

Label Propagation on K-partite Graphs

Chris Ding*

Tao Li†

Dingding Wang‡

Abstract

Label propagation is an approach to assign class labels to unlabeled data given some partially labeled data. In this paper, we systematically generalize the Laplacian matrix based label propagation method from pairwise graph data to data objects described by bipartite and general K -partite graphs. By deriving explicit label propagation formula, we show how information on one type of variables can be transformed to other types of variables. For example, in a word-document-author multi-relational dataset, information on words and on authors can effectively enhance the document labeling. Motivating examples are presented to illustrate these new concepts. Extensive experiments are performed on real-life datasets to show the effectiveness of our label propagation.

1 Introduction

With an explosive amount of data being accumulated, labeling data for many supervised learning and data mining tasks requires extensive human skills and extensive human labors. As a consequence, most of the available data are in fact unlabeled or partially labeled. Thus semi-supervised learning becomes a very active research area. In semi-supervised learning, we have a large amount of unlabeled data, and only a very small fraction of them are labeled. The learning task is to classify the unlabeled data based on the labeled data.

There exists a very large number of semi-supervised learning methods (see a survey [23]): (a) The classification-based approach, in which a classifier is first trained on the small labeled data and is gradually improved by incorporating unlabeled data. Earlier methods mostly follow this approach [3]. (b) The clustering-based approach, in which a clustering algorithm is used on the whole data (labeled and unlabeled), with the labeled data serving as penalty or regularization or prior information. Recent methods mostly follow this approach, such as spectral clustering based methods [11, 1, 4]. (c) Special mechanisms, such as Gaussian process [13], graph mincut [2], entropy minimization [10], label sampling [17], and nonparametric transforms [25], etc. However, most if not all of these methods are focusing on homogeneous data of the same type (the meaning will become clear very shortly).

In many real world applications, however, there are often relational data and a typical task often involves more than one type of data points [15]. For example, in document analysis, there are *terms* and *documents* and the data is represented as word-document matrices. We wish to ask the following question. Suppose we have partially labeled data on the word side, would that be helpful to label the documents? Or if we have partially labeled data on both document and word sides, how to make use both of them simultaneously?

Recently, a mechanism based on matrix factorization was provided in [14] to enable additional information/knowledge on the word side to influence/help the clustering of documents. However, the model described in [14] is only focused on unsupervised learning tasks on bipartite heterogeneous relational data. It does not incorporate dual supervision and thus is not able to make use of labeled data samples.

So far, most of these semi-supervised learning are restricted to data of one type, using a single pairwise similarity matrix (graph). Many other type of data appear often as well. A document-word table is a good example of bipartite type relations among word and documents. More general K -partite graph also occurs frequently, such as document-word-author relations. This problem can be modeled as 3D tensors and tri-partite graph.

A general k -partite graph is also sometimes called multi-relational data. Multi-relational data mining approaches have been developed for association mining and classification from datasets involving multiple tables (relations) from a relational database [7, 21]. Probabilistic relational learning methods are studied in [9, 16]. However, these multi-relational data mining methods are not designed to deal with semi-supervised learning tasks on relational data. To our knowledge, the area of semi-supervised learning for these bipartite, k -partite graphs data has not been widely investigated. We note that the very recent work in [18] proposes a dual supervision model for semi-supervised sentiment analysis using bipartite graph regularization.

In this paper, we explore this new area. Semi-supervised learning tasks on bipartite data have an interesting and unique feature which do not appear on standard data. A bipartite data has two data object types, say word object space and document object space in a word-document relational table. In this paper, our main contribution is to show that partial knowledge on one data space can help knowledge discovery on the other data space. We call this cross-propagation (see §3).

A K -partite graph has several bi-partite tables. We will show both theoretically and experimentally, that partial knowledge on one data space can help knowledge discovery on all other data spaces (see §4.1). Technically, we propose a label propagation

*CSE Department, University of Texas at Arlington.

†School of Computer Science, Florida International University.

‡School of Computer Science, Florida International University.

framework to carry out the semi-supervised learning on bipartite graphs and on general K -partite graphs which is a model for multi-relational data.

