Spectral Clustering for Multi-type Relational Data

Bo Long BLONG1@BINGHAMTON.EDU

Zhongfei (Mark) Zhang ZZHANG@BINGHAMTON.EDU
Computer Science Dept., SUNY Binghamton, Binghamton, NY 13902 USA

Xiaoyun Wu XIAOYUNW@YAHOO-INC.COM
Yahoo! Inc, 701 First Avenue, Sunnyvale, CA 94089 USA

Philip S. Yu PSYU@US.IBM.COM
IBM Watson Research Center, 19 skyline Drive, Hawthorne, NY 10532

Abstract

Clustering on multi-type relational data has attracted more and more attention in recent years due to its high impact on various important applications, such as Web mining, e-commerce and bioinformatics. However, the research on general multi-type relational data clustering is still limited and preliminary. The contribution of the paper is three-fold. First, we propose a general model, the collective factorization on related matrices, for multi-type relational data clustering. The model is applicable to relational data with various structures. Second, under this model, we derive a novel algorithm, the spectral relational clustering, to cluster multi-type interrelated data objects simultaneously. The algorithm iteratively embeds each type of data objects into low dimensional spaces and benefits from the interactions among the hidden structures of different types of data objects. Extensive experiments demonstrate the promise and effectiveness of the proposed algorithm. Third, we show that the existing spectral clustering algorithms can be considered as the special cases of the proposed model and algorithm. This demonstrates the good theoretic generality of the proposed model and algorithm.

1. Introduction

Most clustering approaches in the literature focus on ”flat” data in which each data object is represented as a fixed-length feature vector (R.O.Duda et al., 2000). However, many real-world data sets are much richer in structure, involving objects of multiple types that are related to each other, such as Web pages, search queries and Web users in a Web search system, and papers, key words, authors and conferences in a scientific publication domain. In such scenarios, using traditional methods to cluster each type of objects independently may not work well due to the following reasons.

First, to make use of relation information under the traditional clustering framework, the relation information needs to be transformed into features. In general, this transformation causes information loss and/or very high dimensional and sparse data. For example, if we represent the relations between Web pages and Web users as well as search queries as the features for the Web pages, this leads to a huge number of features with sparse values for each Web page. Second, traditional clustering approaches are unable to tackle with the interactions among the hidden structures of different types of objects, since they cluster data of single type based on static features. Note that the interactions could pass along the relations, i.e., there exists influence propagation in multi-type relational data. Third, in some machine learning applications, users are not only interested in the hidden structure for each type of objects, but also the global structure involving multi-types of objects. For example, in document clustering, except for document clusters and word clusters, the relationship between document clusters and word clusters is also useful information. It is difficult to discover such global structures by clustering each type of objects individually.

Therefore, multi-type relational data has presented a great challenge for traditional clustering approaches. In this study, first, we propose a general model, the collective factorization on related matrices, to discover the hidden structures of multi-types of objects based on both feature information and relation information. By clustering the multi-types of objects simultaneously, the model performs

Appearing in *Proceedings of the 23^{rd} International Conference on Machine Learning*, Pittsburgh, PA, 2006. Copyright 2006 by the author(s)/owner(s).

adaptive dimensionality reduction for each type of data. Through the related factorizations which share factors, the hidden structures of different types of objects could interact under the model. In addition to the cluster structures for each type of data, the model also provides information about the relation between clusters of different types of objects.

Second, under this model, we derive a novel algorithm, the spectral relational clustering, to cluster multi-type interrelated data objects simultaneously. By iteratively embedding each type of data objects into low dimensional spaces, the algorithm benefits from the interactions among the hidden structures of different types of data objects. The algorithm has the simplicity of spectral clustering approaches but at the same time also applicable to relational data with various structures. Theoretic analysis and experimental results demonstrate the promise and effectiveness of the algorithm.

Third, we show that the existing spectral clustering algorithms can be considered as the special cases of the proposed model and algorithm. This provides an unified view to understand the connections among these algorithms.

2. Related Work

Spectral clustering (Ng et al., 2001; Bach & Jordan, 2004) has been well studied in the literature. The spectral clustering methods based on the graph partitioning theory focus on finding the best cuts of a graph that optimize certain predefined criterion functions. The optimization of the criterion functions usually leads to the computation of singular vectors or eigenvectors of certain graph affinity matrices. Many criterion functions, such as the average cut (Chan et al., 1993), the average association (Shi & Malik, 2000), the normalized cut (Shi & Malik, 2000), and the min-max cut (Ding et al., 2001), have been proposed.

Spectral graph partitioning has also been applied to a special case of multi-type relational data, bi-type relational data such as the word-document data (Dhillon, 2001; H.Zha & H.Simon, 2001). These algorithms formulate the data matrix as a bipartite graph and seek to find the optimal normalized cut for the graph. Due to the nature of a bipartite graph, these algorithms have the restriction that the clusters from different types of objects must have one-to-one associations.

Clustering on bi-type relational data is called co-clustering or bi-clustering. Recently, co-clustering has been addressed based on matrix factorization. Both Long et al. (2005) and Li (2005) model the co-clustering as an optimization problem involving a triple matrix factorization. Long et al. (2005) propose an EM-like algorithm based on multiplicative updating rules and Li (2005) proposes a hard clustering algorithm for binary data. Ding et al. (2005) extend the non-negative matrix factorization to symmetric matrices and show that it is equvilent to the Kernel K-means and the Laplacian-based spectral clustering. Several previous efforts related to co-clustering are model based. PLSA (Hofmann, 1999) is a method based on a mixture decomposition derived from a latent class model. A two-sided clustering model is proposed for collaborative filtering by Hofmann and Puzicha (1999). Information-theory based co-clustering has also attracted attention in the literature. El-Yaniv and Souroujon (2001) extend the information bottleneck (IB) framework (Tishby et al., 1999) to repeatedly cluster documents and then words. Dhillon et al. (2003) propose a co-clustering algorithm to maximize the mutual information between the clustered random variables subject to the constraints on the number of row and column clusters. A more generalized co-clustering framework is presented by Banerjee et al. (2004) wherein any Bregman divergence can be used in the objective function.

