Collaborative Filtering: Weighted Nonnegative Matrix Factorization Incorporating User and Item Graphs

Quanquan Gu* Jie Zhou* Chris Ding†

Abstract

Collaborative filtering is an important topic in data mining and has been widely used in recommendation system. In this paper, we proposed a unified model for collaborative filtering based on graph regularized weighted nonnegative matrix factorization. In our model, two graphs are constructed on users and items, which exploit the internal information (e.g. neighborhood information in the user-item rating matrix) and external information (e.g. content information such as user's occupation and item's genre, or other kind of knowledge such as social trust network). The proposed method not only inherits the advantages of model-based method, but also owns the merits of memory-based method which considers the neighborhood information. Moreover, it has the ability to make use of content information and any additional information regarding user-user such as social trust network. Due to the use of these internal and external information, the proposed method is able to find more interpretable low-dimensional representations for users and items, which is helpful for improving the recommendation accuracy. Experimental results on benchmark collaborative filtering data sets demonstrate that the proposed methods outperform the state of the art collaborative filtering methods a lot.

Keywords

Graph Regularization, Weighted Nonnegative Matrix Factorization, Weighted Nonnegative Matrix Tri-Factorization, Collaborative Filtering

1 Introduction

Recommendation systems [1] have received more and more attention since the amount of information on the web is increasing rapidly. They attempt to recommend items, e.g. movies, books, music, news, web pages, etc., which are likely to arise the users' interests. Existing recommendation systems can be roughly classified into content-based [2] and collaborative filtering based [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14]. Content-based systems make use of profiles of users or items to characterize their nature. In contrast, collaborative filtering based systems utilize the past user ratings rather than user profiles to predict the unknown user ratings. In the past decades, on account of their general superior performance, collaborative filtering based systems have been more popular than content-based ones. There are also some works on combining content-based and collaborative filtering based systems [15] [16] [17].

Collaborative filtering can be categorized into memory-based methods (a.k.a. neighborhood-based methods) [3] [4] [5] [11] [13], model-based methods [7] [8] [9] [10] [12] [14] and the hybrid [6]. Memory-based methods include user-oriented [3], item-oriented [4] [11] and hybrid [5] [13]. They mainly use the neighborhood information of users or items in the user-item rating matrix. First, they compute similarities between the active user and other users, or, between the active item and other items, and apply them to identify the k most similar neighbors of the active one. Then the unknown rating is predicted by combining the known rating of the neighbors. Despite of their success in the industry, memory-based methods suffer from both the data sparsity and the scalability problem. Due to the sparsity of the user-item rating matrix, memory-based methods may fail to correctly identify the most similar users or items, which in turn sacrifices the recommender accuracy. On the other hand, when the number of users and items are very large in real application, the search of k most similar neighbors is time consuming.

To overcome the limitation of memory-based methods, model-based approaches have been proposed, which establish a model using the observed ratings that can interpret the given data and predict the unknown ratings. Methods in this category include aspect model [8] [7], matrix factorization based model [9] [10] [12] [14] and so on. Due to its efficiency in handling very huge data sets, matrix factorization based model has become one of the most popular models among the model-based methods, e.g. weighted low rank matrix factorization [9], weighted nonnegative matrix factorization (WNMF) [12], maximum margin matrix factorization (MMMF) [10] and probabilistic matrix factorization (PMF) [14].

Although the internal neighborhood information has been widely used in memory-based methods, it is rarely used in model-based methods. On the other

*State Key Laboratory on Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology(TNList), Department of Automation, Tsinghua University, Beijing, China 100084. gqq03@mails.tsinghua.edu.cn, jzhou@tsinghua.edu.cn.

†Department of Computer Science and Engineering, University of Texas at Arlington, 416 Yates Street, Arlington, TX 76019, chqding@uta.edu

hand, external information such as user's demographic information, item's genre information or social trust network plays very important role in content-based recommendation systems. For instance, users with similar occupations may have similar interest in products. Movies of the same genre may be liked similarly by people. And a user usually asks his/her trusted person for the recommendation on products. Nevertheless, this external information is usually neglected in collaborative filtering based system.

Based on these observations, in this paper, we proposed a unified model for collaborative filtering based on graph regularized weighted nonnegative matrix factorization. We construct two graphs on the users and items respectively, to exploit the internal and external information. Thus our method not only inherits the advantages of model-based methods, but also owns the merits of memory-based methods which take into account the neighborhood information. Moreover, our method has the ability to make use of the user's demographic information and item's genre information which is used by content-based recommendation systems and any additional information regarding user-user such as social trust network. Due to the use of these information, our method is able to find more interpretable low-dimensional representations for users and items, which consequently improve the recommendation performance. In addition, we present a graph regularized weighted nonnegative matrix tri-factorization model as an extension, which is more suitable for graph regularization on both items and users simultaneously. Experiments on benchmark data sets demonstrate that the proposed methods have a better performance than the state of the art collaborative filtering methods.

The remainder of this paper is organized as follows. In Section 2, we introduce problem formulation and notations. In Section 3 and Section 4, we introduce user graph and item graph respectively. In Section 5, we present graph regularized weighted matrix factorization model for collaborative filtering. In Section 6, we extend the model to nonnegative matrix tri-factorization. The experiments on benchmark collaborative filtering data sets are demonstrated in Section 7. Finally, we draw a conclusion and point out the future works in Section 8.

2 Problem Formulation & Notations

Before going any further, let's first introduce the problem formulation and notations. Suppose we have M movies and N users, and integer rating values from 1 to K. $\mathbf{X} \in \mathbb{R}_+^{N\times M}$ is the rating matrix, where $\mathbf{X}_{ij}$ represents the rating of user i for movie j. $\mathbf{X}$ is sparse because many elements are missing, and each such element $\mathbf{X}_{ij}$ is assigned 0 to indicate that item j has not been rated by user i. $\mathbf{Y} \in \mathbb{R}^{N\times M}$ is an indicator matrix where $\mathbf{Y}_{ij} = 1$ if user i rated item j and $\mathbf{Y}_{ij} = 0$ otherwise.

We denote by $\mathbf{x}_{i\cdot}, 1 \le i \le N$ and $\mathbf{x}_{\cdot j}, 1 \le j \le M$ the ith row and jth column of $\mathbf{X}$, which represent the ith user's ratings to all items and all users' ratings to the jth item, respectively.

In matrix factorization based model, we usually seek two low-rank matrices $\mathbf{U} \in \mathbb{R}_+^{N\times d}$ and $\mathbf{V} \in \mathbb{R}_+^{M\times d}$. The row vector $\mathbf{u}_{i\cdot}, 1 \le i \le N$ and $\mathbf{v}_{j\cdot}, 1 \le j \le M$ represent the low-dimensional representations of users and items respectively.

3 User Graph

Graph regularization [18] has been widely used in dimensionality reduction[19] [20], clustering [21] [22], coclustering [23], semi-supervised clustering [24] and semi-supervised learning [25] [26] [27]. In our study, in order to incorporate the internal and external information in model-based collaborative filtering, we adopt graph regularization. Generally speaking, we construct two graphs: one is the user graph, the other is the item graph. In this section, we will introduce user graph. The item graph will be introduced in the next section.

