

Knowledge Transformation by Cross-Domain Belief Propagation

Fei Wang Tao Li

School of Computing & Information Sciences
Florida International University, Miami, FL, 33199
{feiwang,taoli}@cs.fiu.edu

Abstract—Belief propagation is an iterative algorithm for computing marginals of functions on a graphical model most commonly used in information retrieval. In this paper, we consider the problem of performing cross-domain belief propagation on multi-relational data for semi-supervised learning. We demonstrate that partial knowledge on one type of variables can help knowledge discovery on the other type of variables with cross-domain belief propagation by utilizing the existing relationships in multi-relation data. For example, in a word-document data set, information on the word domain can effectively enhance the labeling of document domain. In this paper, we explore this new area, knowledge transformation of multi-relation data for semi-supervised learning tasks. We show that partial knowledge on one data variable domain can help knowledge discovery on the other variable domain with cross-domain belief propagation by utilizing the existing relationships in multi-relation data. The experimental results on several real world data sets are presented to show the effectiveness of our method.

I. INTRODUCTION

In many practical IR applications, one often faces a lack of sufficient labeled data, since labeling often requires expensive human labor and much time. However, in many cases, large numbers of unlabeled data can be far easier to obtain. For example, in text classification, one may have an easy access to a large database of documents (*e.g.* by crawling the web), but only a small part of them are classified manually.

Consequently, semi-supervised learning methods, which aim to learn from partially labeled data, are proposed. Recently, graph-based semi-supervised learning approaches, where the data points are represented as nodes on the graph and their pairwise relationships as represented as edges on the graph, have also been developed [12][11]. For classification, if we treat the data labels as hidden variables, then they can be modeled as a Markov Random Field and the unknown labels can be obtained via efficient inference algorithms, such as *belief propagation* [5].

So far, most if not all of these methods are focusing on homogeneous data of the same type. *i.e.*, data points from a unique domain. However, in real IR applications, data points from different domains are usually coupled with each other (which are usually called relational data). For example, in text classification, the data set is usually represented by a word-document matrix, which encodes the relationships between the data points from the word

domain and the document domain. Note that information in the word domain gives additional information in the document domain, and vice versa. More general multi-relation data also occur frequently, such as document-word-author relations and document-word-citation relations. Then a natural question arises: if we have some partial label information on the word domain, then can this information be used to help the labeling of the documents? Or if we have some partial label information on both word and document domains, then can we make use of them simultaneously to enhance the discrimination on both domains?

To our understanding, such knowledge transformation in relational data has not been investigated extensively for semi-supervised learning [8]. Multi-relational data mining approaches have been developed for clustering and classification from datasets involving multiple tables (relations) from a relational database [4]. Probabilistic relational learning methods are studied in [6]. However, these multi-relational data mining methods are not designed to deal with semi-supervised learning tasks and they also do not enable knowledge transformation across different domains.

In this paper, we explore this new area, knowledge transformation of multi-relation data for semi-supervised learning tasks. We show that partial knowledge on one data variable domain can help knowledge discovery on the other variable domain with cross-domain belief propagation by utilizing the existing relationships in multi-relation data. The rest of the paper is organized as follows: In section 2 we will briefly review the basic procedure of belief propagation. In section 3 we will introduce how to perform cross-domain belief propagation on the data from two domains. In section 4 we will generalize our algorithm to the data set from multiple domains. The detailed experiments on applying our algorithm to some real world problems are presented in section 5, followed by the conclusions and discussions in section 6.

II. BACKGROUND AND NOTATIONS

In this section we will introduce the background knowledge of belief propagation along with the symbols and notations that will be used to derive our algorithm.

Without much loss of generality, we only consider *pairwise* MRFs, and assume that every node has an observation node attached. Denote the nodes as $\mathbf{X} = [x_1, \dots, x_n]^T$, the

observations as $\mathbf{Y} = [y_1, \dots, y_n]^T$, and the compatibility functions as $\psi_{ij}(\cdot, \cdot)$. According the Hammersley-Clifford theorem [10], the joint distribution can be written as

$$P(x, y) = \frac{1}{Z} \prod_{i,j} \psi_{ij}(x_i, x_j) \prod_i \psi_{ii}(x_i, y_i) \quad (1)$$

where Z is called the *partition function* which ensures that the joint distribution is properly normalized. Usually, the compatibility ψ_{ij} is parameterized by the edge weight w_{ij} , which measures the similarity between x_i and x_j . If two nodes are similar, then their states should also be close to achieve a high compatibility in the network.