Comparing with co-clustering, clustering on general relational data, which may consist of more than two types of data objects, has not been well studied in the literature. Several noticeable efforts are discussed as follows. Taskar et al. (2001) extend the the probabilistic relational model to the clustering scenario by introducing latent variables into the model. Gao et al. (2005) formulate star-structured relational data as a star-structured m-partite graph and develop an algorithm based on semi-definite programming to partition the graph. Like bipartite graph partitioning, it has limitations that the clusters from different types of objects must have one-to-one associations and it fails to consider the feature information.

An intuitive idea for clustering multi-type interrelated objects is the mutual reinforcement clustering. The idea works as follows: start with initial cluster structures of the data; derive the new reduced features from the clusters of the related objects for each type of objects; based on the new features, cluster each type of objects with a traditional clustering algorithm; go back to the second step until the algorithm converges. Base on this idea, Zeng et al. (2002) propose a framework for clustering heterogeneous Web objects and Wang et al. (2003) present an approach to improve the cluster quality of interrelated data objects through an iterative reinforcement clustering process. However, there is no sounded objective function and theoretical proof on the effectiveness and correctness (convergence) of the mutual reinforcement clustering.

To summarize, the research on multi-type relational data clustering has attracted substantial attention, especially in the the special cases of relational data. However, there is still limited and preliminary work on the general relational data. This paper attempts to derive a theoretically sounded model and algorithm for general multi-type relational data clustering.

3. Model Formulation

In this section, we propose a general model for clustering multi-type relational data based on factorizing multiple re-

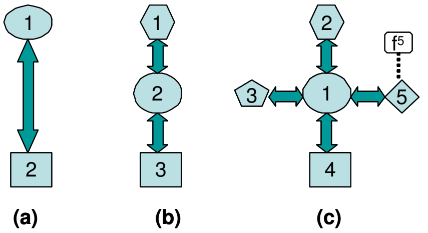

Figure 1. Examples of the structures of multi-type relational data.

lated matrices.

Given m sets of data objects, $\mathcal{X}_1 = \{x_{11}, \ldots, x_{1n_1}\}, \ldots, \mathcal{X}_m = \{x_{m1}, \ldots, x_{mn_m}\}$, which refer to m different types of objects relating to each other, we are interested in simultaneously clustering $\mathcal{X}_1$ into k_1 disjoint clusters, ..., and $\mathcal{X}_m$ into k_m disjoint clusters. We call this task as *collective clustering on multi-type relational data*.

To derive a general model for collective clustering, we first formulate Multi-Type Relational Data (MTRD) as a set of related matrices, in which two matrices are related in the sense that their row indices or column indices refer to the same set of objects. First, if there exist relations between $\mathcal{X}_i$ and $\mathcal{X}_j$ (denoted as $\mathcal{X}_i \sim \mathcal{X}_j$), we represent them as a relation matrix $R^{(ij)} \in \mathbb{R}^{n_i \times n_j}$, where an element $R_{pq}^{(ij)}$ denotes the relation between x_{ip} and x_{jq}. Second, a set of objects $\mathcal{X}_i$ may have its own features, which could be denoted by a feature matrix $F^{(i)} \in \mathbb{R}^{n_i \times f_i}$, where an element $F_{pq}^{(i)}$ denotes the qth feature values for the object x_{ip} and f_i is the number of features for $\mathcal{X}_i$.

Figure 1 shows three examples of the structures of MTRD. Example (a) refers to a basic bi-type of relational data denoted by a relation matrix $R^{(12)}$, such as word-document data. Example (b) represents a tri-type of star-structured data, such as Web pages, Web users and search queries in Web search systems, which are denoted by two relation matrices $R^{(12)}$ and $R^{(23)}$. Example (c) represents the data consisting of shops, customers, suppliers, shareholders and advertisement media, in which customers (type 5) have features. The data are denoted by four relation matrices $R^{(12)}$, $R^{(13)}$, $R^{(14)}$ and $R^{(15)}$, and one feature matrix $F^{(5)}$.

It has been shown that the hidden structure of a data matrix can be explored by its factorization (D.D.Lee & H.S.Seung, 1999; Long et al., 2005). Motivated by this observation, we propose a general model for collective clustering, which is based on factorizing the multiple related matrices. In MTRD, the cluster structure for a type of objects $\mathcal{X}_i$ may be embedded in multiple related matrices; hence it can be exploited in multiple related factorizations. First, if $\mathcal{X}_i \sim \mathcal{X}_j$, then the cluster structures of both $\mathcal{X}_i$ and $\mathcal{X}_j$ are reflected in the triple factorization of their relation matrix $R^{(ij)}$ such that $R^{(ij)} \approx C^{(i)}A^{(ij)}(C^{(j)})^T$ (Long et al., 2005), where $C^{(i)} \in \{0,1\}^{n_i \times k_i}$ is a *cluster indicator matrix* for $\mathcal{X}_i$ such that $\sum_{q=1}^{k_i} C_{pq}^{(i)} = 1$ and $C_{pq}^{(i)} = 1$ denotes that the pth object in $\mathcal{X}_i$ is associated with the qth cluster. Similarly $C^{(j)} \in \{0,1\}^{n_j \times k_j}$. $A^{(ij)} \in \mathbb{R}^{k_i \times k_j}$ is the *cluster association matrix* such that A_{pq}^{ij} denotes the association between cluster p of $\mathcal{X}_i$ and cluster q of $\mathcal{X}_j$. Second, if $\mathcal{X}_i$ has a feature matrix $F^{(i)} \in \mathbb{R}^{n_i \times f_i}$, the cluster structure is reflected in the factorization of $F^{(i)}$ such that $F^{(i)} \approx C^{(i)}B^{(i)}$, where $C^{(i)} \in \{0,1\}^{n_i \times k_i}$ is a cluster indicator matrix, and $B^{(i)} \in \mathbb{R}^{k_i \times f_i}$ is the feature basis matrix which consists of k_i basis (cluster center) vectors in the feature space.