We construct an undirected weighted graph $\mathcal{G}_U = (\mathcal{V}_U, \mathcal{E}_U)$ on users, namely *user graph*, whose vertex set $\mathcal{V}_U$ corresponds to users $\{\mathbf{x}_{1\cdot}, \ldots, \mathbf{x}_{N\cdot}\}$. The graph regularization on user graph is formulated as

$$\begin{aligned} & \frac{1}{2}\sum_{ij} ||\mathbf{u}_{i\cdot} - \mathbf{u}_{j\cdot}||^2 W_{ij}^U \\ = & \sum_{i,j} \mathbf{u}_{i\cdot} W_{ij}^U \mathbf{u}_{i\cdot}^T - \sum_{i,j} \mathbf{u}_{i\cdot} W_{ij}^U \mathbf{u}_{j\cdot}^T \\ = & \sum_{i} \mathbf{u}_{i\cdot} D_{ii}^U \mathbf{u}_{i\cdot}^T - \sum_{i,j} \mathbf{u}_{i\cdot} W_{ij}^U \mathbf{u}_{j\cdot}^T \\ = & \mathrm{tr}(\mathbf{U}^T(\mathbf{D}^U - \mathbf{W}^U)\mathbf{U}) \\ = & \mathrm{tr}(\mathbf{U}^T \mathbf{L}_U \mathbf{U}) \end{aligned} \tag{3.1}$$

where $\mathrm{tr}(\cdot)$ denotes the matrix trace, $\mathbf{W}^U = [W_{ij}^U]$ is the symmetric adjacency matrix which encodes the user information, $D_{ii}^U = \sum_j W_{ij}^U$ is a diagonal matrix, and $\mathbf{L}_U = \mathbf{D}^U - \mathbf{W}^U$ is the graph Laplacian [28] of the user graph $\mathcal{G}_U$. The crucial part of graph regularization is the definition of the adjacency matrix $\mathbf{W}^U$, which encodes desired information. In the following, we will introduce how to define proper adjacency matrix to encode various useful information.

3.1 Neighborhood Information The neighborhood information of users in the user-item rating matrix is widely used in user-oriented memory-based methods [3] [29]. Since the user-item rating matrix is very sparse,

ordinary matrix factorization does not work well. Here we will use the neighborhood information of users to aid matrix factorization to get more interpretable low-dimensional representations. The basic assumption of user-oriented memory-based methods is: *if two users have similar ratings on common items, then they probably have similar ratings on the other items.* This can be embodied by the adjacency matrix on user graph as follows,

$$W_{ij}^U = \begin{cases} sim(\mathbf{x}_{i\cdot}, \mathbf{x}_{j\cdot}), & \text{if } \mathbf{x}_{j\cdot} \in \mathcal{N}(\mathbf{x}_{i\cdot}) \text{ or } \mathbf{x}_{i\cdot} \in \mathcal{N}(\mathbf{x}_{j\cdot}) \\ 0, & \text{otherwise.} \end{cases} \tag{3.2}$$

where $\mathcal{N}(\mathbf{x}_{i\cdot})$ denotes the k-nearest neighbor of $\mathbf{x}_{i\cdot}$ and $sim(\mathbf{x}_{i\cdot}, \mathbf{x}_{j\cdot})$ is the similarity between users. There exist a number of different similarities in the literature [4], e.g. cosine distance, Pearson correlation coefficients and adjusted cosine distance. In our experiment, we use cosine distance for simplicity.

3.2 User's Demographic Information Demographic information of users has been widely used in content-based recommendation system. User demographic information includes age, gender and occupation (e.g. administrator, artist, doctor, educator, engineer, executive and so on). The basic assumption of content-based recommendation system which uses user's demographic information is: *if two users have the same gender, similar ages, and the same occupation, then they probably have the similar ratings on the items.* We denote by $\mathbf{f}_i^U$, the feature vector which characterizes the demographic information of user i, then the assumption can be embodied by the adjacency matrix on user graph as follows,

$$W_{ij}^U = sim(\mathbf{f}_i^U, \mathbf{f}_j^U) \tag{3.3}$$

where $sim(\mathbf{f}_i^U, \mathbf{f}_j^U)$ is the similarity between the feature vectors of user i and user j. In our experiment, we use cosine distance for simplicity.

3.3 Social Trust Network Traditional recommendation systems ignore the social interactions or relationship among users. However, the recommendation is sometimes a social activity. For example, we usually ask a friend for a recommendation of movies to see or books to read. [30] revealed that friends are more qualified to make good and useful recommendations than traditional recommendation system which does not consider the social network of users. To this end, [31] proposed a matrix factorization method to exploit the social network information. In our study, we use a directed graph to characterize the social trust network, whose adjacency matrix is defined as

$$W_{ij}^U = \begin{cases} 1, & \text{if user } i \text{ trusts user } j \\ 0, & \text{otherwise.} \end{cases} \tag{3.4}$$

Since the graph is directed, the graph regularization in Eq.(3.1) does not work anymore. We turn to directed graph regularization [32] instead. A direct graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of a finite vertex set $\mathcal{V}$ together with an edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. An edge of a directed graph is an ordered pair $[i, j]$ where i and j are the vertex indices. Each edge associates a weight W_{ij}. The in-degree of the ith vertex is defined as $D_i^- = \sum_{j \to i} W_{ji}$, where $j \to i$ denotes the jth vertex has a directed link pointing to the ith vertex, while out-degree of the ith vertex is defined as $D_i^+ = \sum_{i \to j} W_{ij}$, where $i \to j$ denotes the ith vertex has a directed link pointing to the jth vertex. Given a weighted directed graph, we define a transition probability function of random walk as $P_{ij} = W_{ij}/D_i^+$ for all $[i, j] \in \mathcal{E}$, and 0 otherwise. It is obvious that it satisfies $\sum_j P_{ij} = 1$. Assume the stationary distribution for ith vertex is Π_i. Then it satisfies $\sum_i \Pi_i = 1$ and $\Pi_j = \sum_{i \to j} \Pi_i P_{ij}$. Then graph regularization on directed graph can be formulated as