1.1 Organization of The Paper

In §2, we review the label propagation theory based on *Reproducing Kernel Hilbert Space* theory, a regularization theory on a simple similarity graph. The emphasis is the importance of the kernel function of the inverse Laplacian for label propagation. In §3, we generalize this to bipartite graph, and derive several concrete relations for label propagation connecting the two sides (two types of nodes on the bipartite graph). We show theoretically how knowledge on one side could help the learning of the other side. In §4, we generalize this to general K -partite graph, and derive several concrete relations for label propagation connecting the K sides. We also show the consistency between these approaches. In §5, we present systematic experimental results on the label propagation on bipartite and tri-partite datasets. Our results demonstrate the benefits of using bi- and tri- partite information to help learning on other sides.

2 Label Propagation Theory

Our label propagation approach is motivated by the work of Zhou et al [22]. We derive it from the theory of Reproducing Kernel Hilbert Space (RKHS) [8] at strong regularization limit.

Suppose we have $n = n_l + n_u$ data $\{\mathbf{x}_i\}_{i=1}^n$ where the first n_l data have class labels $\{h_i\}_{i=1}^{n_l}$, and the second n_u data are unlabeled. Our task is to assign class labels to those unlabeled data.

We first consider 2-class problems, where $h_i = \pm 1$ for labeled data. We initially set $h_i = 0$ for unlabeled data. Besides the partially labeled data, we have a matrix of pairwise similarities $W = (w_{ij})$ among the data point. W which can be viewed as the edge weights on a graph. Our task is to learn the classification function $h_i = f(\mathbf{x}_i)$.

To be consistent, one requirement is that for the error for the labeled data

$$\sum_{i=1}^{n_l} [h_i - f(\mathbf{x}_i)]^2$$

is minimized. For the unlabeled data x_i , since initial label is $h_i = 0$ and $f(\mathbf{x}_i) = \pm 1$, $\sum_{i=n_l+1}^n [h_i - f(\mathbf{x}_i)]^2 = n_u$ is a constant. Thus we may minimize the total squared error:

$$\sum_{i=1}^n [h_i - f(\mathbf{x}_i)]^2 = \|\mathbf{h} - \mathbf{f}\|^2,$$

where $\mathbf{f} = (f_1, \dots, f_n)^T$, $f_i = f(\mathbf{x}_i)$, and $\mathbf{h} = [h_1, h_2, \dots, h_n]^T$. In statistics, we often add a penalty (regularization) term to ensure smoothness of certain quantities such as derivatives. RKHS seeks the function $f(\cdot)$ that minimizes

$$J[\mathbf{f}] = \sum_{i=1}^n [h_i - f(\mathbf{x}_i)]^2 + \beta \mathbf{f}^T \mathcal{K}^{-1} \mathbf{f} \quad (1)$$

RKHS theory is equivalent to the uniform convergence theory of Vapnik [20]. When the loss function $[h_i - f(\mathbf{x}_i)]^2$ is replaced by

the hinge function

$$[1 - h_i f(\mathbf{x}_i)]_+,$$

the dual space solution gives SVM.

Solution of RKHS for the quadratic loss function is

$$\mathbf{f} = \mathcal{K}(\mathcal{K} + \beta I)^{-1} \mathbf{h}. \quad (2)$$

At the large β limit, we get the solution

$$\mathbf{f} = \frac{1}{\beta} (\mathcal{K} - \frac{1}{\beta} \mathcal{K}^2 + \frac{1}{\beta^2} \mathcal{K}^3 - \dots) \mathbf{h}. \quad (3)$$

In this paper, we use only the first term. Since the proportional constant $1/\beta$ is unimportant, we obtain the final label propagation results:

$$\mathbf{f} = \mathcal{K} \mathbf{h}. \quad (4)$$

Standard kernel machines are used for supervised learning with $h_i = \pm 1$. For semi-supervised learning we set $h_i = 0$ for those unlabeled data.

