Information Marginalization on Subgraphs

Jiayuan Huang[1,2], Tingshao Zhu[2], Russell Greiner[2],
Dengyong Zhou[3], and Dale Schuurmans[2]

[1] University of Waterloo, Waterloo, Canada
[2] University of Alberta, Edmonton, Canada
[3] NEC Laboratories America, Inc.

Abstract. Real-world data often involves objects that exhibit multiple relationships; for example, 'papers' and 'authors' exhibit both paper-author interactions and paper-paper citation relationships. A typical learning problem requires one to make inferences about a subclass of objects (e.g. 'papers'), while using the remaining objects and relations to provide relevant information. We present a simple, unified mechanism for incorporating information from multiple object types and relations when learning on a targeted subset. In this scheme, all sources of relevant information are marginalized onto the target subclass via random walks. We show that marginalized random walks can be used as a general technique for combining multiple sources of information in relational data. With this approach, we formulate new algorithms for transduction and ranking in relational data, and quantify the performance of new schemes on real world data—achieving good results in many problems.

1 Introduction

Currently, most text classification and clustering algorithms base their inference on the co-occurrence statistics of terms appearing in documents by representing document-term relations via a *bipartite graph*. Many algorithms have been developed for clustering in bipartite graphs, i.e., [9,2,8,4,3]. The underlying intuition behind these approaches is that the similarities among one type of object can be used by the other type of object for clustering.

One obvious limitation of existing co-clustering methods is that they can only deal with two types of data objects, whereas most data sets contain more than two types of objects. For example, in a paper classification task, beyond the bipartite interaction between papers and authors, it is also useful to consider other sources of relevant information, such as the conferences where the papers were published. Such additional paper-conference information could help enhance the classification performance. In this case, one could construct a *tripartite graph* $G = (\langle A, B, C \rangle, E)$, where the vertex sets correspond to authors, papers, and conferences respectively, and E is the set of edges, as shown in Figure 1–left. One could consider addressing the problem of higher-order-partite graphs in a trivial manner by applying co-clustering on each pair of object types; that is, apply a co-clustering method on A, B, and then on B, C individually. However

J. Fürnkranz, T. Scheffer, and M. Spiliopoulou (Eds.): PKDD 2006, LNAI 4213, pp. 199–210, 2006.

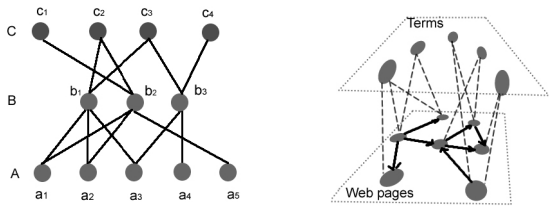

Fig. 1. Left: A tripartite graph. **Right**: A graph of Web pages and terms.

it is hard to ensure the solutions are consistent at the intersection on B. [1] and [5] proposed methods for solving clustering with interactive relationships among multiple object types using ideas from information theory and spectral graph clustering, but they needed to employ sophisticated and computationally expensive methods like semidefinite programming to keep the partitions consistent.

Beyond tripartite clustering, more complex scenarios arise when one considers relationships among data objects of the same type. Previous work on clustering with bipartite and *k-partite* graphs has, for the most part, not taken the relationships *between* objects of the same type into account. Obviously, such information is simply ignored if we present the data as a k-partite graph.

Moving beyond documents and terms, if one considers clustering Web pages, it is clear that the bipartite graph information between Web pages and terms ignores significant relevant information encoded in the hyperlink structure [7,6,10]. When clustering Web pages, it seems clear that both hyperlink structure and term co-occurrence are relevant sources of useful information that one would like to take account of in a unified way. Ideally, one would just model the relationships between Web pages and terms as vertices in a graph like the one shown in Figure 1–right. To the best of our knowledge, clustering in data sets with multiple object types, and multiple relationships between objects of various types has not been well studied in the graph partitioning literature.