$$\begin{aligned}
& \frac{1}{2} \sum_{[i,j] \in \mathcal{E}} ||\mathbf{u}_{i\cdot} - \mathbf{u}_{j\cdot}||^2 \Pi_i P_{ij} \\
= \; & \frac{1}{4} \sum_j (\sum_{i \to j} ||\mathbf{u}_{i\cdot} - \mathbf{u}_{j\cdot}||^2 \Pi_i P_{ij} \\
+ \; & \sum_{j \to i} ||\mathbf{u}_{j\cdot} - \mathbf{u}_{i\cdot}||^2 \Pi_j P_{ji}) \\
= \; & \frac{1}{4} \sum_j (2 \sum_{i \to j} \mathbf{u}_{i\cdot} \Pi_i P_{ij} \mathbf{u}_{i\cdot}^T - 2 \sum_{i \to j} \mathbf{u}_{i\cdot} \Pi_i P_{ij} \mathbf{u}_{j\cdot}^T \\
+ \; & 2 \sum_{j \to i} \mathbf{u}_{j\cdot} \Pi_j P_{ji} \mathbf{u}_{j\cdot}^T - 2 \sum_{j \to i} \mathbf{u}_{j\cdot} \Pi_j P_{ji} \mathbf{u}_{i\cdot}^T) \\
= \; & \sum_j \mathbf{u}_{j\cdot} \Pi_j \mathbf{u}_{j\cdot}^T - \frac{1}{2} \sum_j (\sum_{i \to j} \mathbf{u}_{i\cdot} \Pi_i P_{ij} \mathbf{u}_{j\cdot}^T \\
+ \; & \sum_{j \to i} \mathbf{u}_{j\cdot} \Pi_j P_{ji} \mathbf{u}_{i\cdot}^T) \\
= \; & \mathrm{tr}(\mathbf{U}^T (\mathbf{\Pi} - \frac{1}{2} (\mathbf{\Pi P} + \mathbf{P}^T \mathbf{\Pi})) \mathbf{U}) \\
= \; & \mathrm{tr}(\mathbf{U}^T \mathbf{L} \mathbf{U})
\end{aligned} \tag{3.5}$$

where $\mathbf{L} = \mathbf{\Pi} - \frac{1}{2}(\mathbf{\Pi P} + \mathbf{P}^T \mathbf{\Pi})$ is graph Laplacian of directed graph, $\mathbf{\Pi}_{ii} = \Pi_i$ is a diagonal matrix and $\mathbf{P} = [P_{ij}]$. Till now, we have introduced the graph regularization on the directed graph which is associated to the social trust network.

4 Item Graph

Likewise, we construct an undirected weighted graph $\mathcal{G}_V = (\mathcal{V}_V, \mathcal{E}_V)$ on items, namely *item graph*, whose vertex set $\mathcal{V}_V$ corresponds to items $\{\mathbf{x}_{\cdot 1}, \ldots, \mathbf{x}_{\cdot M}\}$. The graph regularization on item graph is formulated as

$$\begin{aligned} & \frac{1}{2}\sum_{ij} ||\mathbf{v}_{i\cdot} - \mathbf{v}_{j\cdot}||^2 W_{ij}^V \\ = \; & \mathrm{tr}(\mathbf{V}^T(\mathbf{D}^V - \mathbf{W}^V)\mathbf{V}) \\ = \; & \mathrm{tr}(\mathbf{V}^T\mathbf{L}_V\mathbf{V}) \end{aligned} \tag{4.6}$$

where $\mathbf{W}^V = [W_{ij}^V]$ is the symmetric adjacency matrix which encodes the item information, $D_{ii}^V = \sum_j W_{ij}^V$ is a diagonal matrix, and $\mathbf{L}_V = \mathbf{D}^V - \mathbf{W}^V$ is the graph Laplacian [28] of the item graph $\mathcal{G}_V$.

4.1 Neighborhood Information As in user graph, the neighborhood information of items in the user-item rating matrix is also very useful and has been used in item-oriented memory-based method [4] [11]. The basic assumption of item-oriented memory-based methods is: *if two items have similar ratings by common users, then they probably have similar ratings by the other users.* This can be embodied by the adjacency matrix on item graph as follows,

$$W_{ij}^V = \begin{cases} sim(\mathbf{x}_{\cdot i}, \mathbf{x}_{\cdot j}), & \text{if } \mathbf{x}_{\cdot j} \in \mathcal{N}(\mathbf{x}_{\cdot i}) \text{ or } \mathbf{x}_{\cdot i} \in \mathcal{N}(\mathbf{x}_{\cdot j}) \\ 0, & \text{otherwise.} \end{cases} \tag{4.7}$$

where $\mathcal{N}(\mathbf{x}_{\cdot i})$ denotes the k-nearest neighbor of $\mathbf{x}_{\cdot i}$ and $sim(\mathbf{x}_{\cdot i}, \mathbf{x}_{\cdot j})$ is the similarity between items. In our experiment, we use cosine distance for simplicity.

4.2 Item's Genre Information Genre information of items has also been widely used in content-based recommendation system. Take movie for example, its genre may be Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Romance, War and so on. The basic assumption of content-based recommendation system which uses item genre is if two items have the same genre, then they probably have the similar ratings by the users. The basic assumption of content-based recommendation system which uses item's genre information is: *if two items have the same genre, then they probably have the similar ratings by the users.* We denote by $\mathbf{f}_i^V$, the feature vector which characterizes the genre information of item i, then the assumption can be embodied by the adjacency matrix on item graph as follows,

$$W_{ij}^V = sim(\mathbf{f}_i^V, \mathbf{f}_j^V) \tag{4.8}$$

where $sim(\mathbf{f}_i^V, \mathbf{f}_j^V)$ is the similarity between the genre information of item i and item j. In our experiment, we use cosine distance for simplicity.

5 Graph Regularized Weighted Nonnegative Matrix Factorization

Till now, we have introduced graph regularization on user and item graphs. In this section, we propose graph regularized weighted nonnegative matrix factorization, which unifies graph regularization and weighted nonnegative matrix factorization. We first review weighted nonnegative matrix factorization in the following.

5.1 Weighted Nonnegative Matrix Factorization Given a nonnegative rating matrix $\mathbf{X} \in \mathbb{R}_+^{N\times M}$, *Weighted Nonnegative Matrix Factorization* (WNMF) [12] aims to find two nonnegative matrices $\mathbf{U} \in \mathbb{R}_+^{N\times d}$ and $\mathbf{V} \in \mathbb{R}_+^{M\times d}$ which minimize the following objective

$$\begin{aligned} J_{WNMF} &= \sum_{i=1}^{N}\sum_{j=1}^{M} \mathbf{Y}_{ij}(\mathbf{X}_{ij} - (\mathbf{U}\mathbf{V}^T)_{ij})^2 \\ &= ||\mathbf{Y} \odot (\mathbf{X} - \mathbf{U}\mathbf{V}^T)||_F^2, \\ &\text{s.t.} \quad \mathbf{U} \geq 0, \mathbf{V} \geq 0, \end{aligned} \tag{5.9}$$

where $\odot$ is Hadamard product (element-wise product). $||\cdot||_F$ is Frobenius norm, $\mathbf{Y}$ is the indicator matrix. Eq.(5.9) can be optimized by iterative multiplicative updating algorithm as follows

$$\begin{aligned} \mathbf{U}_{ij} &\leftarrow \mathbf{U}_{ij}\frac{(\mathbf{Y} \odot \mathbf{X}\mathbf{V})_{ij}}{(\mathbf{Y} \odot (\mathbf{U}\mathbf{V}^T)\mathbf{V})_{ij}} \\ \mathbf{V}_{ij} &\leftarrow \mathbf{V}_{ij}\frac{((\mathbf{Y} \odot \mathbf{X})^T\mathbf{U})_{ij}}{((\mathbf{Y} \odot \mathbf{U}\mathbf{V}^T)^T\mathbf{U})_{ij}} \end{aligned} \tag{5.10}$$