Based on the above discussions, formally we formulate the task of collective clustering on MTRD as the following optimization problem. Considering the most general case, we assume that in MTRD, every pair of $\mathcal{X}_i$ and $\mathcal{X}_j$ is related to each other and every $\mathcal{X}_i$ has a feature matrix $F^{(i)}$.

Definition 3.1. Given m positive numbers $\{k_i\}_{1 \le i \le m}$ and MTRD $\{\mathcal{X}_1, \ldots, \mathcal{X}_m\}$, which is described by a set of relation matrices $\{R^{(ij)} \in \mathbb{R}^{n_i \times n_j}\}_{1 \le i < j \le m}$, a set of feature matrices $\{F^{(i)} \in \mathbb{R}^{n_i \times f_i}\}_{1 \le i \le m}$, as well as a set of weights $w_a^{(ij)}, w_b^{(i)} \in R_+$ for different types of relations and features, the task of the collective clustering on the MTRD is to minimize

$$\begin{aligned} L &= \sum_{1 \le i < j \le m} w_a^{(ij)} ||R^{(ij)} - C^{(i)}A^{(ij)}(C^{(j)})^T||^2 \\ &\quad + \sum_{1 \le i \le m} w_b^{(i)} ||F^{(i)} - C^{(i)}B^{(i)}||^2 \end{aligned} \tag{1}$$

w.r.t. $C^{(i)} \in \{0,1\}^{n_i \times k_i}$, $A^{(ij)} \in \mathbb{R}^{k_i \times k_j}$, and $B^{(i)} \in \mathbb{R}^{k_i \times f_i}$ subject to the constraints: $\sum_{q=1}^{k_i} C_{pq}^{(i)} = 1$, where $1 \le p \le n_i$, $1 \le i < j \le m$, and $|| \cdot ||$ denotes the Frobenius norm for a matrix.

We call the model proposed in Definition 3.1 as the Collective Factorization on Related Matrices (CFRM).

The CFRM model clusters multi-type interrelated data objects simultaneously based on both relation and feature information. The model exploits the interactions between the hidden structures of different types of objects through the related factorizations which share matrix factors, i.e., cluster indicator matrices. Hence, the interactions between hidden structures work in two ways. First, if $\mathcal{X}_i \sim \mathcal{X}_j$, the interactions are reflected as the duality of row clustering and column clustering in $R^{(ij)}$. Second, if two types of objects are indirectly related, the interactions pass along the relation ”chains” by a chain of related factorizations, i.e., the model is capable of dealing with influence propagation. In addition to local cluster structure for each type of objects, the model also provides the global structure information by the cluster association matrices, which represent the relations among the clusters of different types of objects.

4. Algorithm Derivation

In this section, we derive a spectral clustering algorithm for MTRD under the CFRM model. First, without loss of generality, we re-define the cluster indicator matric $C^{(i)}$ as the following vigorous cluster indicator matrix,

$$C_{pq}^{(i)} = \begin{cases} \frac{1}{|\pi_q^{(i)}|^{\frac{1}{2}}} & \text{if } x_{ip} \in \pi_q^{(i)} \\ 0 & \text{otherwise} \end{cases}$$

where $|\pi_q^{(i)}|$ denotes the number of objects in the qth cluster of $\mathcal{X}^{(i)}$. Clearly $C^{(i)}$ still captures the disjoint cluster memberships and $(C^{(i)})^T C^{(i)} = I_{k_i}$ where I_{k_i} denotes $k_i \times k_i$ identity matrix. Hence our task is the minimization:

$$\min_{\substack{\{(C^{(i)})^T C^{(i)} = I_{k_i}\}_{1 \le i \le m} \\ \{A^{(ij)} \in \mathbb{R}^{k_i \times k_j}\}_{1 \le i < j \le m} \\ \{B^{(i)} \in \mathbb{R}^{k_i \times f_i}\}_{1 \le i \le m}}} L \tag{2}$$

where L is the same as in Eq. (1).

Then, we prove the following lemma, which is useful in proving our main theorem.

Lemma 4.1. *If* $\{C^{(i)}\}_{1 \le i \le m}$, $\{A^{(ij)}\}_{1 \le i < j \le m}$, *and* $\{B^{(i)}\}_{1 \le i \le m}$ *are the optimal solution to Eq.* (2), *then*

$$A^{(ij)} = (C^{(i)})^T R^{(ij)} C^{(j)} \tag{3}$$

$$B^{(i)} = (C^{(i)})^T F^{(i)} \tag{4}$$

for $1 \le i \le m$.