Belief Propagation (BP) [10] utilizes the conditional independence properties in the network to derive efficient solutions. Corresponding to the MM and MAP inferences, there are two types of BP [13]. One is *belief update* (BU) *a.k.a.* the *sum-product* algorithm for MM inferences, and another is *belief revision* (BR) *aka* the *max-product* algorithm for MAP inferences. The BP algorithm can be summarized as: 1) nodes deliver their distribution information (*messages*) to others through (and affected by) edges; 2) the distribution (*belief*) at a node is formed by combining messages it received. A detailed introduction on BP can be found in [13]. Here we show that the BU rules are

$$m_{ij}(x_j) \leftarrow \alpha \sum_{x_i} \psi_{ii}(x_i, y_i) \psi_{ij}(x_i, x_j) m_{ii}(x_i) \prod_{x_k \in N(x_i) \setminus x_j} m_{ki}(x_i) \quad (2)$$

$$b_i(x_i) \leftarrow \beta m_{ii}(x_i) \prod_{x_k \in N(x_i)} m_{ki}(x_i) \quad (3)$$

where m_{ij} is the message from x_i to x_j , m_{ii} is the message from y_i to x_i , and b_i is the belief at x_i . α and β are normalization constants and $N(x_i) \setminus x_j$ means all the neighboring nodes of x_i except x_j . The BR rules can be obtained by replacing \sum_{x_i} with \max_{x_i} in (2).

The computational load of belief propagation is concentrated on calculating the messages which has the complexity of $O(ek^2T)$, where e is the number of edges, k is the number of possible states ($k = 1$ for Gaussian MRFs) and T is the number of iterations which usually equals to the graph's diameter. It can be seen that belief propagation could be rather slow on densely connected graphs or graphs with large diameters. To decrease this cost, we could cut down the scale of the graph to reduce e and T , and further reduce T by providing a good start point for belief propagation.

III. CROSS-DOMAIN BELIEF PROPAGATION ON RELATIONAL DATA SETS

In this section, we will first review the basic concepts in bipartite graph and then introduce how to generalize belief propagation on bipartite graph.

A. Bipartite Graph and Its Graphical Model

The traditional data mining and machine learning algorithms usually deal with homogeneous data, *i.e.*, the data

items are all from the same source. For example, we can apply the Gaussian random field method [15] to classify a set of documents, however, the documents usually have close relationships with the words they contain. If we can incorporate the knowledge on the word domain to the process of document classification, the final classification performance could be greatly improved.

Based on the above considerations, some researchers have proposed the idea of "co-clustering" [1], [2], which aims at clustering the interrelated data points from different domains simultaneously. Taking the *Spectral Bipartite Graph Partitioning* (SBGP) method [2] as an example, SBGP first model the data set as a bipartite graph (it assumes the data set comes from two domains) as in Figure ??(a), where the squares represent the data points from one domain, and the triangles represents the data points from the other domain. An edge connecting two data points implies that there is a relationship between the two points (*i.e.* a document contains a word). The goal of SBGP is to find an optimal cut on such bipartite graph, which automatically partitions the data points from different domains simultaneously.

However, to the best of our knowledge, there are rarely any research works on how to perform semi-supervised classification simultaneously on different domains. In this section, we will propose a probabilistic framework based on belief propagation to carry out such procedure. Our approach is based on a basic assumption that if there exists some relationships between different domains, then these domains must have something in common.

Therefore, the remaining problem is how to properly define such an Markov random field on which we can perform the label inference. In the following subsection we will address this issue in detail.

B. Cross-Domain Belief Propagation on Bipartite Graphs

To begin with, we first derive our cross-domain belief propagation algorithm on the data set from two domains in this section. In the next section we will introduce how to generalize our algorithm to the data set from multiple domains. We first introduce some notations.