In this paper, we propose a simple, unified mechanism for learning in complex scenarios, like the ones shown above, in a graph based approach. We model all data objects as vertices in a graph; e.g., a k-partite graph or a mixed graph as shown in Figure 1–right. The graph based representation allows a simple mechanism for propagating useful information globally throughout a large database of objects: based on the graph, a natural random walk model can be defined that communicates information in a Markov chain. To summarize information from multiple object types and relations when making inferences about one object type, we marginalize the transition probability of the random walk onto the target subset, based on the transition probability of the *induced subgraph* and the transition probability between the subset and its *complement*. In this way, we obtain a valid, new random walk model on the induced subgraph that summarizes all external sources of relevant information. Two objects in the target subgraph that share a lot of common external information will be highly linked in the induced random walk, even if they share no direct links in the induced

subgraph. Once a valid random walk model has been defined, one can derive algorithms for transductive classification, clustering and ranking, by performing random walks over a Markov Chain [10]. The idea of marginalization is a simple and elegant way of dealing with many types of complex scenarios uniformly. Interestingly, when dealing with graphs that happen to be bipartite, the clustering method implied by marginalization is equivalent to the spectral co-clustering method proposed in [9,2]. That is, we recover prominent bipartite graph based inference methods as a special case.

Furthermore, the marginalization idea can be extended to solve more general and interesting types of inference problems on graphs than having been commonly studied in graph partitoning. Consider the problem of clustering the set of blog pages on the Web. In a conventional approach, one could use the induced subgraph on blog pages (namely the subgraph of all the blog pages and their hyperlink structure) to classify the blog pages with respect to their common topics. However, the difficulty with this approach is that there is not much information in the hyperlinks between blog pages, as the owners of the blogs typically do not add links to other blogs if they do not know each other. Therefore, the information obtained directly from the subgraph is not enough to identify blogs of common interest. It therefore makes sense to explore the hyperlinks that *connect blog pages to other general web pages*. For example, people who are interested in computer programming might add a link from their blogs to the page "the art of computer programming" created by Donald Knuth. Although the blogs themselves may have only a few direct links, the blogs can still be clustered into identifiable communities by detecting the pages of common interest linked from the blogs. The scheme we propos can fully exploit all sources of relevant information in a graph of heterogeneous objects to achieve better performance on the target subset.

2 Preliminaries

A *bipartite graph* $G = (\langle A, B \rangle, E)$ is a graph that consists of two *disjoint* sets of vertices, A and B, and a set of edges, E, between A and B. (Typically, the two sets represent different objects, e.g. documents and terms.) Each edge (a, b) is associated with a similarity weight $w(a, b)$. One can generalize bipartite graphs to higher order *k-partite graphs*, whose vertices are divided into k disjoint sets.

Given an undirected graph, a natural random walk can be defined by the transition probability $p : V \times V \rightarrow \Re^{\geq 0}$ such that $p(a, b) = w(a, b)/d(a)$ for all $(a, b) \in E$, where $d(a) = \sum_b w(a, b)$. If the edges have directions, then p is defined by $p(u, v) = w(u, v)/d^+(u)$ for all $(u, v) \in E$ and 0 otherwise, where $d^+(u) = \sum_{u \rightarrow v} w(u, v)$. The random walk on a connected graph has unique *stationary distribution* π that satisfies the *balance equation* $\pi p = \pi$.

Given a general graph $G = (V, E)$ (directed or undirected), and a subset $S \subset V$ of the vertices, the *induced subgraph* with respect to S is the subset V of vertices of G together with any edges whose endpoints are both in V.

3 Learning on a Ergodic Markov Chain

Before presenting our approach in detail, we briefly review related techniques for clustering and transductive learning in graphs involved with Markov chain properties of natural random walks [10]. A graph $G = (V, E)$ can be associated with a Markov chain defined via a random walk on the graph. The stationary distribution of this random walk gives a probability distribution over the vertices v in the graph.