Proof. The objective function in Eq. (2) can be expanded as follows.

$$\begin{aligned} L = & \sum_{1 \le i < j \le m} w_a^{(ij)} \text{tr}((R^{(ij)} - C^{(i)} A^{(ij)} (C^{(j)})^T) \\ & (R^{(ij)} - C^{(i)} A^{(ij)} (C^{(j)})^T)^T) + \\ & \sum_{1 \le i \le m} w_b^{(i)} \text{tr}((F^{(i)} - C^{(i)} B^{(i)})(F^{(i)} - C^{(i)} B^{(i)})^T) \\ = & \sum_{1 \le i < j \le m} w_a^{(ij)} (\text{tr}(R^{(ij)} (R^{(ij)})^T) + \\ & \text{tr}(A^{(ij)} (A^{(ij)})^T) - 2\text{tr}(C^{(i)} A^{(ij)} (C^{(i)})^T (R^{(ij)})^T)) \\ & + \sum_{1 \le i \le m} w_b^{(i)} (\text{tr}(F^{(i)} (F^{(i)})^T) + \text{tr}(B^{(i)} (B^{(i)})^T) \\ & -2\text{tr}(C^{(i)} B^{(i)} (F^{(i)})^T)) \end{aligned} \tag{5}$$

where tr denotes the trace of a matrix; the terms tr$(A^{(ij)} (A^{(ij)})^T)$ and tr$(B^{(i)} (B^{(i)})^T)$ result from the communicative property of the trace and $(C^{(i)})^T (C^{(i)}) = I_{k_i}$. Based on Eq. (5), solving $\frac{\partial L}{\partial A^{(ij)}} = 0$ and $\frac{\partial L}{\partial B^{(i)}} = 0$ leads to Eq. (3) and Eq. (4). This completes the proof of the lemma. □

Lemma 4.1 implies that the objective function in Eq. (1) can be simplified to the function of only $C^{(i)}$. This leads to the following theorem, which is the basis of our algorithm.

Theorem 4.2. *The minimization problem in Eq.* (2) *is equivalent to the following maximization problem:*

$$\begin{aligned} \max_{\substack{\{(C^{(i)})^T C^{(i)} \\ = I_{k_i}\}_{1 \le i \le m}}} & \sum_{1 \le i \le m} w_b^{(i)} tr((C^{(i)})^T F^{(i)} (F^{(i)})^T C^{(i)}) + \\ & \sum_{1 \le i < j \le m} w_a^{(ij)} tr((C^{(i)})^T R^{(ij)} C^{(j)} (C^{(j)})^T (R^{(ij)})^T C^{(i)}) \end{aligned} \tag{6}$$

Proof. From Lemma 4.1, we have Eq. (3) and (4). Plugging them into Eq. (5), we obtain

$$\begin{aligned} L = & \sum_{1 \le i \le m} w_b^{(i)} (\text{tr}(F^{(i)} (F^{(i)})^T) - \\ & \text{tr}((C^{(i)})^T F^{(i)} (F^{(i)})^T C^{(i)})) + \\ & \sum_{1 \le i < j \le m} w_a^{(ij)} (\text{tr}(R^{(ij)} (R^{(ij)})^T) - \\ & \text{tr}((C^{(i)})^T R^{(ij)} C^{(j)} (C^{(j)})^T (R^{(ij)})^T C^{(i)})). \end{aligned} \tag{7}$$

Since in Eq. (7), tr$(F^{(i)} (F^{(i)})^T)$ and tr$(R^{(ij)} (R^{(ij)})^T)$ are constants, the minimization of L in Eq. (2) is equivalent to the maximization in Eq. (6). This completes the proof of the theorem. □

We propose an iterative algorithm to determine the optimal (local) solution to the maximization problem in Theorem 4.2, i.e., at each iterative step we maximize the objective function in Eq. (6) w.r.t. only one matrix $C^{(p)}$ and fix other $C^{(j)}$ for $j \neq p$ where $1 \le p, j \le m$. Based on Eq. (6), after a little algebraic manipulation, the task at each iterative step is equivalent to the following maximization,

$$\max_{(C^{(p)})^T C^{(p)} = I_{k_p}} \text{tr}((C^{(p)})^T M^{(p)} C^{(p)}) \tag{8}$$

where

$$\begin{aligned} M^{(p)} = & w_b^{(p)} (F^{(p)} (F^{(p)})^T) + \\ & \sum_{p < j \le m} w_a^{(pj)} (R^{(pj)} C^{(j)} (C^{(j)})^T (R^{(pj)^T})) + \\ & \sum_{1 \le j < p} w_a^{(jp)} ((R^{(jp)})^T C^{(j)} (C^{(j)})^T (R^{(jp)})). \end{aligned} \tag{9}$$

Clearly $M^{(p)}$ is a symmetric matrix. Since $C^{(p)}$ is a vigorous cluster indicator matrix, the maximization problem in Eq. (8) is still NP-hard. However, as in the spectral graph partitioning, if we apply real relaxation to $C^{(p)}$ to let $C^{(p)}$ be an arbitrary orthonormal matrix, it turns out that the maximization in Eq. (8) has a closed-form solution.