5.2 Objective Based on the graph regularizations on item graph and user graph introduced in previous sections, we propose a *Graph Regularized Weighted Nonnegative Matrix Factorization* (GWNMF), which minimizes the following objective,

$$\begin{aligned} J_{GWNMF} &= ||\mathbf{Y} \odot (\mathbf{X} - \mathbf{U}\mathbf{V}^T)||_F^2 \\ &+ \lambda\mathrm{tr}(\mathbf{U}^T\mathbf{L}_U\mathbf{U}) + \mu\mathrm{tr}(\mathbf{V}^T\mathbf{L}_V\mathbf{V}) \\ &\text{s.t.} \quad \mathbf{U} \geq 0, \mathbf{V} \geq 0 \end{aligned} \tag{5.11}$$

where $\lambda, \mu \geq 0$ are regularization parameters balancing the reconstruction error of WNMF in the first term and graph regularizations in the second and third term. When letting $\lambda = \mu = 0$, GWNMF degenerates to ordinary weighted nonnegative matrix factorization in Eq.(5.9). Note that similar models have been proposed in [21] [22], which address clustering on data manifold.

If we look closely, we can find that there is a potential problem in Eq.(5.11), when we use graph regularization on user graph and item graph together.

That is, when setting λ and μ to positive simultaneously, there is a contradiction between minimizing the three terms. To overcome this problem, we will extend our model to *Graph Regularized Weighted Nonnegative Matrix Tri-Factorization* in the next section. Here, we set λ or μ to positive and the other to zero when we use GWNMF. To make the objective in Eq.(5.11) lower bounded, when $\lambda = 0$, we use L_2 normalization on columns of $\mathbf{V}$ in the optimization, and compensate the norms of $\mathbf{V}$ to $\mathbf{U}$. And when $\mu = 0$, we use L_2 normalization on columns of $\mathbf{U}$ in the optimization, and compensate the norms of $\mathbf{U}$ to $\mathbf{V}$.

5.3 Optimization In the following, we will give the solution to Eq.(5.11). For the sake of convenience, we will see both λ and μ as positive in the derivation. As we see, minimizing Eq.(5.11) is with respect to $\mathbf{U}$ and $\mathbf{V}$, and we cannot give a closed-form solution. We will present an alternating scheme to optimize the objective. In other words, we will optimize the objective with respect to one variable while fixing the other one. This procedure repeats until convergence.

5.3.1 Computation of U Optimizing Eq.(5.11) with respect to $\mathbf{U}$ is equivalent to optimizing

$$\begin{aligned} L(\mathbf{U}) &= ||\mathbf{Y} \odot (\mathbf{X} - \mathbf{U}\mathbf{V}^T)||_F^2 + \lambda \mathrm{tr}(\mathbf{U}^T\mathbf{L}_U\mathbf{U}) \\ \text{s.t.} \quad & \mathbf{U} \geq 0 \end{aligned} \tag{5.12}$$

The derivative of $L(\mathbf{U})$ with respect to $\mathbf{U}$ is

$$\frac{\partial L(\mathbf{U})}{\partial \mathbf{U}} = -2\mathbf{Y}\odot\mathbf{X}\mathbf{V} + 2\mathbf{Y}\odot(\mathbf{U}\mathbf{V}^T)\mathbf{V} + 2\lambda\mathbf{L}_U\mathbf{U} \tag{5.13}$$

Using the Karush-Kuhn-Tucker complementary condition [33] for the nonnegativity of $\mathbf{U}$, we get

$$[-\mathbf{Y}\odot\mathbf{X}\mathbf{V} + \mathbf{Y}\odot(\mathbf{U}\mathbf{V}^T)\mathbf{V} + \lambda\mathbf{L}_U\mathbf{U}]_{ij}\mathbf{U}_{ij} = 0 \tag{5.14}$$

Since $\mathbf{L}_U$ may take any signs, we decompose it as $\mathbf{L}_U = \mathbf{L}_U^+ - \mathbf{L}_U^-$, where $\mathbf{A}_{ij}^+ = (|\mathbf{A}_{ij}| + \mathbf{A}_{ij})/2$ and $\mathbf{A}_{ij}^- = (|\mathbf{A}_{ij}| - \mathbf{A}_{ij})/2$, then

$$\begin{aligned} &[-\mathbf{Y}\odot\mathbf{X}\mathbf{V} + \mathbf{Y}\odot(\mathbf{U}\mathbf{V}^T)\mathbf{V} \\ &\quad +\lambda\mathbf{L}_U^+\mathbf{U} - \lambda\mathbf{L}_U^-\mathbf{U}]_{ij}\mathbf{U}_{ij} = 0 \end{aligned} \tag{5.15}$$

Eq.(5.15) leads to the following updating formula

$$\mathbf{U}_{ij} \leftarrow \mathbf{U}_{ij}\sqrt{\frac{[\mathbf{Y}\odot\mathbf{X}\mathbf{V} + \lambda\mathbf{L}_U^-\mathbf{U}]_{ij}}{[\mathbf{Y}\odot(\mathbf{U}\mathbf{V}^T)\mathbf{V} + \lambda\mathbf{L}_U^+\mathbf{U}]_{ij}}} \tag{5.16}$$

5.3.2 Computation of V Optimizing Eq.(5.11) with respect to $\mathbf{V}$ is equivalent to optimizing

$$\begin{aligned} L(\mathbf{V}) &= ||\mathbf{Y} \odot (\mathbf{X} - \mathbf{U}\mathbf{V}^T)||_F^2 + \mu \mathrm{tr}(\mathbf{V}^T\mathbf{L}_V\mathbf{V}) \\ \text{s.t.} \quad & \mathbf{V} \geq 0 \end{aligned} \tag{5.17}$$

The derivative of $L(\mathbf{V})$ with respect to $\mathbf{V}$ is

$$\frac{\partial L(\mathbf{V})}{\partial \mathbf{V}} = -2(\mathbf{Y}\odot\mathbf{X})^T\mathbf{U} + 2(\mathbf{Y}\odot(\mathbf{U}\mathbf{V}^T))^T\mathbf{U} + 2\mu\mathbf{L}_V\mathbf{V} \tag{5.18}$$

Using the Karush-Kuhn-Tucker complementary condition [33] for the nonnegativity of $\mathbf{V}$, we get

$$[-(\mathbf{Y}\odot\mathbf{X})^T\mathbf{U} + (\mathbf{Y}\odot(\mathbf{U}\mathbf{V}^T))^T\mathbf{U} + \mu\mathbf{L}_V\mathbf{V}]_{ij}\mathbf{V}_{ij} = 0 \tag{5.19}$$