Let $\mathcal{H}(V)$ denote the space of partitions of the vertices V, in that each $f \in \mathcal{H}(V)$ maps each $v \in V$ into real values between -1 and 1. We assume that most linked vertices as similar—that is, belong to the same class. This means, in particular, that all vertices from a densely linked subgraph are likely to have the same label. This motivates us to define the functional

$$\Omega(f) := \frac{1}{2} \sum_{[u,v] \in E} \pi(u) p(u,v) \left(\frac{f(u)}{\sqrt{\pi(u)}} - \frac{f(v)}{\sqrt{\pi(v)}} \right)^2$$

that sums the weighted variation of a function on each edge of the directed graph. The labels are smoothed over the entire graph by minimizing the variation.

There is a equivalent way to express $\Omega(f)$. Let Π denote the diagonal matrix with $\Pi(v, v) = \pi(v)$ for all $v \in V$; let P denote the transition probability matrix; and let P^T the transpose of P. Then

$$\Theta = \frac{\Pi^{1/2} P \Pi^{-1/2} + \Pi^{-1/2} P^T \Pi^{1/2}}{2}.$$

Using I for the identity matrix, it can be proved that $\Omega(f) = f^T(I - \Theta)f$. The functional $\Omega(f)$ can also be derived with respect to a normalized cut criterion that generalizes the standard spectral clustering criterion to directed graphs [10].

4 Marginalized Random Walks on a Subgraph

We can model many versions of graph-based inference problems as learning on an induced subgraph. Typical learning tasks in this setting are classification and clustering on a target subset, where one would like to utilize not only the original structure of the subgraph, but also the global structure and the interactions between the subgraph and its complement. To propagate the information needed to perform these tasks, the graph based approach depends upon a random walk model to communicate the relevant information globally throughout the graph. In the case where the inference problem is to be localized on a focused subset of the graph, we need a new random walk model that communicates the sources of relevant information to the subset. With an appropriate *marginalized* random walk model, we can then derive principled techniques for transductive classification, clustering and ranking.

Given a graph $G = (V, E)$ (either directed or undirected), and a subset of vertices $A \subset V$, we are interested in performing a learning task in A, e.g.,

learning a classification of A's vertices. We let A^c denote the complement of A. For example, in the blog example where A is the set of blog pages we want to classify based on topic, A^c is the set of non-blog Web pages that have connections to the blog pages. In the example of a tripartite graph for a citation network including papers, authors and conferences, A is the set of papers and A^c includes all the authors of the papers and the conferences.

Typically, the transition probability P of a natural random walk model on the graph can be written as in Section 2. Here we can equivalently rewrite the transition probability in a blockwise form with respect to A and A^c

$$P = \begin{pmatrix} P_{AA} & P_{AA^c} \\ P_{A^cA} & P_{A^cA^c} \end{pmatrix}$$

where P_{AA^c} denotes the transition probability between vertices in A and A^c, etc.

One could attempt to perform classification in A based only on P_{AA}, by applying the framework reviewed in Section 3. However this ignores the information that connects A and A^c, which could be significant. A extreme case is that when we have no interactive relationships in either A or A^c but only P_{AA^c} and P_{A^cA}; that is, a bipartite graph (when edges between A and A^c are undirected). We will see later in Section 4.1 that co-clustering methods utilize P_{AA^c} and P_{A^cA} in an undirected case. Now our goal is to define a new random walk in A incorporating all relevant information.

Given a vertex u in A, we first assume it has outlinks to a vertex v in A and a vertex v_c in A^c. The random walk has the following two options starting from u: it can follow the outlink to v (and so stay within A), or to v_c (and so leave A). If it stays in A, the random surfer follows the transition probability P_{AA}. If the random surfer jumps out of A to A^c, its walk will follow the transition probability P_{AA^c}. Once it enters A^c, there is a non-zero chance it will take any number of steps in A^c before possibly returning to A. Therefore, we can write the transition probability between u and v in A, if the surfer re-entered A after transiting from A to A^c and back to A as,

$$P_{out} = P_{AA^c}\left(I + \sum_{i=1}^{n\to\infty} P_{A^cA^c}^i\right) P_{A^cA} = P_{AA^c}(I - P_{A^cA^c})^{-1} P_{A^cA}$$

In addition, define $P_{in} = P_{AA}$ if the surfer stays within A. Combining these two transition models yields a new random walk on the *subgraph* A, whose transition probability P_{AA}^* is given by

$$P_{AA}^* = P_{in} + P_{out}$$

To ensure P_{out} and P_{AA}^* are well defined, we assume P is ergodic. We then have the following claims.