Let $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ be the whole data set, where $\mathcal{X}_1 = \{\mathbf{x}_{11}, \mathbf{x}_{12}, \dots, \mathbf{x}_{1n_1}\}$ is the set of data points from domain 1, $\mathcal{X}_2 = \{\mathbf{x}_{21}, \mathbf{x}_{22}, \dots, \mathbf{x}_{2n_2}\}$ is the data points from domain 2. Denoting $\mathbf{f}_1 = [f_{11}, f_{12}, \dots, f_{1n_1}]^T$ as the label configuration vector of \mathcal{X}_1 , and $\mathbf{f}_2 = [f_{21}, f_{22}, \dots, f_{2n_2}]^T$ as the label configuration vector of \mathcal{X}_2 . For notational convenience, we first consider the two class classification problem here, therefore $f_{ij} \in \{1, -1\}$. Since the problem we considered is semi-supervised, we use $\mathbf{y}_1 \in \mathcal{R}^{n_1 \times 1}$ to represent the initial label vector of \mathcal{X}_1 , where

$$f_{1i} = \begin{cases} 1, & \text{if } \mathbf{x}_{1i} \text{ belongs to class 1} \\ -1, & \text{if } \mathbf{x}_{1i} \text{ belongs to class 2} \\ 0, & \text{if } \mathbf{x}_{1i} \text{ is unlabeled} \end{cases} \quad (4)$$

Similarly, we can define the initial label vector for \mathcal{X}_2 as $\mathbf{y}_2 \in \mathcal{R}^{n_2 \times 1}$. Then we can define the joint probability distribution over the graphical model of the bipartite graph as

$$P(\mathbf{f}_1, \mathbf{f}_2 | \mathbf{y}_1, \mathbf{y}_2) = e^{-G} / Z \quad (5)$$

where Z is the normalization constant, and the energy function G is defined as

$$\begin{aligned} G &= \alpha \sum_{i,j} [f_{1i}, f_{2j}] C_{ij}^{12} [f_{1i}, f_{2j}]^T \\ &+ \sum_i [f_{1i}, y_{1i}] C_{ii}^1 [f_{1i}, y_{1i}]^T \\ &+ \sum_j [f_{2j}, y_{2j}] C_{jj}^2 [f_{2j}, y_{2j}]^T \end{aligned} \quad (6)$$

where C_{ij}^{12} is the 2×2 potential matrix between f_{1i} and f_{2j} ¹, C_{ii}^1 is the potential matrix between f_{1i} and its corresponding observation note y_{1i} , C_{jj}^2 is defined similarly. $\alpha > 0$ is a tradeoff parameter.

Inspired by the derivations in [2], we define the concrete form of G as

$$G = \alpha \sum_{i,j} (f_{1i} - f_{2j})^2 w_{ij}^{12} + \sum_{i: \mathbf{x}_{1i} \in \mathcal{L}_1} (f_{1i} - y_{1i})^2 + \sum_{j: \mathbf{x}_{2j} \in \mathcal{L}_2} (f_{2j} - y_{2j})^2 \quad (7)$$

where w_{ij}^{12} denotes the similarity (or the strength of the relationship) between \mathbf{x}_{1i} and \mathbf{x}_{2j} , and we use \mathcal{L}_1 to denote the labeled subset in \mathcal{X}_1 , and \mathcal{L}_2 to denote the labeled subset in \mathcal{X}_2 . Now we define an $n_1 \times n_2$ relational matrix \mathbf{W}^{12} as

$$\mathbf{W}^{12} = \begin{bmatrix} w_{11}^{12} & w_{12}^{12} & \cdots & w_{1n_2}^{12} \\ w_{21}^{12} & w_{22}^{12} & \cdots & w_{2n_2}^{12} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n_1}^{12} & w_{n_1 2}^{12} & \cdots & w_{n_1 n_2}^{12} \end{bmatrix} \quad (8)$$

where the meaning of w_{ij}^{12} is the same as in Eq.(7). If we further define two diagonal matrices

$$\begin{aligned} \mathbf{D}^1 &= \text{diag} \left(\sum_i w_{1i}^{12}, \sum_i w_{2i}^{12}, \dots, \sum_i w_{n_1 i}^{12} \right) \in \mathbb{R}^{n_1 \times n_1} \\ \mathbf{D}^2 &= \text{diag} \left(\sum_i w_{i1}^{12}, \sum_i w_{i2}^{12}, \dots, \sum_i w_{in_2}^{12} \right) \in \mathbb{R}^{n_2 \times n_2} \end{aligned}$$