2.1 Combinatorial Laplacian

Given a graph with n nodes and edge weights $W = (W_{ij})$, where W_{ij} represents similarity between nodes i, j . The *combinatorial Laplacian* is defined to be

$$L = D - W,$$

where the diagonal matrix contains row sums of W : $D = \text{diag}(d_1, \dots, d_n)$, $d_i = \sum_j W_{ij}$. Let eigenvectors of L be:

$$L \mathbf{q}_k = \lambda_k \mathbf{q}_k, \quad \mathbf{q}_p^T \mathbf{q}_k = \delta_{pk}. \quad (5)$$

where $0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ are the eigenvalues. We assume the graph is connected (otherwise we deal with each connected component one at a time). L has a zero mode, i.e., the eigenvector

$$\mathbf{q}_1 = \mathbf{e}/\sqrt{n}, \quad \mathbf{e} = (1 \dots 1)^T. \quad (6)$$

with zero eigenvalue: $(D - W) \mathbf{q}_1 = 0$. Note this is a constant function on the nodes of the graph. This is a critical fact that will be critically important later on.

We define the Laplacian Kernel as

$$K = (D - W)_+^{-1} \quad (7)$$

where the positive part $(\cdot)_+$ implies that the zero mode is excluded.

\mathcal{K} is a kernel. First, \mathcal{K} is clearly a semi-positive definite function. Second, any function $\mathbf{f} \in \mathfrak{R}^n$ can be expanded in the basis of \mathcal{K} , i.e. $(\mathbf{q}_2, \dots, \mathbf{q}_n)$ plus a constant $\mathbf{e}/\sqrt{n} = \mathbf{q}_1$.

Using the Laplacian kernel, our label propagation results Eq.(4) becomes

$$\mathbf{f} = \frac{1}{(D - W)_+} \mathbf{h}. \quad (8)$$

Our work is partially motivated by the work of Zhou et al [22] who propose the consistency framework that minimizes the functional

$$J_1[\mathbf{f}] = \mu \sum_{i=1}^n [h_i - f_i]^2 + \frac{1}{2} \sum_{i,j=1}^n \left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right)^2 w_{ij}. \quad (9)$$

Note that d_i explicitly appear in the second term of Eq.(9), but not in the first term — somehow d_i is not treated symmetric. More importantly, the strong regulation limit of the solutions are different.

2.2 Label Propagation for Multi-class Problems

In above analysis for 2-class semi-supervised learning, we use a single vector \mathbf{h} on n data points:

$h_i = \pm 1$ for data point \mathbf{x}_i with known labels;

$h_i = 0$, for data point \mathbf{x}_i with unknown labels.

as the input. After computing \mathbf{f} according to Eq.(8), we interpret the results as x_i belongs to C_+ class if $f_i > 0$; x_i belongs to C_- class if $f_i < 0$.

In label propagation for multi-class problem, we write

$$F = [D - W]^{-1}H$$

where $H = (\mathbf{h}_1, \dots, \mathbf{h}_K)$ for K classes. The input H is specified as: If labeled data point \mathbf{x}_i has a known class label k , $H_{ik} = 1$; $H_{ik} = 0$ otherwise. After F is computed, we assign x_i to class $k = \arg \max_j F_{ij}$.

3 Cross Propagation on Bipartite Graphs

A bipartite graph can be represented by an $m \times n$ rectangular nonnegative matrix (a contingency table) P . we view it as the adjacency matrix of a bipartite graph. There are two type of nodes: x -nodes (each is represented by a column in the table P , and y -nodes (each is represented by a row in the table P). P_{ij} denotes the weight (similarity) between y -node i and x -node j .

Suppose we have $n = n_l + n_u$ x -node data $\{\mathbf{x}_i\}_{i=1}^n$ where the first n_l data has class labels $\{h_i^x\}_{i=1}^{n_l}$. Suppose we have $m = m_l + m_u$ data $\{\mathbf{y}_i\}_{i=1}^m$ where the first m_l data has class labels $\{h_i^y\}_{i=1}^{m_l}$. The task is to assign class labels to unlabeled x -nodes and unlabeled y -nodes.

Although we expect that the labeled x -nodes will affect the labeling of unlabeled x -nodes, we will show that the labeled y -nodes will affect the labeling of unlabeled x -nodes. This interesting feature of learning on a bipartite graph is called cross-propagation.