Claim. $I - P_{A^cA^c}$ is invertible.

Proof (of claim). Assume $I - P_{A^cA^c}$ is singular. Then $(I - P_{A^cA^c})x = 0$ has a non-trivial solution $x = P_{A^cA^c}x$. Taking norms, we have $|x| = |P_{A^cA^c}x| \leq$

$|P_{A^cA^c}|\,|x| \;<\; |x|$. The last inequality follows because the row sum of $P_{A^cA^c}$ is less than 1. Contradiction.

Claim. P^*_{AA} is a valid transition probability; i.e. the sum of each row equals 1.

Proof. Consider the ways a random surfer can start from a vertex u in A and return to another vertex v in A. In the first step, u has two choices, either follow links in A or jump out of A to A^c. If it stays in A, the transition probability is P_{in}. If it jumps out of A, then the surfer has an infinite number of paths lenghts that stay in A^c, before (possibly) returning to A. Here, P_{out} is the probability of transiting from u to v via A^c and P_{in} is the transition probability from u to v without entering A^c. Thus the sum of these two disjoint transition probabilities is a valid transition probability.

We let P^*_{AA} denote the new transition probability on A by marginalizing the random walk on subset A, taking all sources of information into account. The similarity among vertices in A is measured by a combination of the transition probability within A, P_{in}, and the probability of escaping from A to A^c and then returning to A, P_{out}. Therefore, we define a new Markov Chain over the subset of the graph. We can use the functional (3), to produce graph-based algorithms for transductive classification, clustering and ranking on complex graphs:

$$f^* = \arg\min_f \{\Omega(f) \;+\; \mu\,||f-y||^2\}$$

Here $y = \langle\, y_i \,\rangle$ is the partially labeled vector; where each labeled data is either 1 or -1, and $y_i = 0$ for each unlabeled data point. For ranking, we label the *root* data as 1 and the rest as 0. Also, μ is a tuning parameter; where for clustering tasks we set $\mu = 0$ since we do not have any label information.

4.1 Learning with a Bipartite Graph

In this section, we will show that the original spectral co-clustering on a bipartite graph [9,2] can be equivalently interpreted as defining new random walk models on each subset of the bipartite graph in our scheme.

Given a bipartite graph $G = (\langle\, A, B \,\rangle, E)$, where A and B are disjoint subsets of vertices, the transition probability P over G has the following blockwise form:

$$P \;=\; \begin{pmatrix} 0 & P_{AB} \\ P_{BA} & 0 \end{pmatrix}$$

Thus, as in the previous section, we can define new random walk in A and B as

$$P^A = P_{AB}P_{BA}, \tag{1}$$

$$P^B = P_{BA}P_{AB} \tag{2}$$

Intuitively, such random walks can be also understood as a two step random walk, motivated by the Hub and Authority model. We take vertices in B as

the evidence of existing similarities between nodes in A. The similarities are mutually reinforced via the random walk between them, as follows.

First consider the random walk among vertices in A (B will be isomorphic). If the random surfer is currently at vertex $a_i \in A$, it first takes a backward step along edge (a_i, b) to some vertex $b \in B$. Then if b also has an edge connected to a_j, the surfer will visit a_j along the edge (b, a_j).

The two-step transition probability $p^A(a_i, a_j)$ is determined by the surfer taking one backward step and one forward step. Therefore,

$$p^A(a_i, a_j) = \sum_b p(a_i, b)p(b, a_j) = \sum_b \frac{w(a_i, b)w(b, a_j)}{d(a_i)d(b)} \tag{3}$$

which is exactly the same as the P^A obtained in (1).