Since $\mathbf{L}_V$ may take any signs, we decompose it as $\mathbf{L}_V = \mathbf{L}_V^+ - \mathbf{L}_V^-$, then

$$\begin{aligned} &[-(\mathbf{Y}\odot\mathbf{X})^T\mathbf{U} + (\mathbf{Y}\odot(\mathbf{U}\mathbf{V}^T))^T\mathbf{U} \\ &\quad +\mu\mathbf{L}_V^+\mathbf{V} - \mu\mathbf{L}_V^-\mathbf{V}]_{ij}\mathbf{V}_{ij} = 0 \end{aligned} \tag{5.20}$$

Eq.(5.20) leads to the following updating formula

$$\mathbf{V}_{ij} \leftarrow \mathbf{V}_{ij}\sqrt{\frac{[(\mathbf{Y}\odot\mathbf{X})^T\mathbf{U} + \mu\mathbf{L}_V^-\mathbf{V}]_{ij}}{[(\mathbf{Y}\odot(\mathbf{U}\mathbf{V}^T))^T\mathbf{U} + \mu\mathbf{L}_V^+\mathbf{V}]_{ij}}} \tag{5.21}$$

5.4 Convergence Analysis In the following, we will investigate the convergence of the updating formulas in Eq.(5.16) and Eq.(5.21). We use the auxiliary function approach [34] to prove the convergence of the algorithm.

DEFINITION 5.1. *[34] $Z(h, h')$ is an auxiliary function for $F(h)$ if the conditions*

$$Z(h, h') \geq F(h), Z(h, h) = F(h),$$

are satisfied.

LEMMA 5.1. *[34] If Z is an auxiliary function for F, then F is non-increasing under the update*

$$h^{(t+1)} = \arg\min_h Z(h, h^{(t)})$$

Proof. $F(h^{(t+1)}) \leq Z(h^{(t+1)}, h^{(t)}) \leq Z(h^{(t)}, h^{(t)}) = F(h^{(t)})$

LEMMA 5.2. *[35] For any nonnegative matrices $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{k\times k}$, $\mathbf{S} \in \mathbb{R}^{n\times k}$,$\mathbf{S}' \in \mathbb{R}^{n\times k}$, and $\mathbf{A}$, $\mathbf{B}$ are symmetric, then the following inequality holds*

$$\sum_{i=1}^{n}\sum_{p=1}^{k}\frac{(\mathbf{A}\mathbf{S}'\mathbf{B})_{ip}\mathbf{S}_{ip}^2}{\mathbf{S}'_{ip}} \geq tr(\mathbf{S}^T\mathbf{A}\mathbf{S}\mathbf{B})$$

THEOREM 5.1. *Let*

$$J(\mathbf{U}) = tr(\lambda\mathbf{U}^T\mathbf{L}_U\mathbf{U} - 2\mathbf{Y}\odot\mathbf{X}\mathbf{V}\mathbf{U}^T + \mathbf{Y}\odot(\mathbf{U}\mathbf{V}^T)\mathbf{V}\mathbf{U}^T) \tag{5.22}$$

Then the following function

$$\begin{aligned} & Z(\mathbf{U},\mathbf{U}') \\ = & \lambda\sum_{ij}\frac{(\mathbf{L}_U^+\mathbf{U}')_{ij}\mathbf{U}_{ij}^2}{\mathbf{U}'_{ij}} \\ - & \lambda\sum_{ijk}(\mathbf{L}_U^-)_{jk}\mathbf{U}'_{ji}\mathbf{U}'_{ki}(1+\log\frac{\mathbf{U}_{ji}\mathbf{U}_{ki}}{\mathbf{U}'_{ji}\mathbf{U}'_{ki}}) \\ - & 2\sum_{ij}(\mathbf{Y}\odot\mathbf{X}\mathbf{V})_{ij}\mathbf{U}'_{ij}(1+\log\frac{\mathbf{U}_{ij}}{\mathbf{U}'_{ij}}) \\ + & \sum_{ij}\frac{(\mathbf{Y}\odot(\mathbf{U}'\mathbf{V}^T)\mathbf{V})_{ij}\mathbf{U}_{ij}^2}{\mathbf{U}'_{ij}} \end{aligned}$$

is an auxiliary function for $J(\mathbf{U})$. Furthermore, it is a convex function in $\mathbf{U}$ and its global minimum is

$$\mathbf{U}_{ij} = \mathbf{U}_{ij}\sqrt{\frac{[\mathbf{Y}\odot\mathbf{X}\mathbf{V}+\lambda\mathbf{L}_U^-\mathbf{U}]_{ij}}{[\mathbf{Y}\odot(\mathbf{U}\mathbf{V}^T)\mathbf{V}+\lambda\mathbf{L}_U^+\mathbf{U}]_{ij}}} \tag{5.23}$$

Proof. It can be proved by Lemma 5.2 and the inequality: $z \geq 1+\log z, \forall z > 0$. Please refer to [22] [36] for more details.

THEOREM 5.2. *Updating $\mathbf{U}$ using Eq.(5.16) will monotonically decrease the value of the objective in Eq.(5.11), hence it converges.*

Proof. By Lemma 5.1 and Theorem 5.1, we can get that $J(\mathbf{U}^0) = Z(\mathbf{U}^0,\mathbf{U}^0) \geq Z(\mathbf{U}^1,\mathbf{U}^0) \geq J(\mathbf{U}^1) \geq \cdots$ So $J(\mathbf{U})$ is monotonically decreasing. Since $J(\mathbf{U})$ is obviously bounded below, we prove this theorem.

THEOREM 5.3. *Updating $\mathbf{V}$ using Eq.(5.21) will monotonically decrease the value of the objective in Eq.(5.11), hence it converges.*

Proof. Note the symmetry of $\mathbf{U}$ and $\mathbf{V}$ in Eq.(5.11), it can be proved analogously as Theorem 5.2.

6 Graph Regularized Weighted Nonnegative Matrix Tri-Factorization

We have introduced graph regularized weighted nonnegative matrix factorization in last section. However, as we previously mentioned, there is a potential problem in GWNMF when we use item graph regularization and user graph regularization together. We will solve this problem by extending GWNMF to *Graph Regularized Weighted Nonnegative Matrix Tri-Factorization*.

6.1 Weighted Nonnegative Matrix Tri-Factorization Given a nonnegative rating matrix $\mathbf{X} \in \mathbb{R}_+^{N\times M}$, *Weighted Nonnegative Matrix Tri-Factorization* (WNMTF) aims to find three nonnegative matrices $\mathbf{U} \in \mathbb{R}_+^{N\times m}$, $\mathbf{S} \in \mathbb{R}_+^{m\times d}$ and $\mathbf{V} \in \mathbb{R}_+^{M\times d}$ which minimize the following objective

$$\begin{aligned} J_{WNMTF} &= \sum_{i=1}^{N}\sum_{j=1}^{M}\mathbf{Y}_{ij}(\mathbf{X}_{ij}-(\mathbf{U}\mathbf{S}\mathbf{V}^T)_{ij})^2 \\ &= ||\mathbf{Y}\odot(\mathbf{X}-\mathbf{U}\mathbf{S}\mathbf{V}^T)||_F^2, \\ \text{s.t.} \quad & \mathbf{U}\geq 0, \mathbf{S}\geq 0, \mathbf{V}\geq 0, \end{aligned} \tag{6.24}$$

where $\odot$ is Hadamard product (element-wise product). $||\cdot||_F$ is Frobenius norm, $\mathbf{Y}$ is the indicator matrix.