Theorem 4.3. *(Ky-Fan thorem) Let* M *be a symmetric matrix with eigenvalues* $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_k$, *and the corresponding eigenvectors* $U = [u_1, \ldots, u_k]$. *Then* $\sum_{i=1}^k \lambda_i = \max_{X^T X = I_k} tr(X^T M X)$. *Moreover, the optimal* X *is given by* $[u_1, \ldots, u_k]Q$ *where* Q *is an arbitrary orthogonal matrix.*

Algorithm 1 Spectral Relational Clustering

Input: Relation matrices $\{R^{(ij)} \in \mathbb{R}^{n_i \times n_j}\}_{1 \leq i < j \leq m}$, feature matrices $\{F^{(i)} \in \mathbb{R}^{n_i \times f_i}\}_{1 \leq i \leq m}$, numbers of clusters $\{k_i\}_{1 \leq i \leq m}$, weights $\{w_a^{(ij)}, w_b^{(i)} \in R_+\}_{1 \leq i < j \leq m}$.
Output: Cluster indicator matrices $\{C^{(p)}\}_{1 \leq p \leq m}$.
Method:

```
1:  Initialize $\{C^{(p)}\}_{1 \leq p \leq m}$ with othonormal matrices.
2:  repeat
3:      for $p = 1$ to $m$ do
4:          Compute the matrix $M^{(p)}$ as in Eq. (9).
5:          Update $C^{(p)}$ by the leading $k_p$ eigenvectors of $M^{(p)}$.
6:      end for
7:  until convergence
8:  for $p = 1$ to $m$ do
9:      transform $C^{(p)}$ into a cluster indicator matrix by the k-means.
10: end for
```

Based on Theorem 4.3 (Bhatia, 1997), at each iterative step we update $C^{(p)}$ as the leading k_p eigenvectors of the matix $M^{(p)}$. After the iteration procedure converges, since the resulting eigen-matrices are not indicator matrices, we need to transform them into cluster indicator matrices by postprocessing (Bach & Jordan, 2004; Zha et al., 2002; Ding & He, 2004). In this paper, we simply adopt the k-means for the postprocessing.

The algorithm, called Spectral Relational Clustering (SRC), is summarized in Algorithm 1. By iteratively updating $C^{(p)}$ as the leading k_p eigenvectors of $M^{(p)}$, SRC makes use of the interactions among the hidden structures of different type of objects. After the iteration procedure converges, the hidden structure for each type of objects is embedded in an eigen-matrix. Finally, we postprocess each eigen-matrix to extract the cluster structure.

To illustrate the SRC algorithm, we describe the specific update rules for the tri-type relational data as shown in Figure 1(b): update $C^{(1)}$ as the leading k_1 eigenvectors of $w_a^{(12)} R^{(12)} C^{(2)} (C^{(2)})^T (R^{(12)})^T$; update $C^{(2)}$ as the leading k_2 eigenvectors of $w_a^{(12)} (R^{(12)})^T C^{(1)} (C^{(1)})^T R^{(12)} + w_a^{(23)} R^{(23)} C^{(3)} (C^{(3)})^T (R^{(23)})^T$; update $C^{(3)}$ as the leading k_3 eigenvectors of $w_a^{(23)} (R^{(23)})^T C^{(2)} (C^{(2)})^T R^{(23)}$.

the computational complexity of SRC can be shown to be $O(tmn^2k)$ where t denotes the number of iterations, $n = \Theta(n_i)$ and $k = \Theta(k_i)$. For sparse data, it could be reduced to $O(tmzk)$ where z denotes the number of nonzero elements.

The convergence of SRC algorithm can be proved. We describe the main idea as follows. Theorem 4.2 and Eq. (8) imply that the updates of the matrices in Line 5 of Algorithm 1 increase the objective function in Eq. (6), and hence equivalently decrease the objective function in Eq.(2). Since the objective function in Eq. (2) has the lower bound 0, the convergence of SRC is guaranteed.

5. Special Cases and Discussions

In this section we discuss special cases of the CFRM model and the SRC algorithm to show that they provide a unified view for the existing spectral clustering algorithms.

5.1. K-means and Spectral Clustering

Traditional ”flat” data can be viewed as a special MTRD with only one feature matrix. In this situation, the objective function in Definition 3.1 is reduced to $L = ||F - CB||^2$, which is the matrix representation for the objective function of the k-means algorithm (Zha et al., 2002). Therefore, by Theorem 4.2, k-means is equivalent to the maximization:

$$\max_{C^T C = I_k} \text{tr}(C^T F F^T C). \tag{10}$$

If we treat FF^T as a graph affinity matrix, the above objective function is equivalent to the objective function of graph partitioning based on average association cut (Shi & Malik, 2000). If we normalize F to be $D^{-1/2}F$ where $D = \text{diag}(FF^T\mathbf{e})$, $\mathbf{e} = [1, 1, \ldots, 1]^T$, the objective function in Eq. (10) is equivalent to the objective function of graph partitioning based on normalized cut (Shi & Malik, 2000). Other versions of graph partitioning can also be formulated to be equivalent to Eq. (10). For the objective function in Eq. (10), SRC iterates only once to compute the leading k eigenvectors of FF^T and postprocesses them to extract the cluster structure. This is exactly the procedure described by Ng et al. (2001). Hence spectral clustering algorithms based on graph partitioning are naturally accommodated in the SRC algorithm.

If we consider FF^T in Eq.(10) as a general similarity matrix which denotes similarities or relations within the same type of objects, SRC is naturally extended to a more general case. In some applications, besides features and relations to other types of objects, a type of objects $\mathcal{X}^{(p)}$ in MTRD may have intra-type relations (here we assume undirected relations), which can be denoted by a symmetric matrix $S^{(p)} \in \mathbb{R}^{n_p \times n_p}$. By treating $S^{(p)}$ the same as $F^{(p)}(F^{(p)})^T$, it is easy to extend SRC to this situation by simply adding an extra term $w_s^{(p)} S^{(p)}$ to $M^{(p)}$ in Eq.(9), where $w_s^{(p)} \in \mathbb{R}$ denotes the weight for $S^{(p)}$. Due to space limitation, theoretic analysis for this extension is omitted.