Then we can rewrite Eq.(7) in its matrix form as

$$G = \alpha \mathbf{f}^T (\mathbf{D} - \mathbf{W}) \mathbf{f} + \|\mathbf{J}\mathbf{f} - \mathbf{y}\|^2 \quad (9)$$

where

$$\begin{aligned} \mathbf{f} &= \begin{bmatrix} \mathbf{f}_1^T \\ \mathbf{f}_2^T \end{bmatrix} \in \mathbb{R}^{(n_1+n_2) \times 1}, \quad \mathbf{y} = \begin{bmatrix} \mathbf{y}_1^T \\ \mathbf{y}_2^T \end{bmatrix} \in \mathbb{R}^{(n_1+n_2) \times 1} \\ \mathbf{W} &= \begin{bmatrix} \mathbf{0} & \mathbf{W}^{12} \\ (\mathbf{W}^{12})^T & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{(n_1+n_2) \times (n_1+n_2)} \\ \mathbf{D} &= \begin{bmatrix} \mathbf{D}^1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}^2 \end{bmatrix} \in \mathbb{R}^{(n_1+n_2) \times (n_1+n_2)} \end{aligned}$$

¹Note that what we considered is a bipartite graph, therefore the data in \mathcal{X}_1 can only have relationships with the data in \mathcal{X}_2 , so do their associated label nodes.

$\mathbf{J} \in \mathbb{R}^{(n_1+n_2) \times (n_1+n_2)}$ is a diagonal matrix with $J_{ii} = 1$ if the corresponding data point is labeled (no matter it is in \mathcal{L}_1 or \mathcal{L}_2).

Now the remaining problem is to derive a MAP configuration of the data labels, *i.e.*,

$$\mathbf{f}^* = \text{argmax}_{\mathbf{f}} P(\mathbf{f} | \mathbf{y}) \quad (10)$$

where $P(\mathbf{f} | \mathbf{y})$ has exactly the same form as in Eq.(5). As introduced in section II, such MAP estimation problem can be solved via belief revision by the following rules.

$$m_{ij}(f_j) \leftarrow \alpha \max_{f_i} \psi_{ii}(f_i, y_i) \psi_{ij}(f_i, f_j) m_{ii}(f_i) \prod_{f_k \in N(f_i) \setminus f_j} m_{ki}(f_i) \quad (11)$$

$$b_i(f_i) \leftarrow \beta m_{ii}(f_i) \prod_{f_k \in N(f_i)} m_{ki}(f_i) \quad (12)$$

where we use f_i to denote the i -th element in the concatenated label vector \mathbf{f} , and y_i is the i -th element of \mathbf{y} , W_{ij} is the (i, j) -th element of \mathbf{W} . $N(f_i)$ contains the set of nodes which are connected to f_i . The potential functions

$$\psi(f_i, y_i) = (f_i - y_i)^2 \quad (13)$$

$$\psi(f_i, f_j) = W_{ij} (f_i - f_j)^2 \quad (14)$$

For multi-class classification problem (assume there are totally C classes), we use an $n_1 \times C$ classification matrix \mathbf{F}^1 to represent the data labels in \mathcal{X}_1 , such that

$$F_{ij}^1 = \begin{cases} 1, & \text{if } \mathbf{x}_{1i} \text{ belongs to class } j \\ 0, & \text{otherwise} \end{cases}$$

Similarly we can define an $n_2 \times C$ classification matrix \mathbf{F}^2 for \mathcal{X}_2 . We can also similarly define an $n_1 \times C$ initial classification matrix \mathbf{Y}^1 for \mathcal{X}_1 and an $n_2 \times C$ initial classification matrix \mathbf{Y}^2 for \mathcal{X}_2 . Then using the concatenated matrices