We first consider 2-class problems, where $h_i = \pm 1$ for labeled data. The label propagation is to minimize the following clustering objective

$$J_2 = \sum_{i=1}^m (f_i^x - h_i^x)^2 + \sum_{j=1}^n (f_j^y - h_j^y)^2 + \beta \sum_{ij} (f_i^x - f_j^y)^2 P_{ij}, \quad (10)$$

which is a generalization of Eq.(1). Let

$$d_i^x = \sum_{j=1}^n P_{ij}, \quad d_j^y = \sum_{i=1}^m P_{ij}.$$

Denote

$$\mathbf{f} = \begin{pmatrix} \mathbf{f}^x \\ \mathbf{f}^y \end{pmatrix}, \quad \mathbf{h} = \begin{pmatrix} \mathbf{h}^x \\ \mathbf{h}^y \end{pmatrix}, \quad (11)$$

where

$$W = \begin{pmatrix} 0 & P \\ P^T & 0 \end{pmatrix}, \quad D = \begin{pmatrix} D^x & 0 \\ 0 & D^y \end{pmatrix}, \quad (12)$$

and

$$D^x = \text{diag}(d_1^x, \dots, d_n^x), \quad D^y = \text{diag}(d_1^y, \dots, d_m^y).$$

We have

$$\mathbf{f} = [D - W]_+^{-1} \mathbf{h}. \quad (13)$$

It can be shown that

$$(D - W)^{-1} =$$

$$\begin{pmatrix} (D_x - P D_y^{-1} P^T)^{-1} & (D_x - P D_y^{-1} P^T)^{-1} P D_y^{-1} \\ (D_y - P^T D_x^{-1} P)^{-1} P D_x^{-1} & (D_y - P^T D_x^{-1} P)^{-1} \end{pmatrix}$$

Thus we obtain

$$\begin{bmatrix} \mathbf{f}^x \\ \mathbf{f}^y \end{bmatrix} = \begin{bmatrix} (D_x - P D_y^{-1} P^T)^{-1} (\mathbf{h}_x + P D_y^{-1} \mathbf{h}_y) \\ (D_y - P^T D_x^{-1} P)^{-1} (\mathbf{h}_y + P^T D_x^{-1} \mathbf{h}_x) \end{bmatrix} \quad (14)$$

where \mathbf{h}_x represents the partial labels for rows of P and \mathbf{h}_y represents the partial labels for columns of P .

This equation is interesting in several aspects: (a) For row labels \mathbf{h}^x , $(D_x - P D_y^{-1} P^T)^{-1}$ is the appropriate kernel function. (b) There are two terms contributing to \mathbf{f}^x . The first term involving \mathbf{h}_x is the standard one. The second term involving \mathbf{h}_y suggests that the partial labels on column variables propagate to row variables. Suppose P is a word-document matrix. This says partial labels on documents can also determine the labels on words. Similar results apply to column labels \mathbf{f}^y .

This equation is one of our main results of this work. It shows how the label information on one type of nodes is propagated to the other type of nodes. In a word-document relation, the results suggest label information on words (such as partially labeled information) can influence the labeling of documents. We call this *cross propagation*.

4 Label Propagation on K-partite Graphs

The label propagation on bipartite graph can be generalized to K -partite graph. A K -partite graph has K type nodes \mathbf{x}^p , $p = 1, \dots, K$. Each type has $(x_1^p, \dots, x_{n_p}^p)$ nodes. The K -partite graph contains up to $K(K-1)/2$ bipartite (two-way) similarity relations P^{pq} , $p, q = 1, \dots, K$. For example, P^{12} contains all $n^1 n^2$ similarities between type-1 nodes $(x_1^1, \dots, x_{n_1}^1)$ and type-2 nodes $(x_1^2, \dots, x_{n_2}^2)$.

In semi-supervised learning, we assume that the first n_l^p of type- p nodes are labeled with class labels $(h_1^p, \dots, h_{n_l^p}^p)$. The task is to predict the labels for the unlabeled nodes of all types simultaneously.