5.2. Bipartite Spectral Graph Partitioning

Bipartite Spectral Graph Partitioning (BSGP) (Dhillon, 2001; H.Zha & H.Simon, 2001) was proposed to co-cluster bi-type relational data, which can be denoted as one relation matrix $R \in \mathbb{R}^{n_1 \times n_2}$, such as word-document co-occurrence matrix. The BSGP formulates the data as a bipartite graph, whose adjacency matrix can be written as

$\begin{bmatrix} 0 & R \\ R^T & 0 \end{bmatrix}$. After the deduction, spectral partitioning on the bipartite graph is converted to a singular value decomposition (SVD) (Dhillon, 2001; H.Zha & H.Simon, 2001). Under the CFRM model, clustering on bi-type relational data is equivalent to

$$\min_{\substack{(C^{(1)})^T C^{(1)}=I_{k_1} \\ (C^{(2)})^T C^{(2)}=I_{k_2}}} ||R - C^{(1)} A (C^{(2)})^T||^2 \qquad (11)$$

BSGP has the restriction that clusters of different types of objects must have one-to-one associations. Under the CFRM model, this is equivalent to adding an extra constraint on cluster association matrix A to let A be a $k \times k$ diagonal matrix. It immediately follows from the standard result of linear algebra (G.Golub & Loan, 1989) that the minimization in Eq.(11) with the diagonal constraint on A is equivalent to partial SVD. Hence, the CFRM model provides a simple way to understand BSGP. Moreover, since in SRC there are no constraints on A, it provides a novel co-clustering algorithm which does not require that different types of objects have equal number of clusters and one-to-one cluster associations.

6. Experimental Results

In this section, we evaluate the effectiveness of the SRC algorithm on two types of MTRD, bi-type relational data and tri-type star-structured data as shown in Figure 1(a) and Figure 1(b), which represent two basic structures of MTRD and arise frequently in real applications.

The data sets used in the experiments are mainly based on the 20-Newsgroup data (Lang, 1995) which contains about $20,000$ articles from 20 newsgroup. We pre-process the data by removing stop words and file headers and selecting top 2000 words by the mutual information. The word-document matrix R is based on *tf.idf* and each document vector is normalized to the unit norm vector. In the experiments the classis k-means is used for initialization and the final performance score for each algorithm is the average of the 20 test runs unless stated otherwise.

6.1. Clustering on bi-type relational Data

In this section we conduct experiments on a bi-type relational data, word-document data, to demonstrate the effectiveness of SRC as a novel co-clustering algorithm. A representative spectral clustering algorithm, Normalized-Cut (NC) spectral clustering (Ng et al., 2001; Shi & Malik, 2000), and BSGP (Dhillon, 2001), are used as comparisons.

The graph affinity matrix for NC is $R^T R$, i.e., the cosine similarity matrix. In NC and SRC, the leading k eigenvectors are used to extract the cluster structure, where k is the number of document clusters. For BSGP, the second to the $(\lceil \log_2 k \rceil + 1)$th leading singular vectors are used (Dhillon, 2001). K-means is adopted to postprocess the eigenvectors.

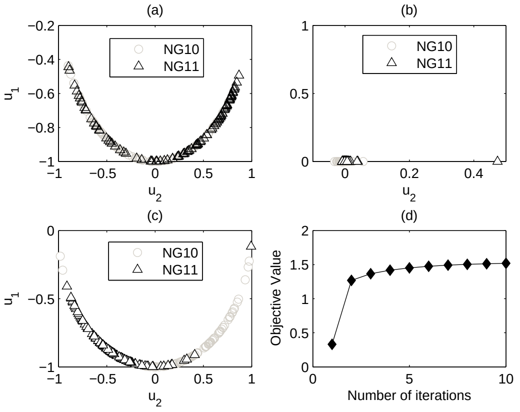

Figure 2. (a), (b) and (c) are document embeddings of multi2 data set produced by NC, BSGP and SRC, respectively (u_1 and u_2 denote first and second eigenvectors, respectively). (d) is an iteration curve for SRC.

Before postprocessing, the eigenvectors from NC and SRC are normalized to the unit norm vector and the eigenvectors from BSGP are normalized as described by Dhillon (2001). Since all the algorithms have random components resulting from k-means or itself, at each test we conduct three trials with random initializations for each algorithm and the optimal one provides the performance score for that test run. To evaluate the quality of document clusters, we elect to use the Normalized Mutual Information (NMI) (Strehl & Ghosh, 2002), which is a standard way to measure the cluster quality.

At each test run, five data sets, multi2 (NG 10, 11), multi3(NG 1,10,20), multi5 (NG 3, 6, 9, 12, 15), multi8 (NG 3, 6, 7, 9, 12, 15, 18, 20) and multi10 (NG 2, 4, 6, 8, 10, 12 ,14 ,16 ,18,20), are generated by randomly sampling 100 documents from each newsgroup. Here NG i means the ith newsgroup in the original order. For the numbers of document clusters, we use the numbers of the true document classes. For the numbers of word clusters, there are no options for BSGP, since they are restricted to equal to the numbers of document clusters. For SRC, it is flexible to use any number of word clusters. Since how to choose the optimal number of word clusters is beyond the scope of this paper, we simply choose one more word clusters than the corresponding document clusters, i.e., 3,4, 6, 9, and 11. This may not be the best choice but it is good enough to demonstrate the flexibility and effectiveness of SRC.

In Figure 2, (a), (b) and (c) show three document embeddings of a multi2 sample, which is sampled from two close newsgroups, *rec.sports.baseball* and *rec.sports.hockey*. In this example, when NC and BSGP fail to separate the document classes, SRC still provides a satisfied separation. The possible explanation is that the adaptive interactions among the hidden structures of word clusters and document clusters remove the noise to lead to better embeddings. (d) shows a typical run of the SRC algorithm. The objective value is the trace value in Theorem 4.2.