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}^1 \\ \mathbf{F}^2 \end{bmatrix} \in \mathbb{R}^{(n_1+n_2) \times C} \quad \mathbf{Y} = \begin{bmatrix} \mathbf{Y}^1 \\ \mathbf{Y}^2 \end{bmatrix} \in \mathbb{R}^{(n_1+n_2) \times C}$$

we can define the energy function for multi-class problem as

$$G = \alpha \text{tr}(\mathbf{F}^T (\mathbf{D} - \mathbf{W}) \mathbf{F}) + \|\mathbf{J}\mathbf{F} - \mathbf{Y}\|_F^2 \quad (15)$$

where $\text{tr}(\cdot)$ denotes the matrix trace, and $\|\cdot\|_F$ is the Frobenius norm of a matrix. Then similar to Eq.(10), the optimal label configurations can be achieved by

$$\mathbf{F}^* = \text{argmax}_{\mathbf{F}} P(\mathbf{F} | \mathbf{Y}) \quad (16)$$

which can be solved via belief revision by the following rules.

$$m_{ij}(\mathbf{f}_j) \leftarrow \alpha \max_{\mathbf{f}_i} \psi_{ii}(\mathbf{f}_i, \mathbf{y}_i) \psi_{ij}(\mathbf{f}_i, \mathbf{f}_j) m_{ii}(\mathbf{f}_i) \prod_{\mathbf{f}_k \in N(\mathbf{f}_i) \setminus \mathbf{f}_j} m_{ki}(\mathbf{f}_i) \quad (17)$$

$$b_i(\mathbf{f}_i) \leftarrow \beta m_{ii}(\mathbf{f}_i) \prod_{\mathbf{f}_k \in N(\mathbf{f}_i)} m_{ki}(\mathbf{f}_i) \quad (18)$$

Note that the only difference between the above two equations and Eq.(11), Eq.(12) is that we replace all the label

scalars with label indicator vectors, since in the multi-class case, we need to use a $C \times 1$ vector to denote one data label.

One issue that is worthy of being mentioned here is that by observing Eq.(11) and Eq.(17), we may find that the messages only pass through the edges that connect the data from different domains (since we considered here is a bipartite graph, then the edges only exist between the data points from different domains). In this way, the knowledge in one domain can be efficiently transformed to the knowledge on the other domain, and that is also the reason why we call our method *cross-domain belief propagation* (CDBP).

C. Cross-Domain Belief Propagation on Multi-Relational Data

In the last subsection we have derived an efficient cross-domain belief propagation method to solve the semi-supervised classification problem on bipartite graphs. A natural question would be how to generalize such procedure to the data set from multiple domains. Usually this kind of multi-relational data can be represented by a *k-partite graph* [9]. Similar to the bipartite case, we can define a joint probability distribution over all the data labels as

$$P(\mathbf{f}_1, \dots, \mathbf{f}_k | \mathbf{y}_1, \dots, \mathbf{y}_k) = e^{-G}/Z$$

where \mathbf{f}_i denotes the labels of the data from domain i , and \mathbf{y}_i denotes the initial label information of the data from domain i . The energy G can be defined as

$$G = \alpha \sum_{p,q} \sum_{i,j} (f_{pi} - f_{qj})^2 w_{ij}^{pq} + \sum_p \sum_i (f_{pi} - y_{pi})^2 \quad (19)$$

Similar as in Eq.(9), we can define the matrices

$$\mathbf{f} = \begin{bmatrix} \mathbf{f}_1^T \\ \mathbf{f}_2^T \\ \vdots \\ \mathbf{f}_k^T \end{bmatrix} \in \mathbb{R}^{(\sum_{i=1}^k n_i) \times 1} \quad \mathbf{y} = \begin{bmatrix} \mathbf{y}_1^T \\ \mathbf{y}_2^T \\ \vdots \\ \mathbf{y}_k^T \end{bmatrix} \in \mathbb{R}^{(\sum_{i=1}^k n_i) \times 1}$$

$$\mathbf{W} = \begin{bmatrix} \mathbf{0} & \mathbf{W}^{12} & \dots & \mathbf{W}^{1k} \\ (\mathbf{W}^{12})^T & \mathbf{0} & \dots & \mathbf{W}^{2k} \\ \vdots & \vdots & \ddots & \vdots \\ (\mathbf{W}^{1k})^T & (\mathbf{W}^{2k})^T & \dots & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{(\sum_{i=1}^k n_i) \times (\sum_{i=1}^k n_i)}$$