Denote

$$\mathbf{f}^p = \begin{pmatrix} f_1^p \\ \vdots \\ f_{n_p}^p \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} \mathbf{f}^1 \\ \vdots \\ \mathbf{f}^K \end{pmatrix}.$$

Let

$$D = \text{diag}(D^1, D^2, \dots, D^K), \quad D^p = \text{diag}(d_1^p, d_2^p, \dots, d_{n_p}^p),$$

where $d_i^p = \sum_{q \neq p} \sum_j P_{ij}^{pq}$; and

$$W = \begin{pmatrix} 0 & P^{12} & \dots & P^{1K} \\ (P^{12})^T & 0 & \dots & P^{2K} \\ \dots & \dots & \dots & \dots \\ (P^{1K})^T & (P^{2K})^T & \dots & 0 \end{pmatrix}.$$

The corresponding solution is given by Eq.(13).

4.1 Cross Propagation on a Tri-partite Graph

Let us consider the tri-partite graph. An example application is word-document-author relationship. Let x represents words, y represents documents, and z represents authors. In many situations, the relation $A = P^{12}$ between words and documents is known, the relation $B = P^{23}$ between authors and documents is known, but the relation C between word and authors is unknown. Thus we set $C = P^{32} = 0$.

The similarity matrix W and D matrix are

$$W = \begin{pmatrix} 0 & A & 0 \\ A^T & 0 & B \\ 0 & B^T & 0 \end{pmatrix}, \quad D = \begin{pmatrix} D_x & & \\ & D_y & \\ & & D_z \end{pmatrix} \quad (15)$$

In general, we directly solve the kernel function $\mathcal{K} = (D - W)_+^{-1}$, and obtain the label propagation

$$\begin{pmatrix} \mathbf{f}_x \\ \mathbf{f}_y \\ \mathbf{f}_z \end{pmatrix} = \mathcal{K} \begin{pmatrix} \mathbf{h}_x \\ \mathbf{h}_y \\ \mathbf{h}_z \end{pmatrix} \quad (16)$$

For analytic solution, we have

$$(D - W)^{-1} = D^{-1/2}(I - \widehat{W})^{-1}D^{-1/2} \quad (17)$$

We can show that

$$(I - \widehat{W})^{-1} = \begin{bmatrix} I + \widehat{A}\Delta\widehat{A}^T & \widehat{A}\Delta & \widehat{A}\Delta\widehat{B} \\ \Delta\widehat{A}^T & \Delta & \Delta\widehat{B} \\ \widehat{B}^T\Delta\widehat{A}^T & \widehat{B}^T\Delta & I + \widehat{B}^T\Delta\widehat{B} \end{bmatrix} \quad (18)$$

where

$$\Delta = (I - \widehat{A}^T\widehat{A} - \widehat{B}\widehat{B}^T)^{-1}$$

represents the influence propagation between different classes.

From this, we obtain the components \mathcal{K} . For example,

$$\mathcal{K}_{xx} = D_x^{-1/2}(I + \widehat{A}\Delta\widehat{A}^T)D_x^{-1/2} \quad (19)$$

is the self-propagation from \mathbf{h}_x to \mathbf{f}_x .

$$\mathcal{K}_{xz} = D_x^{-1/2}\widehat{A}\Delta\widehat{B}D_x^{-1/2}$$

is the double cross-propagation from \mathbf{h}_z to \mathbf{f}_x . We call this *double cross propagation* because $P^{13} = 0$, and the propagation is a two-step propagation: first from z nodes to y nodes using $B = P^{23}$ and second from y nodes to x nodes using $A = P^{12}$.

This results have several interesting features. First, the label or partial label information on authors (z) can infer the label information on words (x), even though there is no direct link between words and documents. This is indicated in the term $A\Delta B\mathbf{h}_z$, which shows the author label information is passed from z -nodes on to y -nodes using weight matrix B . Then using Δ to transfer into matrix A and to nodes x .

Consistency with bi-partite propagation

Second, we demonstrate that when $B = 0$, the tri-partite propagation reduces to the bi-partite propagation. We show that \mathcal{K}_{xx} of Eq.(19) is identical to

$$(D_x - PD_y^{-1}P^T)^{-1} \quad (20)$$

according to Eq.(14). The proof is omitted due to the space limit.

5 Experiments

In this section, we present our experimental results on four real world datasets to evaluate our proposed cross label propagation methods.

5.1 Dataset Description

The following four datasets are used in our experiments.