Table 1. NMI comparisons of SRC, NC and BSGP algorithms

DATA SET	SRC	NC	BSGP
MULTI2	0.4979	0.1036	0.1500
MULTI3	0.5763	0.4314	0.4897
MULTI5	0.7242	0.6706	0.6118
MULTI8	0.6958	0.6192	0.5096
MULTI10	0.7158	0.6292	0.5071

Table 1 shows NMI scores on all the data sets. We observe that SRC performs better than NC and BSGP on all data sets. This verifies the hypothesis that benefiting from the interactions of the hidden structures of different types of objects, the SRC's adaptive dimensionality reduction has advantages over the dimensionality reduction of the existing spectral clustering algorithms.

6.2. Clustering on Tri-type Relational Data

In this section, we conduct experiments on tri-type star-structured relational data to evaluate the effectiveness of SRC in comparison with other two algorithms for MTRD clustering. One is based on m-partite graph partitioning, Consistent Bipartite Graph Co-partitioning (CBGC) (Gao et al., 2005) (we thank the authors for providing the executable program of CBGC). The other is Mutual Reinforcement K-means (MRK), which is implemented based on the idea of mutual reinforcement clustering as discussed in Section 2.

The first data set is synthetic data, in which two relation matrices, $R^{(12)}$ with 80-by-100 dimension and $R^{(23)}$ with 100-by-80 dimension, are binary matrices with 2-by-2 block structures. $R^{(12)}$ is generated based on the block structure $\begin{bmatrix} 0.9 & 0.7 \\ 0.8 & 0.9 \end{bmatrix}$, i.e., the objects in cluster 1 of $\mathcal{X}^{(1)}$ is related to the objects in cluster 1 of $\mathcal{X}^{(2)}$ with probability 0.9, and so on so forth. $R^{(23)}$ is generated based on the block structure $\begin{bmatrix} 0.6 & 0.7 \\ 0.7 & 0.6 \end{bmatrix}$. Each type of objects has two equal size clusters. It is not a trivial task to identify the cluster structure of this data, since the block structures are subtle. We denote this data as Binary Relation Matrices (TRM) data.

Other three data sets are built based on the 20-newsgroups data for hierarchical taxonomy mining and document clustering. In the field of text categorization, hierarchical taxonomy classification is widely used to obtain a better trade-off between effectiveness and efficiency than flat taxonomy classification. To take advantage of hierarchical classification, one must mine a hierarchical taxonomy from the data set. We can see that words, documents and categories formulate a tri-type relational data, which consists of two relation matrices, a word-document matrix $R^{(12)}$ and a document-category matrix $R^{(23)}$ (Gao et al., 2005).

Table 2. Taxonomy structures for three data sets

DATA SET	TAXONOMY STRUCTURE
TM1	{NG10, NG11}, {NG17, NG18, NG19}
TM2	{NG2, NG3}, {NG8, NG9}, {NG12, NG13}
TM3	{NG4, NG5}, {NG8, NG9}, {NG14, NG15}, {NG17, NG18}

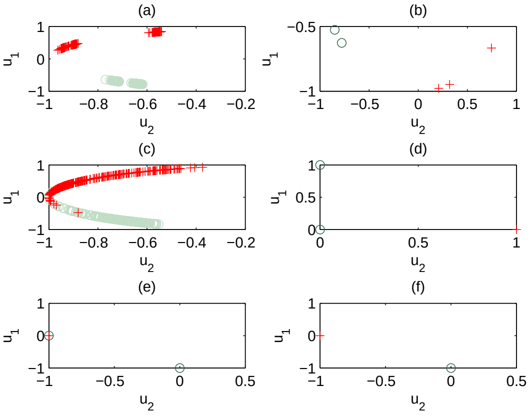

Figure 3. Three pairs of embeddings of documents and categories for the TM1 data set produced by SRC with different weights: (a) and (b) with $w_a^{(12)} = 1, w_a^{(23)} = 1$; (c) and (d) with $w_a^{(12)} = 1, w_a^{(23)} = 0$; (e) and (f) with $w_a^{(12)} = 0, w_a^{(23)} = 1$.

The true taxonomy structures for three data sets, TM1, TM2 and TM3, are listed in Table 2. For example, TM1 data set is sampled from five categories, in which NG10 and NG11 belong to the same high level category *res.sports* and NG17, NG18 and NG19 belong to the same high level category *talk.politics*. Therefore, for the TM1 data set, the expected clustering result on categories should be {NG10, NG11} and {NG17, NG18, NG19} and the documents should be clustered into two clusters according to their categories. The documents in each data set are generated by sampling 100 documents from each category.

The number of clusters used for documents and categories are 2, 3 and 4 for TM1, TM2 and TM3, respectively. For the number of word clusters, we adopt the number of categories, i.e., 5, 6 and 8. For the weights $w_a^{(12)}$ and $w_a^{(23)}$, we simply use equal weight, i.e., $w_a^{(12)} = w_a^{(23)} = 1$. Figure 3 illustrates the effects of different weights on embeddings of documents and categories. When $w_a^{(12)} = w_a^{(23)} = 1$, i.e., SRC makes use of both word-document relations and document-category relations, both documents and categories are separated into two clusters very well as in (a) and (b) of Figure 3, respectively; when SRC makes use of only the word-document relations, the documents are separated with partial overlapping as in (c) and the categories are randomly mapped to a couple of points as in (d); when SRC makes use of only the document-category relations, both documents and categories are incorrectly overlapped as in (e) and (f), respectively, since the document-category matrix itself does not provide any useful information for the taxonomy structure.