6.2 Objective Again, we propose a *Graph Regularized Weighted Nonnegative Matrix Tri-Factorization* (GWNMTF), which minimizes the following objective,

$$\begin{aligned} J_{GWNMTF} &= ||\mathbf{Y}\odot(\mathbf{X}-\mathbf{U}\mathbf{S}\mathbf{V}^T)||_F^2 \\ &+ \lambda\text{tr}(\mathbf{U}^T\mathbf{L}_U\mathbf{U})+\mu\text{tr}(\mathbf{V}^T\mathbf{L}_V\mathbf{V}) \\ \text{s.t.} \quad & \mathbf{U}\geq 0, \mathbf{S}\geq 0, \mathbf{V}\geq 0 \end{aligned} \tag{6.25}$$

where $\lambda, \mu \geq 0$ are regularization parameters balancing the reconstruction error of WNMTF in the first term and graph regularizations in the second and third term. When letting $\lambda = \mu = 0$, GWNMTF degenerates to ordinary weighted nonnegative matrix tri-factorization in Eq.(6.24). Note that similar matrix factorization models have been proposed in [23] and [24], which address co-clustering on manifolds, and semi-supervised co-clustering respectively[1]. To make the objective in Eq.(6.25) lower bounded, we use L_2 normalization on columns of $\mathbf{U}$ and $\mathbf{V}$ in the optimization, and compensate the norms of $\mathbf{U}$ and $\mathbf{V}$ to $\mathbf{S}$.

6.3 Optimization

6.3.1 Computation of S Optimizing Eq.(6.25) with respect to $\mathbf{S}$ is equivalent to optimizing

$$\begin{aligned} L(\mathbf{S}) &= ||\mathbf{Y}\odot(\mathbf{X}-\mathbf{U}\mathbf{S}\mathbf{V}^T)||_F^2 \\ \text{s.t.} \quad & \mathbf{S}\geq 0 \end{aligned} \tag{6.26}$$

The derivative of $L(\mathbf{S})$ with respect to $\mathbf{S}$ is

$$\frac{\partial L(\mathbf{S})}{\partial \mathbf{S}} = -2\mathbf{U}^T(\mathbf{Y}\odot\mathbf{X})\mathbf{V}+2\mathbf{U}^T(\mathbf{Y}\odot(\mathbf{U}\mathbf{S}\mathbf{V}^T))\mathbf{V} \tag{6.27}$$

[1] In [23] [24], the authors omit the nonnegative constraint on $\mathbf{S}$, thus their models can deal with general input data with mixed signs rather than nonnegative data. In collaborative filtering, the elements in the rating matrix are usually positive or missing, i.e. nonnegative, hence we add nonnegative constraint on $\mathbf{S}$

Using the Karush-Kuhn-Tucker complementary condition [33] for the nonnegativity of $\mathbf{S}$, we get

$$[-2\mathbf{U}^T(\mathbf{Y}\odot\mathbf{X})\mathbf{V}+2\mathbf{U}^T(\mathbf{Y}\odot(\mathbf{U}\mathbf{S}\mathbf{V}^T))\mathbf{V}]_{ij}\mathbf{S}_{ij}=0 \tag{6.28}$$

Eq.(6.28) leads to the following updating formula

$$\mathbf{S}_{ij}\leftarrow\mathbf{S}_{ij}\sqrt{\frac{[\mathbf{U}^T(\mathbf{Y}\odot\mathbf{X})\mathbf{V}]_{ij}}{[\mathbf{U}^T(\mathbf{Y}\odot(\mathbf{U}\mathbf{S}\mathbf{V}^T))\mathbf{V}]_{ij}}} \tag{6.29}$$

6.3.2 Computation of U Optimizing Eq.(6.25) with respect to $\mathbf{U}$ is equivalent to optimizing

$$\begin{aligned} L(\mathbf{U}) &= ||\mathbf{Y}\odot(\mathbf{X}-\mathbf{U}\mathbf{S}\mathbf{V}^T)||_F^2+\lambda\mathrm{tr}(\mathbf{U}^T\mathbf{L}_U\mathbf{U}) \\ \text{s.t.}\quad & \mathbf{U}\geq 0 \end{aligned} \tag{6.30}$$

The derivative of $L(\mathbf{U})$ with respect to $\mathbf{U}$ is

$$\frac{\partial L(\mathbf{U})}{\partial\mathbf{U}}=-2\mathbf{Y}\odot\mathbf{X}\mathbf{V}\mathbf{S}^T+2\mathbf{Y}\odot(\mathbf{U}\mathbf{S}\mathbf{V}^T)\mathbf{V}\mathbf{S}^T+2\lambda\mathbf{L}_U\mathbf{U} \tag{6.31}$$

Using the Karush-Kuhn-Tucker complementary condition [33] for the nonnegativity of $\mathbf{U}$, we get

$$[-\mathbf{Y}\odot\mathbf{X}\mathbf{V}\mathbf{S}^T+\mathbf{Y}\odot(\mathbf{U}\mathbf{S}\mathbf{V}^T)\mathbf{V}\mathbf{S}^T+\lambda\mathbf{L}_U\mathbf{U}]_{ij}\mathbf{U}_{ij}=0 \tag{6.32}$$

Since $\mathbf{L}_U$ may take any signs, we decompose it as $\mathbf{L}_U=\mathbf{L}_U^+-\mathbf{L}_U^-$, then

$$\begin{aligned} &[-\mathbf{Y}\odot\mathbf{X}\mathbf{V}\mathbf{S}^T+\mathbf{Y}\odot(\mathbf{U}\mathbf{S}\mathbf{V}^T)\mathbf{V}\mathbf{S}^T \\ &\qquad +\lambda\mathbf{L}_U^+\mathbf{U}-\lambda\mathbf{L}_U^-\mathbf{U}]_{ij}\mathbf{U}_{ij} = 0 \end{aligned} \tag{6.33}$$

Eq.(6.33) leads to the following updating formula

$$\mathbf{U}_{ij}\leftarrow\mathbf{U}_{ij}\sqrt{\frac{[\mathbf{Y}\odot\mathbf{X}\mathbf{V}\mathbf{S}^T+\lambda\mathbf{L}_U^-\mathbf{U}]_{ij}}{[\mathbf{Y}\odot(\mathbf{U}\mathbf{S}\mathbf{V}^T)\mathbf{V}\mathbf{S}^T+\lambda\mathbf{L}_U^+\mathbf{U}]_{ij}}} \tag{6.34}$$