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}^1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{D}^2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{D}^k \end{bmatrix} \in \mathbb{R}^{(\sum_{i=1}^k n_i) \times (\sum_{i=1}^k n_i)}$$

where $\mathbf{W}^{pq}(i, j) = w_{ij}^{pq}$, and

$$\mathbf{D}^i = \text{diag}(\sum_p \sum_i w_{1i}^{1p}, \sum_p \sum_i w_{2i}^{1p}, \dots, \sum_p \sum_i w_{n_1 i}^{1p}) \in \mathbb{R}^{n_1 \times n_1}$$

is a diagonal matrix. We can also similarly define a diagonal matrix $\mathbf{J} \in \mathbb{R}^{(\sum_{i=1}^k n_i) \times (\sum_{i=1}^k n_i)}$ with $J_{ii} = 1$ if the

corresponding data point is labeled (no matter it is in which domain), then we can reformulate Eq.(19) as

$$G = \alpha \mathbf{f}^T (\mathbf{D} - \mathbf{W}) \mathbf{f} + \|\mathbf{J} \mathbf{f} - \mathbf{y}\|^2$$

To get an optimal \mathbf{f} , we can use Eq.(11) and Eq.(12) to iteratively update it. For multi-class problems, we can apply Eq.(17) and Eq.(18) as the updating rules.

IV. EXPERIMENTS

In this section we will present a set of experiments to show the effectiveness of the proposed method.

A. Data Sets

We use the following data sets in our experiments.

- **DBLP Data set:** This data set is obtained from DBLP Computer Science Bibliography². We extract the paper titles published by 552 relatively productive researchers from 9 categories. For easy comparison purpose, we only consider the publications over the last 20 years (from 1988 to 2007, inclusive). Using the ACM Keywords Taxonomy³, we get the term category information and use it as the prior knowledge in the word space.
- **CSTR Data set:** This data set 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.
- **Citeseer Data set:** A real-world data set for experimentation was generated by sampling documents from CiteSeer using combined document meta-data from CiteSeer and another two sources (the ACM Guide, <http://portal.acm.org/guide.cfm>, and the DBLP, <http://www.informatik.uni-trier.de/ley/db/>) for enhanced data accuracy and coverage. The sampled data set includes 1000 documents, 2500 words and 681 authors.
- **BBS Data set:** This is a data set sampled from the Bulletin Board Systems (BBS) data in [7]. The data set includes 1309 users, 1200 topics and 20 boards.

Compared Methods

To demonstrate the superiority of our method, we also conducted a set of competitive methods that are closely related to our method including:

- *Harmonic Gaussian Random Field* (Harmonic) [15]. This method can only tackle homogeneous data, *i.e.*, all the data points should come from the same domain.
- *Learning with Local and Global Consistency* (Consistency) [14]. The difference between Consistency and Harmonic is that Consistency adopts normalized graph Laplacian as the smoothness matrix while Harmonic

²The dblp.xml file is available for download at <http://www.informatik.uni-trier.de/~ley/db/>.

³Available on the page of <http://www.computer.org/portal/pages/ieeecs/publications/author/ACMtaxonomy.html>.

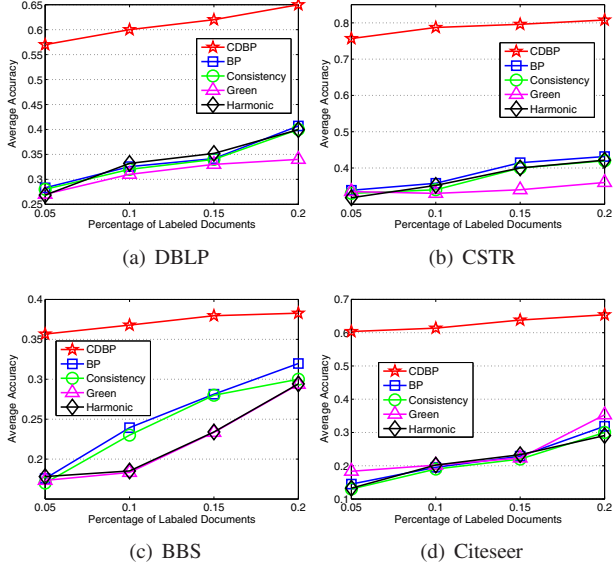


Figure 1. Average classification accuracy comparisons on the four different data sets.