The stationary distribution π^A of this random walk is

$$\pi^A(a) \;=\; \frac{d(a)}{\operatorname{vol} G_A} \tag{4}$$

where $\operatorname{vol} G_A = \sum_{a \in A} d(a)$. This means

$$\begin{aligned}\sum_{a_i \in A} \pi^A(a_i)p^A(a_i, a_j) &= \sum_{a_i \in A} \frac{d(a_i)}{\operatorname{vol} G_A} \sum_{b \in B} \frac{w(a_i, b)w(b, a_j)}{d(a_i)d(b)} \\ &= \frac{1}{\operatorname{vol} G_A} \sum_{b \in B} \frac{w(b, a_j)}{d(b)} \sum_{a_i \in A} w(a_i, b) = \frac{d(a_j)}{\operatorname{vol} G_A} = \pi^A(a_j)\end{aligned}$$

Similarly, we can define the two step transition process among nodes in B, yielding the transition probability

$$p^B(b_i, b_j) = \sum_a p(b_i, a)p(a, b_j) = \sum_a \frac{w(b_i, a)w(a, b_j)}{d(b_i)d(a)} \tag{5}$$

which corresponds to (2). Moreover, the stationary distribution π^B is

$$\pi^B(b) = \frac{d(b)}{\operatorname{vol} G_B} \tag{6}$$

To obtain classification or clustering results on both subsets simultaneously, we define a smoothness function f over A from (3) that is measured by

$$S_A(f) = \frac{1}{2} \sum_{a_i, a_j} P^A(a_i, a_j)\pi(a_i) \left(\frac{f(a_i)}{\sqrt{\pi(a_i)}} - \frac{f(a_j)}{\sqrt{\pi(a_j)}} \right)^2$$

Similarly, the smoothness function g over B is defined as

$$S_B(g) = \frac{1}{2} \sum_{b_i, b_j} P^B(b_i, b_j)\pi(b_i) \left(\frac{g(b_i)}{\sqrt{\pi(b_i)}} - \frac{g(b_j)}{\sqrt{\pi(b_j)}} \right)^2$$

We can use (3), (4), (5) and (6) to prove that

$$S_A(f) = \frac{1}{\mathrm{vol}\, G_A} f^T \Delta_A f, S_B(g) = \frac{1}{\mathrm{vol}\, G_B} g^T \Delta_B g$$

where

$$\begin{aligned} \Delta_A = I - D_A^{-1/2} W^T D_B^{-1} W D_A^{-1/2} &= I - MM^T \\ \Delta_B = I - D_B^{-1/2} W D_A^{-1} W^T D_B^{-1/2} &= I - M^T M \end{aligned}$$

where $D_A = We$, $D_B = W^T e$ and $M = D_A^{-1/2} W^T D_B^{-1/2}$ using the all-1 vector e. W is the weight matrix between A and B. The solutions for f and g are the eigenvectors of MM^T and $M^T M$ with second largest eigenvalues.

It is known the solution of spectral co-clustering on A and B is the second largest left and right singular vectors of M [9,2]. It is easy to see that from the singular value decomposition, that the non-zero left singular eigenvalues of M are the square roots of the non-zero eigenvalues of MM^T with the same eigenvector space. The eigenvector space of M's right eigenvectors is the same as the one of $M^T M$. Therefore, the two solutions are exactly the same, but with different motivations.

The advantage of having marginalized random walk models on each subset is that we can treat each set individually while using their mutual relationships. As expected, the solution is exactly the same as when we considered the combinatorial cut problem in bipartite graphs. In spectral co-clustering method, the goal is to define a cut criterion for the weight matrix that minimizes the cut over the unmatched edges and maximizes the matched vertices in the subgraphs. Such cuts naturally partition the bipartite graph into two parts in each set. The solution is not clear though if we want different number of partitions on each subset. While using our scheme, we can obtain k-cluster results using the first k eigenvectors of Δ_A and Δ_B. Moreover, as discussed in Section 4, this method can be easily generalized into more complex graphs, which would have been difficult from graph cut perspective.

5 Experiments

In this section, we demonstrate several problem settings that involve data represented in complex graph structures. We evaluate our information marginalization approach by applying it to two datasets; see Sections 5.1 and 5.2.

The first dataset is from WebKB (`www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data`), which includes pages from four universities: Cornell, Texas, Washington and Wisconsin. After removing isolated pages, the Web pages have been manually classified into seven categories: *student, faculty, staff, department, course, project* and *other*. We take advantage of the link structure and page-word relationships for the following two learning tasks.