- **DBLP Dataset:** This dataset is obtained from DBLP Computer Science Bibliography¹. We extract the paper titles published by 552 relatively productive researchers from 9 categories. Using the ACM Keywords Taxonomy², we obtain the category information for terms and use it as the prior knowledge in the word space.
- **CSTR Dataset:** This dataset contains the abstracts of technical reports (TRs) published in the Department of Computer Science at a research university from 1991 to 2007. There are 550 abstracts and they are divided into four research areas. We also use the category information of terms obtained from ACM Keywords Taxonomy as prior knowledge.
- **Citeseer Dataset:** The real-world data set for experimentation was generated by sampling documents from CiteSeer. We collected the documents by those top authors in CiteSeer ranked by their numbers of documents. Then we collected the venues of these documents. We kept those venues with most documents in the prepared subset and discarded the venues that include fewer documents. The sampled dataset includes 1000 documents, 2500 words and 681 authors.
- **BBS Dataset:** This is a dataset sampled from the *Bulletin Board Systems (BBS)* data in [12]. A BBS system contains many boards with similar themes. Once a user posts an initial article on a board, the others can show their opinions using reply articles. The initial article and reply articles constitute a topic. People's behaviors on the BBS usually reflect their interests. The dataset contains 1309 users, 1200 topics and 12 boards.

5.2 Evaluation Measure

We use accuracy and normalized mutual information (NMI) as performance measures. Accuracy discovers the one-to-one relationship between clusters and classes and measures the extent to which each cluster contains data points from the corresponding class and it has been used as performance measures for clustering analysis [6]. Generally, the greater accuracy, the better clustering performance. NMI [19] measures the amount of statistical information shared by two random variables representing cluster assignment and underlying class label. In general, the larger the NMI value, the better the clustering quality.

¹The dblp.xml file is available for download at <http://www.informatik.uni-trier.de/~ley/db/>.

²Available on the page of <http://www.computer.org/portal/pages/ieeecs/publications/author/ACMtaxonomy.html>.

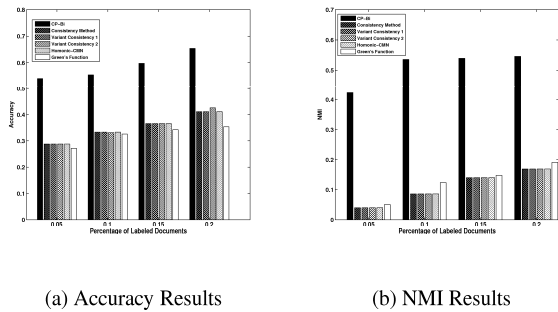


Figure 1. Performance results on DBLP dataset.

5.3 Experimental Results on Bi-partite Graphs

In this section, we compare our method (denoted by CP-Bi: Cross Propagation on Bi-partite graphs) with five semi-supervised clustering approaches: (1 ~ 3) The algorithm proposed in [22] which conducts semi-supervised learning with local and global consistency (Consistency Method), and two of its variants (Variant Consistency 1 and Variant Consistency 2) [22]; (4) Zhu et al.’s harmonic Gaussian field method coupled with the Class Mass Normalization (Harmonic-CMN) [24]; (5) Green’s function learning algorithm (Green’s Function) proposed in [5]. All of these methods do not make use of the knowledge in the word space.

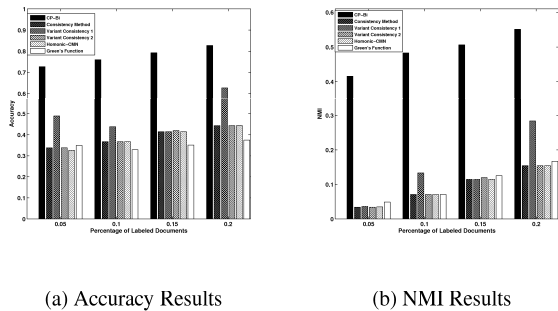


Figure 2. Performance results on CSTR dataset.