6.3.3 Computation of V Optimizing Eq.(6.25) with respect to $\mathbf{V}$ is equivalent to optimizing

$$\begin{aligned} L(\mathbf{V}) &= ||\mathbf{Y}\odot(\mathbf{X}-\mathbf{U}\mathbf{S}\mathbf{V}^T)||_F^2+\mu\mathrm{tr}(\mathbf{V}^T\mathbf{L}_V\mathbf{V}) \\ \text{s.t.}\quad & \mathbf{V}\geq 0 \end{aligned} \tag{6.35}$$

The derivative of $L(\mathbf{V})$ with respect to $\mathbf{V}$ is

$$\frac{\partial L(\mathbf{V})}{\partial\mathbf{V}}=-2(\mathbf{Y}\odot\mathbf{X})^T\mathbf{U}\mathbf{S}+2(\mathbf{Y}\odot(\mathbf{U}\mathbf{S}\mathbf{V}^T))^T\mathbf{U}\mathbf{S}+2\mu\mathbf{L}_V\mathbf{V} \tag{6.36}$$

Using the Karush-Kuhn-Tucker complementary condition [33] for the nonnegativity of $\mathbf{V}$, we get

$$[-(\mathbf{Y}\odot\mathbf{X})^T\mathbf{U}\mathbf{S}+(\mathbf{Y}\odot(\mathbf{U}\mathbf{S}\mathbf{V}^T))^T\mathbf{U}\mathbf{S}+\mu\mathbf{L}_V\mathbf{V}]_{ij}\mathbf{V}_{ij}=0 \tag{6.37}$$

Since $\mathbf{L}_V$ may take any signs, we decompose it as $\mathbf{L}_V=\mathbf{L}_V^+-\mathbf{L}_V^-$, then

$$\begin{aligned} &[-(\mathbf{Y}\odot\mathbf{X})^T\mathbf{U}\mathbf{S}+(\mathbf{Y}\odot(\mathbf{U}\mathbf{S}\mathbf{V}^T))^T\mathbf{U}\mathbf{S} \\ &\qquad +\mu\mathbf{L}_V^+\mathbf{V}-\mu\mathbf{L}_V^-\mathbf{V}]_{ij}\mathbf{V}_{ij} = 0 \end{aligned} \tag{6.38}$$

Eq.(6.38) leads to the following updating formula

$$\mathbf{V}_{ij}\leftarrow\mathbf{V}_{ij}\sqrt{\frac{[(\mathbf{Y}\odot\mathbf{X})^T\mathbf{U}\mathbf{S}+\mu\mathbf{L}_V^-\mathbf{V}]_{ij}}{[(\mathbf{Y}\odot(\mathbf{U}\mathbf{S}\mathbf{V}^T))^T\mathbf{U}\mathbf{S}+\mu\mathbf{L}_V^+\mathbf{V}]_{ij}}} \tag{6.39}$$

6.4 Convergence Analysis In the following, we will investigate the convergence of the updating formulas in Eq.(6.29), Eq.(6.34) and Eq.(6.39).

THEOREM 6.1. *Let*

$$J(\mathbf{S})=tr(-2\mathbf{U}^T(\mathbf{Y}\odot\mathbf{X})\mathbf{V}\mathbf{S}^T+\mathbf{U}^T(\mathbf{Y}\odot(\mathbf{U}\mathbf{S}\mathbf{V}^T))\mathbf{V}\mathbf{S}^T) \tag{6.40}$$

Then the following function

$$\begin{aligned} &Z(\mathbf{S},\mathbf{S}') \\ =\ & -2\sum_{ij}(\mathbf{U}^T(\mathbf{Y}\odot\mathbf{X})\mathbf{V})_{ij}\mathbf{S}'_{ij}(1+\log\frac{\mathbf{S}_{ij}}{\mathbf{S}'_{ij}}) \\ +\ & \sum_{ij}\frac{(\mathbf{U}^T(\mathbf{Y}\odot(\mathbf{U}\mathbf{S}'\mathbf{V}^T))\mathbf{V})_{ij}\mathbf{S}_{ij}^2}{\mathbf{S}'_{ij}} \end{aligned}$$

is an auxiliary function for $J(\mathbf{S})$. Furthermore, it is a convex function in $\mathbf{S}$ and its global minimum is

$$\mathbf{S}_{ij}=\mathbf{S}_{ij}\sqrt{\frac{[\mathbf{U}^T(\mathbf{Y}\odot\mathbf{X})\mathbf{V}]_{ij}}{[\mathbf{U}^T(\mathbf{Y}\odot(\mathbf{U}\mathbf{S}\mathbf{V}^T))\mathbf{V}]_{ij}}} \tag{6.41}$$

Proof. Please refer to [35].

THEOREM 6.2. *Updating $\mathbf{S}$ using Eq.(6.29) will monotonically decrease the value of the objective in Eq.(6.25), hence it converges.*

Proof. By Lemma 5.1 and Theorem 6.1, we can get that $J(\mathbf{S}^0)=Z(\mathbf{S}^0,\mathbf{S}^0)\geq Z(\mathbf{S}^1,\mathbf{S}^0)\geq J(\mathbf{S}^1)\geq\ldots$ So $J(\mathbf{S})$ is monotonically decreasing. Since $J(\mathbf{S})$ is obviously bounded below, we prove this theorem.

THEOREM 6.3. *Let*

$$\begin{aligned} L(\mathbf{U}) &= tr(\lambda\mathbf{U}^T\mathbf{L}_U\mathbf{U}-2\mathbf{Y}\odot\mathbf{X}\mathbf{V}\mathbf{S}^T\mathbf{U}^T \\ &+ \mathbf{Y}\odot(\mathbf{U}\mathbf{S}\mathbf{V}^T)\mathbf{V}\mathbf{S}^T\mathbf{U}^T) \end{aligned} \tag{6.42}$$

Then the following function

$$\begin{aligned} &Z(\mathbf{U},\mathbf{U}') \\ =\ & \lambda\sum_{ij}\frac{(\mathbf{L}_U^+\mathbf{U}')_{ij}\mathbf{U}_{ij}^2}{\mathbf{U}'_{ij}} \\ -\ & \lambda\sum_{ijk}(\mathbf{L}_U^-)_{jk}\mathbf{U}'_{ji}\mathbf{U}'_{ki}(1+\log\frac{\mathbf{U}_{ji}\mathbf{U}_{ki}}{\mathbf{U}'_{ji}\mathbf{U}'_{ki}}) \\ -\ & 2\sum_{ij}(\mathbf{Y}\odot\mathbf{X}\mathbf{V}\mathbf{S}^T)_{ij}\mathbf{U}'_{ij}(1+\log\frac{\mathbf{U}_{ij}}{\mathbf{U}'_{ij}}) \\ +\ & \sum_{ij}\frac{(\mathbf{Y}\odot(\mathbf{U}'\mathbf{S}\mathbf{V}^T)\mathbf{V}\mathbf{S}^T)_{ij}\mathbf{U}_{ij}^2}{\mathbf{U}'_{ij}} \end{aligned}$$

is an auxiliary function for $L(\mathbf{U})$. *Furthermore, it is a convex function in* $\mathbf{U}$ *and its global minimum is*

$$\mathbf{U}_{ij} = \mathbf{U}_{ij}\sqrt{\frac{[\mathbf{Y}\odot\mathbf{X}\mathbf{V}\mathbf{S}^T + \lambda\mathbf{L}_U^-\mathbf{U}]_{ij}}{[\mathbf{Y}\odot(\mathbf{U}\mathbf{S}\mathbf{V}^T)\mathbf{V}\mathbf{S}^T + \lambda\mathbf{L}_U^+\mathbf{U}]_{ij}}} \tag{6.43}$$

Proof. The proof of this theorem is similar with that of Theorem 5.1, hence we omit it here.