(1a) Given the link structure of all the pages and the words used in them, discriminate student (course) pages from non-student (non-course) pages. Here, A corresponds to the web pages, and A^c to the words. See Figure 1.

(1b) Given only the link structure, discriminate student pages (labeled as 1) from course pages (labeled as -1). For this task, A corresponds the pages of students and courses, and A^c to the web pages from other classes.

The second dataset is based on CiteSeer (`citeseer.ist.psu.edu/`)—a well-known scientific digital library that catalogues primarily computer and information science literature. We construct our citation networks based on paper-paper and paper-author relationships from CiteSeer. We extract a set of papers P with authors U. Here, we focus on two kinds of ranking.

(2a) Given some papers (i.e., *seed* papers) in P labeled as relevant to a specific topic T, rank the rest of the papers based on their relevance to T. Here, A is P, A^c is U.

(2b) Given some authors (i.e., *seed* authors) in A identified as relevant since they share similar research interests, rank the remaining authors based on how much they share the research interests with these seed authors. A is U, A^c is P.

To build citation networks, we scout ahead following the paper citation and corresponding authors information from the OAI records (`citeseer.ist.psu.edu/oai.html`). We start a crawl from a set of pre-selected authors (i.e., *root authors*), then collect all their papers and the co-authors of these papers. The co-authors are added to a growing set of authors that is used in the next iteration. We repeat this iteration $n = 3$ times to collect a number of related authors and papers. In our experiment, we choose the root authors from two different areas:

Root authors	# Authors	# Papers
"Berhard Scholkopf" + "John Kleinberg"	7156	4979
"Vladimir Vapnik" + "Jianbo Shi"	3048	2097

Therefore, the citation network contains authors with different research subjects, which is more realistic.

5.1 Results: Web Classification

We compare the performance of two algorithms for Web page classification in transductive setting. It is well-known that transductive classification typically outperforms supervised one because it take advantages of unlabled data in the learning procedure. The first transductive algorithm uses our marginalized random walk P^*, and the second one uses hyperlink structure P_{AA} only. We use canonical 0-1 weights over the directed hyperlinks. We set the tuning parameter $\mu = 2.5$ for both algorithms. We increase the size of the labeled data sample at each iteration. The comparison is based on 0/1 classification error, averaged by 20 iterations.

Figures 2 and 3 show the comparison results for problem (1a), and Figure 4, for problem (1b). It is clear that the methods using information marginalization outperforms the one with only the local hyperlink information from subset. Specifically, this implies that the marginalized random walk is able to convey more global information onto the subset, efficiently improving the performance in classification.s

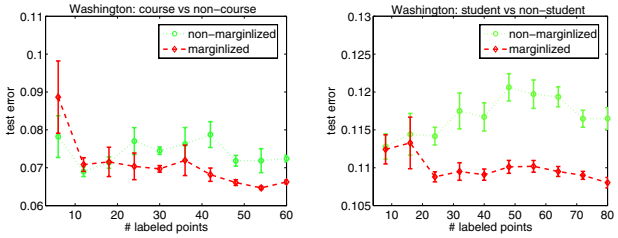

Fig. 2. Classification error on discriminating course pages from non-course pages (left) and student pages from non-student pages (right) from Washington

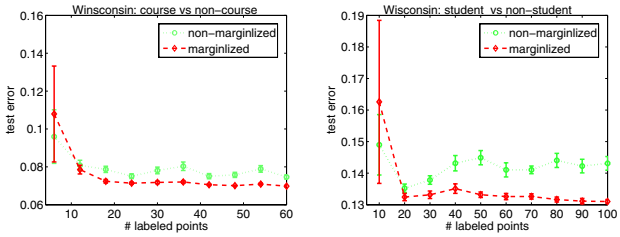

Fig. 3. Classification error on discriminating course pages from non-course pages (left) and student pages from non-student pages (right) from Wisconsin

5.2 Results: Ranking in Citation Networks

For problem (2a), Table 1 shows the top 20 results of paper ranking with respect to the labeled paper "Kernel Principal Component Analysis"; and Table 2 shows the top 10 papers ranked with respect to "Authoritative Sources in a Hyperlinked Environment". We can see that the information maginalization method works better than only using citation links information as the highly ranked papers are closer to the labeled paper in information marginalization scheme. If we only consider citation links, some papers from slightly different domain may be included in the top ranking list because they may have citations with similar papers. With the help of author-paper relationships, the relationship between the labeled paper and other papers become more clear thus lead more accurate ranking results.

