

[Shi and Malik, 2000] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE TPAMI*, 22(8):888–905, 2000.

[Sugiyama *et al.*, 2007] M. Sugiyama, M. Krauledat, and K.-R. Mller. Covariate shift adaptation by importance weighted cross validation. *JMLR*, 8(8):985–1005, 2007.

[Williams and Seeger, 2001] C. K. I. Williams and M. Seeger. Using the nyström method to speed up kernel machines. In *NIPS*, pages 682–688, 2001.

[Yu and Shi, 2003] Stella X. Yu and Jianbo Shi. Multiclass spectral clustering. In *ICCV*, pages 313–319, 2003.

[Zha *et al.*, 2001] Hongyuan Zha, Chris Ding, Ming Gu, Xiaofeng He, and Horst Simon. Spectral relaxation for k-means clustering. In *NIPS*, pages 1057–1064, 2001.

[Zhong and Ghosh, 2005] Shi Zhong and Joydeep Ghosh. Generative model-based clustering of documents: a comparative study. *KAIS*, 8:374–384, 2005.

[Zhu, 2005] Xiaojin Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.