Table 3. NMI comparisons of SRC, MRK and CBGC algorithms

DATA SET	SRC	MRK	CBGC
BRM	0.6718	0.6470	0.4694
TM1	1	0.5243	–
TM2	0.7179	0.6277	–
TM3	0.6505	0.5719	–

The performance comparison is based on the cluster quality of documents, since the better it is, the more accurate we can identify the taxonomy structures. Table 3 shows NMI comparisons of the three algorithms on the four data sets. The NMI score of CBGC is available only for BRM data set because the CBGC program provided by the authors only works for the case of two clusters and small size matrices. We observe that SRC performs better than MRK and CBGC on all data sets. The comparison shows that among the limited efforts in the literature attempting to cluster multi-type interrelated objects simultaneously, SRC is an effective one to identify the cluster structures of MTRD.

7. Conclusions and Future Work

In this paper, we propose a general model CFRM for clustering MTRD. The model is applicable to relational data with various structures. Under this model, we derive a novel algorithm SRC to cluster multi-type interrelated data objects simultaneously. SRC iteratively embeds each type of data objects into low dimensional spaces. Benefiting from the interactions among the hidden structures of different types of data objects, the iterative procedure amounts to adaptive dimensionality reduction and noise removal leading to better embeddings. Extensive experiments demonstrate the promise and effectiveness of SRC. We also show that the CFRM model and SRC algorithm provide a unified view to the existing spectral clustering algorithms in the literature. There are a number of interesting potential directions for future research in the CFRM model and SRC algorithm, such as extending CFRM to more general cases with soft clustering or other distance functions and exploring more applications for SRC.

Acknowledgments

This work is supported in part by NSF (IIS-0535162), AFRL Information Institute (FA8750-04-1-0234, FA8750-05-2-0284), and a summer research internship at Yahoo! Research.

References

Bach, F. R., & Jordan, M. I. (2004). Learning spectral clustering. *Advances in Neural Information Processing Systems 16*.

Banerjee, A., Dhillon, I. S., Ghosh, J., Merugu, S., & Modha, D. S. (2004). A generalized maximum entropy approach to bregman co-clustering and matrix approximation. *KDD* (pp. 509–514).

Bhatia, R. (1997). *Matrix analysis*. New York: Springeer-Cerlag.

Chan, P. K., Schlag, M. D. F., & Zien, J. Y. (1993). Spectral k-way ratio-cut partitioning and clustering. *DAC '93* (pp. 749–754).

D.D.Lee, & H.S.Seung (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*, 788–791.

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

Dhillon, I. S., Mallela, S., & Modha, D. S. (2003). Information-theoretic co-clustering. *KDD'03* (pp. 89–98).

Ding, C., He, X., & Simon, H. (2005). On the equivalence of non-negative matrix factorization and spectral clustering. *SDM'05*.

Ding, C. H. Q., & He, X. (2004). Linearized cluster assignment via spectral ordering. *ICML*.

Ding, C. H. Q., He, X., Zha, H., Gu, M., & Simon, H. D. (2001). A min-max cut algorithm for graph partitioning and data clustering. *Proceedings of ICDM 2001* (pp. 107–114).

El-Yaniv, R., & Souroujon, O. (2001). Iterative double clustering for unsupervised and semi-supervised learning. *ECML* (pp. 121–132).

Gao, B., Liu, T.-Y., Zheng, X., Cheng, Q.-S., & Ma, W.-Y. (2005). Consistent bipartite graph co-partitioning for star-structured high-order heterogeneous data co-clustering. *KDD '05* (pp. 41–50).

G.Golub, & Loan, C. (1989). *Matrix computations*. Johns Hopkins University Press.

Hofmann, T. (1999). Probabilistic latent semantic analysis. *Proc. of Uncertainty in Artificial Intelligence, UAI'99*. Stockholm.

Hofmann, T., & Puzicha, J. (1999). Latent class models for collaborative filtering. *IJCAI'99*. Stockholm.

H.Zha, C.Ding, M. X., & H.Simon (2001). Bi-partite graph partitioning and data clustering. *ACM CIKM'01*.

Lang, K. (1995). News weeder: Learning to filter netnews. *ICML*.

Li, T. (2005). A general model for clustering binary data. *KDD'05*.

Long, B., Zhang, Z. M., & Yu, P. S. (2005). Co-clustering by block value decomposition. *KDD'05*.

Ng, A., Jordan, M., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems 14*.

R.O.Duda, P.E.Hart, & D.G.Stork. (2000). *Pattern classification*. New York: John Wiley & Sons.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*, 888–905.

Strehl, A., & Ghosh, J. (2002). Cluster ensembles – a knowledge reuse framework for combining partitionings. *AAAI 2002* (pp. 93–98). AAAI/MIT Press.

Taskar, B., Segal, E., & Koller, D. (2001). Probabilistic classification and clustering in relational data. *Proceeding of IJCAI-01*.

Tishby, N., Pereira, F., & Bialek, W. (1999). The information bottleneck method. *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing* (pp. 368–377).

Wang, J., Zeng, H., Chen, Z., Lu, H., Tao, L., & Ma, W.-Y. (2003). Recom: reinforcement clustering of multi-type interrelated data objects. *SIGIR '03* (pp. 274–281).

Zeng, H.-J., Chen, Z., & Ma, W.-Y. (2002). A unified framework for clustering heterogeneous web objects. *WISE '02* (pp. 161–172).

Zha, H., Ding, C., Gu, M., He, X., & Simon, H. (2002). Spectral relaxation for k-means clustering. *Advances in Neural Information Processing Systems*, *14*.