For problem (2b),Table 3 lists the ranking results of authors with respect to Vladimir Vapnik in the second citation network. The information from the citation links moves some authors—Chris Burges, Bernhard Scholkopf, Olivier Chapelle and Alex Smola—to higher ranking positions than only using author-paper relationships. The reason is that these authors also have many citation links among their papers that strengthen the similarities with respect to the labeled author.

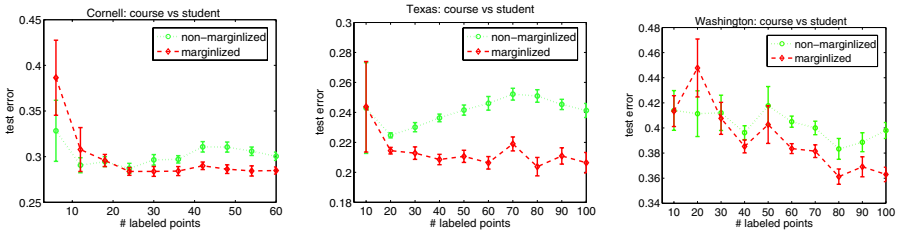

Fig. 4. Classification error on discriminating course pages from student pages

Table 1. Papers Ranked closest to "Kernel Principal Component Analysis"

marginalized random walk	**use only citation links**
title	title
1. Regression Estimation with Support Vector Learning Machines	1. Model Selection for Support Vector Machines
2. Model Selection for Support Vector Machines	2. SV Estimation of a Distribution's Support
3. Support Vector Method for Novelty Detection	3. Support Vector Method for Novelty Detection
4. A Generalized Representer Theorem	4. Optimal Hyperplane Classifier with Adaptive Norm
5. Optimal Hyperplane Classifier with Adaptive Norm	5. Inclusional Theories in Declarative Programming
6. Incorporating Invariances in Support Vector Learning Machines	6. Studies on the Formal Semantics of Pictures
7. Latent Semantic Kernels	7. A Noise-Tolerant Hybrid Model of a Global and a Local Learning Module
8. Sparse Kernel Feature Analysis	8. Latent Semantic Kernels
9. Extracting Support Data for a Given Task	9. Incorporating Invariances in Support Vector Learning Machines
10.Support-Vector Networks	10.A Generalized Representer Theorem
11.Kernel Methods: A Survey of Current Techniques	11.Equivalent Conditions for the Solvability of Nonstandard LQ-Problems with Applications to Partial Differential Equations with Continuous Input-Output Solution Map
12.A Training Algorithm for Optimal Margin Classifiers	12.Hyperbolic Conservation Laws with a Moving Source
13.Improving the Accuracy and Speed of Support Vector Machines	13.Extracting Support Data for a Given Task
14.The Connection between Regularization Operators and Support Vector Kernels	14.Support-Vector Networks
15.Generalization Performance of Regularization Networks and Support Vector Machines	15.On Molecular Approximation Algorithms for NP Optimization Problems
16.Statistical Learning and Kernel Methods	16.Kernel Methods:A Survey of Current Techniques
17.The Kernel Trick for Distances	17.CPU Management for UNIX-based MPEG Video Applications
18.On a Kernel-based Method for Pattern "Recognition," "Regression," "Approximation"	18.Efficient Lossless Compression of Trees and Graphs
19.Advances in Kernel Methods - Support Vector Learning	19.A Precise Semantics For Vague Diagrams
20.Estimating the Support of a High-Dimensional Distribution	20.Redescription, Information And Access