- uses standard graph Laplacian. Consistency can also only tackle homogeneous data.
- *Learning with Green's Function* (Green) [3]. This method applies the Green's function to propagate the data labels through the data graph, rather than graph Laplacian. This is also a method to analyze the data from a unique domain.
 - *Standard Belief Propagation* (BP). All the implementation details of BP is the same as the *cross-domain belief propagation* (CDBP) method introduced in this paper except for that the MRF is constructed on the data set from a unique domain.

All the above methods can only analyze homogeneous data and we can hardly find any semi-supervised methods which can cope with multi-relational data. We use classification accuracy and Normalized Mutual Information to measure the classification performance. Our work is the first step towards semi-supervised learning on multi-relational data. In the following experimental results we will see that by incorporate the transformed knowledge from different interrelated domains, the classification performances on one domain can be improved a lot.

C. Experimental Results on Bipartite Graphs

In this section, we will present the experimental results on applying our method to the classification problem on bipartite graphs. Note that on the Citeseer data set we only use the document-words information, and for the BBS data set we only use the user-topic information.

In our experiments, we randomly label a small portion of documents (or users) (from 5% to 20%), and then apply BP, Harmonic, Consistency and Green methods to classify them.

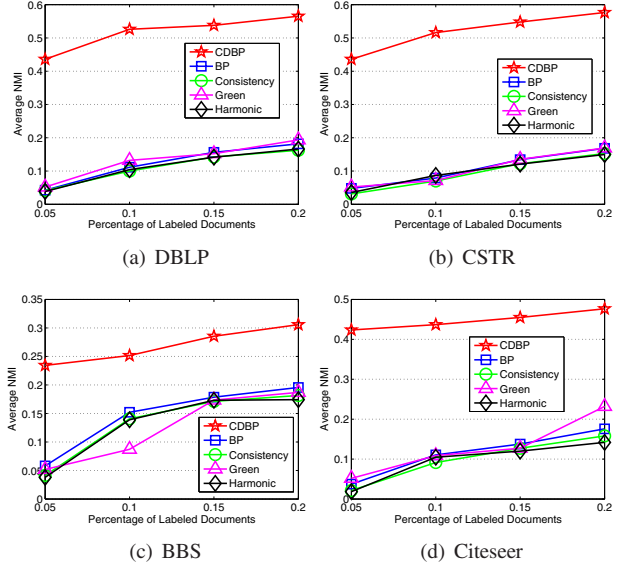


Figure 2. Average NMI comparisons on the four different data sets.

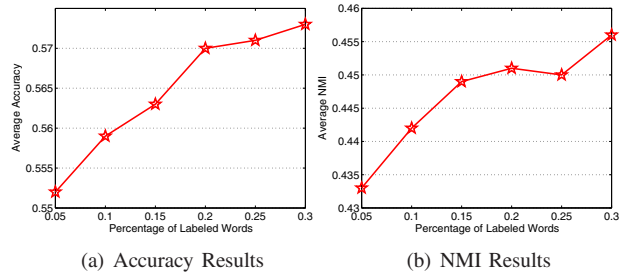


Figure 3. Average accuracy and NMI results on DBLP data set when the size of the labeled words varies.

For each portion value, the experiment is repeated 50 times and the average classification accuracy and NMI is recorded. For our CDBP method, we label 5% of the data points in the word (topic) domain. The results are summarized in Figure 1 and Figure 2. From these figures we observe that by incorporating the knowledge on the word domain, the classification performance on the document domain can be significantly improved.

In another set of experiments, we investigate how the classification result of CDBP varies when the number of labeled words increased. The results on DBLP and CSTR data sets are summarized in Figure 3 and Figure 4, where all the values on the curves are also averaged over 50 independent trials. From the figures we can observe that when the labeled words become more, the classification performance on the document domain would also increase.