5.3.1 Experiments Using Prior Knowledge

In this set of experiments, we compare our proposed CP-Bi algorithm with other semi-supervised algorithms on four datasets as described in Section 5.1. Ten percent of the words in each dataset are labeled and used as prior knowledge. For each dataset, we also vary the percentage of the labeled documents from 5% to 20%. The performance results on the four datasets are presented in Figure 1, Figure 2, Figure 3, and Figure 4, respectively. The results are

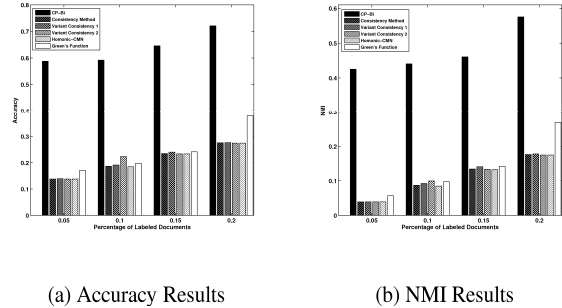


Figure 3. Performance results on Citeseer dataset.

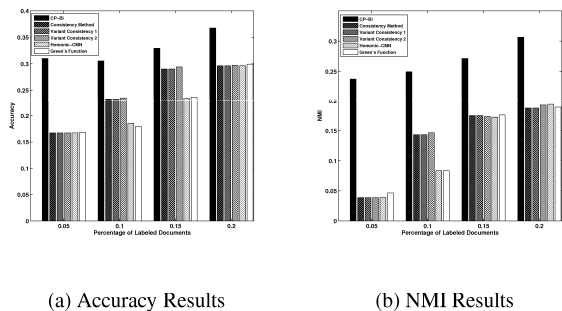


Figure 4. Performance results on BBS dataset.

obtained by averaging 15 trials. From the experimental comparison, we observe that: (1) The performance of the semi-supervised clustering algorithms generally increases as the percentage of labeled documents increases. (2) Our cross propagation method utilizing word space prior knowledge outperforms other algorithms which do not take such knowledge into consideration. The comparison demonstrates that cross label propagation enables the information on one type of variables to be transferred to other types of variables.

5.3.2 Percentage of Labeled Words

In this set of experiments, we perform CP-Bi on DBLP dataset and CSTR dataset to investigate the effects of the percentage of labeled words on clustering quality. We fix the percentage of labeled documents in each dataset as 10% and the results are also obtained by averaging 15 trials. Figure 5 and Figure 6 clearly show that while the percentage of labeled words increases, the accuracy and NMI results on both datasets are improved.

5.4 Experimental Results on Tri-partite Graph

In this section, we compare the proposed cross label propagation method (CP-Tri) which uses both word space and author space

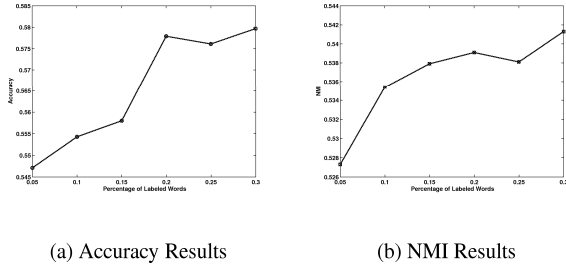


Figure 5. Results with different percentage of labeled words on DBLP dataset.

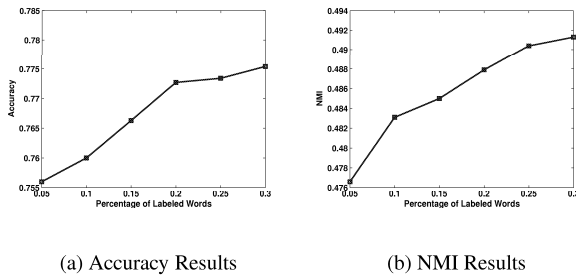


Figure 6. Results with different percentage of labeled words on CSTR dataset.

knowledge with the proposed method (CP-Bi) which uses word space knowledge and other semi-supervised algorithms which do not make use of any external information on BBS and Citeseer datasets. In each dataset, ten percent of documents, words and authors are labeled and they are used as prior knowledge. The results are obtained by averaging 15 trials. From Figure 7, we observe that using more external prior knowledge can improve the clustering performance.

6 Conclusion

In this paper, we show how information on one type of variables can be transformed to other types of variables by deriving explicit label propagation formula. We systematically generalize the Laplacian embedding to bipartite and general K -partite graphs and present algorithms for simultaneous clustering of multi-relational data, and semi-supervised label propagation. Extensive experiments performed on real-life datasets on demonstrate the effectiveness of our label propagation approach.