Table 2. Papers Ranked closest to "Authoritative Sources in a Hyperlinked Environment"

marginalized random walk	**use only citation links**
title	title
1. Fast Monte-Carlo Algorithms for finding low-rank approximations	1. volutionary Strategies For Solving Frustrated Problems
2. Evolutionary Strategies For Solving Frustrated Problems	2. Fast Monte-Carlo Algorithms for finding low-rank approximations
3. The Anatomy of a Large-Scale Hypertextual Web Search Engine	3. Reconstruction From The Multi-Component Am-Fm Image
4. Latent Semantic Indexing: A Probabilistic Analysis	4. The Anatomy of a Large-Scale Hypertextual Web Search Engine
5. Challenges in Web Search Engines	5. Latent Semantic Indexing: A Probabilistic Analysis
6. How to Personalize the Web	6. Learning Decision Strategies with Genetic Algorithms
7. Efficient and Effective Metasearch for Text Databases Incorporating Linkages among Documents	7. A Model for Sequence Databases
8. The PageRank Citation Ranking: Bringing Order to the Web	8. Semantically Driven Automatic Hyperlinking
9. New Results for Online Page Replication	9. Applications of a Web Query Language
10Searching the Web: General and Scientific Information Access	10Efficient and Effective Metasearch for Text Databases Incorporating Linkages among Documents

Table 3. Author ranking result in network 2

marginalized	**only author-paper relationships**	**marginalized**	**only author-paper relationships**
name	name	name	name
1.Chris Burges	1.Sayan Mukherjee	11.Mark Stitson	11.Vladimir Vovk
2.Bernhard E.Boser	2.Chris Burges	12.Alex Gammerman	12.Alex Gammerman
3.Isabelle M. Guyon	3.Bernhard E. Boser	13.Vladimir Vovk	13.Mark Stitson
4.Sayan Mukherjee	4.Isabelle M.Guyon	14.Chris Watkins	14.Klaus-Robert Muller
5.Donghui Wu	5.Donghui Wu	15.Partha Niyogi	15.Federico Girosi
6.Bernhard Scholkopf	6.Steven E.Golowich	16.Olivier Chapelle	16.Koh.Sung
7.Heinrich H.Bulthoff	7.Volker Blanz	17.Alex Smola	17.Partha Niyogi
8.Thomas Vetter	8.Bernhard Scholkopf	18.Adnan Aziz	18.Jason Weston
9.Volker Blanz	9.Thomas Vetter	19.Jason Weston	19.Olivier Chapelle
10.Steven Golowich	10.Chris Watkins	20.Koh.Sung	20.Alex Smola

6 Conclusions

We have proposed a unified mechanism for incorporating information from multiple object types and relations when making inferences about a targeted subset. Our technique can be applied to learning problems with data embedded in complex graphs. We quantify the performance of our new schemes on two real world relational data and achieve good results in challenging inference problems. Future work will deeply explore more interesting applications of this method.

Acknowledgments

Work supported by Alberta Ingenuity (AICML), NSERC, and MITACS.

References

1. R. Bekkerman, E. El-Yaniv, and A. McCallum. Multiway distributional clustering via pairwise interactions. In *ICML*, 2005.
2. I. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioing. In *KDD*, 2001.
3. I. S. Dhillon, S. Mallela, and D. S. Modha. Information-theoretic co-clustering. In *KDD*, 2003.
4. Ran El-Yaniv and Oren Souroujon. Iterative double clustering for unsupervised and semi-supervised learning. In *ECML*, 2001.
5. B. Gao, T. Liu, X. Zheng, Q. Cheng, and W. Ma. Consistent bipartite graph co-partitioning for star-structured high-order heterogeneous data co-clustering. In *KDD*, 2005.
6. J. Kleinberg. Authoritative sources in a hyperlinked environment. *JACM*, 46, 1999.
7. L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford, 1998.
8. N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings 37th Allerton Conference*, 1999.
9. H. Zhang, X. He, C. Ding, and M. Gu. Bipartite graph partitioning and data clustering. In *Proceedings of ACM CIKM 2001*, 2001.
10. D. Zhou, J. Huang, and B. Scholkopf. Learning from labeled and unlabeled data on a directed graph. In *ICML*, 2005.