D. Experiments on Tripartite Graphs

We also carry out a set of experiments on the three-way relational data, which can be formulated as tripartite graphs. For the Citeseer data set, the results of the algorithms are

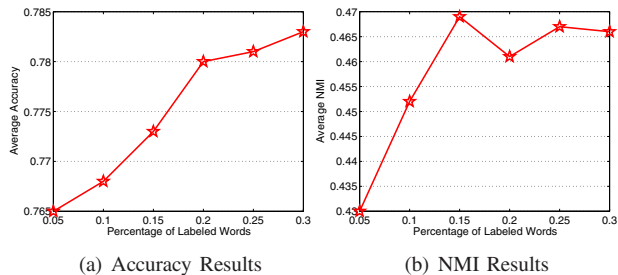


Figure 4. Average accuracy and NMI results on CSTR data set when the size of the labeled words varies.

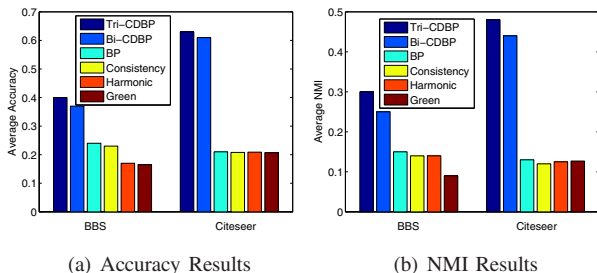


Figure 5. Average accuracy and NMI comparisons on tripartite graphs.

evaluated on the task of classifying the documents, where 10% labeled documents are used, and for the tri-partite CDBP method, we use 5% labeled words and 5% labeled authors. For the BBS data set, the results of the algorithms are evaluated on the task of classifying the users, where 10% labeled users are used, and for the tri-partite CDBP method, we use 5% labeled words and 10% labeled boards. For comparison, we also present the experimental results of CDBP using only words-documents (or users-topics) domain knowledge (denoted as bi-CDBP). The results are summarized in Figure 5, from which observe that the more the domain knowledge we used, the better the classification results would be.

V. CONCLUSIONS

Traditionally belief propagation can only be performed on the data set from one unique domain, however, most of the real world data sets come from different domains and usually they are interrelated with each other. In this paper, we derive a novel algorithm that can perform belief propagation cross the data from different domains, and our experiments show that the knowledge in one domain may be very helpful to the classification task in another domain. In the future we will concentrate on the theoretical analysis of our algorithm on how can such improvements be made.

ACKNOWLEDGEMENT

The work is partially supported by NSF grants NSF grants IIS-0546280 and CCF-0939179.

REFERENCES

- [1] S. Chen, F. Wang, and C. Zhang. Simultaneous heterogeneous data clustering based on higher order relationships. In *Workshops Proceedings of the 7th IEEE International Conference on Data Mining*, pages 387–392, 2007.
- [2] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *SIGKDD*, pages 269–274, 2001.
- [3] C. Ding, H. D. Simon, R. Jin, and T. Li. A learning framework using green’s function and kernel regularization with application for recommender system. In *SIGKDD*, pages 260–269, 2007.
- [4] S. Džeroski. Multi-relational data mining: an introduction. *SIGKDD Explor. Newsl.*, 5(1):1–16, 2003.
- [5] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient belief propagation for early vision. *International Journal of Computer Vision*, 70:41–54, 2006.
- [6] L. Getoor, N. Friedman, D. Koller, and B. Taskar. Learning probabilistic models of relational structure. In *ICML*, 2001.
- [7] Z. Kou and C. Zhang. Reply networks on a bulletin board system. *Physical Review E*, 67(3):036117.1–036117.6, 2003.
- [8] T. Li, C. Ding, Y. Zhang, and B. Shao. Knowledge transformation from word space to document space. In *SIGIR*, pages 187–194, 2008.
- [9] B. Long, X. Wu, Z. Zhang, and P. S. Yu. Unsupervised learning on k-partite graphs. In *SIGKDD*, pages 317–326, 2006.
- [10] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.
- [11] F. Wang, J. Wang, C. Zhang, and H. Shen. Semi-supervised classification using linear neighborhood propagation. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 160–167, 2006.
- [12] F. Wang and C. Zhang. Label propagation through linear neighborhoods. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 985–992, 2006.
- [13] Y. Weiss. Correctness of local probability propagation in graphical models with loops. *Neural Computation*, 12(1):1–41, 2000.
- [14] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. 2003.
- [15] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *The 20th International Conference on Machine Learning*, pages 912–919.