Acknowledgement

The work of C. Ding is partially supported by NSF grants CCF-0939187 and CCF-0830780 and the work of D. Wang and T. Li

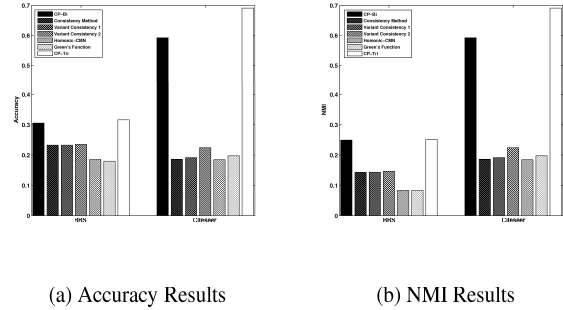


Figure 7. Results on Tri-partite graph.

is partially supported by NSF grants IIS-0546280, CCF-0939179, and CCF-0830659.

References

- [1] M. Belkin and P. Niyogi. Semi-supervised learning on riemannian manifolds. *Machine Learning*, pages 209–239, 2004.
- [2] A. Blum, J. Lafferty, M. Rwebangira, and R. Reddy. Semi-supervised using randomized mincuts. *Proc. ICML 2004*.
- [3] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. *Proc. CLT1998*.
- [4] O. Chapelle, J. Weston, and B. Schölkopf. Cluster kernels for semi-supervised learning. *Proc. NIPS 2002*.
- [5] Chris Ding, Rong Jin, Tao Li, and Horst D Simon. A learning framework using green's function and kernel regularization with application to recommender system. In *Proceedings of KDD 2007*.
- [6] Chris Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proceedings of KDD 2006*.
- [7] Sašo Džeroski. Multi-relational data mining: an introduction. *SIGKDD Explorer Newsl.*, 2003.
- [8] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Adv. Comp. Math.*, 2000.
- [9] Lise Getoor, Nir Friedman, Daphne Koller, and Benjamin Taskar. Learning probabilistic models of relational structure. In *ICML 2001*.
- [10] Y. Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. In *NIPS 2004*.
- [11] T. Joachims. Transductive learning via spectral graph partitioning. *Proc. ICML 2003*.
- [12] Zhongbao Kou and Changshui Zhang. Reply networks on a bulletin board system. *Phys. Rev. E*, 2003.
- [13] N.D. Lawrence and M.I. Jordan. Semi-supervised learning via gaussian process. *NIPS 2004*.
- [14] Tao Li, Chris Ding, Yi Zhang, and Bo Shao. Knowledge Transformation from Word Space to Document Space. *SIGIR 2008*
- [15] Bo Long, Xiaoyun Wu, Zhongfei (Mark) Zhang, and Philip S. Yu. Unsupervised learning on k-partite graphs. In *Proceedings of KDD 2006*.
- [16] Qing Lu and Lise Getoor. Link-based classification using labeled and unlabeled data. In *ICML 2003 Workshop on The Continuum from Labeled to Unlabeled Data*, 2003.
- [17] S. Rosset, J. Zhu, H. Zou, and T. Hastie. A method for inferring label sampling mechanisms in semi-supervised learning. *Proc. NIPS 2004*.
- [18] Vikas Sindhwani and Jianying Hu and Aleksandra Mojsilovic. Regularized Co-Clustering with Dual Supervision. *NIPS 2008*
- [19] Alexander Strehl and Joydeep Ghosh. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 2003.
- [20] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.
- [21] Xiaoxin Yin, Jiawei Han, Jiong Yang, and Philip S. Yu. Efficient classification across multiple database relations: A crossmine approach. *IEEE Transactions on Knowledge and Data Engineering*, 2006.
- [22] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. *Proc. NIPS 2003*.
- [23] X. Zhu. Semi-supervised learning literature survey. *University of Wisconsin CS TR-1530*, 2006.
- [24] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of ICML*, 2003.
- [25] X. Zhu, J. Kandola, Z. Ghahramani, and J. Lafferty. Nonparametric transforms of graph kernels for semi-supervised learning. *NIPS 2000*.