THEOREM 6.4. *Updating* $\mathbf{U}$ *using Eq.(6.34) will monotonically decrease the value of the objective in Eq.(6.25), hence it converges.*

Proof. By Lemma 5.1 and Theorem 6.3, we can get that $J(\mathbf{U}^0) = Z(\mathbf{U}^0, \mathbf{U}^0) \geq Z(\mathbf{U}^1, \mathbf{U}^0) \geq J(\mathbf{U}^1) \geq \dots$ So $J(\mathbf{U})$ is monotonically decreasing. Since $J(\mathbf{U})$ is obviously bounded below, we prove this theorem.

THEOREM 6.5. *Updating* $\mathbf{V}$ *using Eq.(6.39) will monotonically decrease the value of the objective in Eq.(6.25), hence it converges.*

Proof. Note the symmetry of $\mathbf{U}$ and $\mathbf{V}$ in Eq.(6.25), it can be proved analogously as Theorem 6.4.

7 Experiments

In this section, we conduct several experiments to compare our methods with some state of the art collaborative filtering methods, e.g. *Probabilistic Matrix Factorization* (PMF) [14] and WNMF [12] on benchmark collaborative filtering data sets. We will investigate the impacts of internal and external information. All of our experiments have been performed on a Intel Core2 Duo 2.8GHz Windows XP machine with 3GB memory.

7.1 Data Sets In order to evaluate our methods, we use 2 widely used collaborative filtering data sets.

MovieLens100K[2] MovieLens100K consists of 100000 movie ratings of 943 users for 1682 movies. The preference of the user for a specific movie is rated from 1 to 5, and 0 indicates that the movie is not rated by the user. It is very suitable to evaluate the impacts of user demographic information and item genre information because it consists of demographic information (e.g. gender, age and occupation) of users and genre of movies. In detail, we use 2 dimensional feature vector to characterize the user's gender, that is, if the user is male, then the first feature is 1 while the second is 0, and vice versa. We partition the user into 7 age group: 1-17, 18-24, 25-34, 35-44, 45-49, 50-55, 56+. We use 7 dimensional feature vector to describe the user's age group. There are totally 21 occupations: administrator, artist, doctor, educator, engineer, entertainment, executive, healthcare, homemaker, lawyer, librarian, marketing, programmer, retired, salesman, scientist, student, technician, writer, other and none. So we use a 21 dimensional feature vector to describe the user's occupation.We concatenate the 3 feature vectors mentioned above together to get a totally 30 dimensional feature vector $\mathbf{f}_i^U$ for user i. On the other hand, there are 19 genres of movies. Likewise, we use a 19 dimensional feature vector $\mathbf{f}_i^V$ for movie i. We also evaluate the impact of neighborhood information of users and items on this data set.

Epinions[3] In Epinion.com, users can also assign products or reviewers integer ratings form 1 to 5. These ratings and reviews will influence future users when they are deciding whether a product is worth buying or a movie is worth watching. Every user of Epinions maintains a "trust" list which presents a network of trust relationships between users. This network is called the "Web of trust", and is used by Epinions to reorder the product reviews such that a user first sees reviews by users that he or she trusts. So we choose Epinions data set to investigate the impact of social trust network for recommendation. The impact of neighborhood information is also evaluated on this data set. Note that the original data set consists of 49290 uses who have rated on 139738 different items. For the memory limitation in our machine, we chose the users who rated more than 200 ratings and the items which has more than 100 ratings by the users. Thus we obtain a subset consists of 2671 users and 1375 items.

Table 1 summarizes the characteristics of the data sets used in this experiment.

Table 1: Description of the data sets

Data Sets	#users	#items	#ratings
MovieLens100K	943	1682	100000
Epinions	2671	1375	75308

7.2 Evaluation Metrics We use mean absolute error, which is widely used for evaluating collaborative filtering results.

Mean Absolute Error MAE is defined as

$$MAE = \frac{\sum_{i,j}|\mathbf{X}_{ij} - \tilde{\mathbf{X}}_{ij}|}{N} \tag{7.44}$$

where $\mathbf{X}_{ij}$ denotes the rating user i gave to item j, $\tilde{\mathbf{X}}_{ij}$ denotes the predicted rating user i gave to item j.

[2] http://www.grouplens.org/node/73

[3] http://www.trustlet.org/wiki/Downloaded_Epinions_dataset

7.3 Methods & Parameter Settings Here we will introduce the methods which we compared with and their associated parameter settings.

PMF [14]: The dimensionality of the low dimensional representations is set by the grid $\{5, 10, 20\}$. The regularization parameters in the model is set by the grid $\{0.1, 1, 10, 100\}$. And the number of iterations is set to 1000.

WNMF [12]: The parameter settings of WNMF is the same as that of PMF.

GWNMF: To investigate the impact of neighborhood information, user demographic information, item genre information and social trust network, we use GWNMF. In detail, for neighborhood information among users, we set $\mu = 0$ and $\mathbf{W}_U$ as in Eq.(3.2), referred to User Neighbor GWNMF. For user demographic information, we set $\mu = 0$ and $\mathbf{W}_U$ as in Eq.(3.3), referred to User Demographic GWNMF. For neighborhood information of items, we set $\lambda = 0$ and $\mathbf{W}_V$ as in Eq.(4.7), referred to Item Neighbor GWNMF. For item genre information, we set $\lambda = 0$ and $\mathbf{W}_V$ as in Eq.(4.8), referred to Item Genre GWNMF. And for social trust network, we set $\mu = 0$ and $\mathbf{W}_U$ as in Eq.(3.4), referred to Social Trust Network GWNMF. In GWNMF, the regularization parameter λ and μ are set by searching the grid $\{0.1, 1, 10, 100\}$. The dimensionality of the low-dimensional representations, i.e. d, is set by the grid $\{5, 10, 20\}$. For Eq.(3.2) and Eq.(4.7), to avoid parameter tuning, we set the number of nearest neighbors k to $N - 1$ and $M - 1$ respectively, i.e. complete graph. The number of iterations is set to 1000.

GWNMTF: In addition, in order to investigate the impact of using neighborhood information of users and items together, and using user demographic and item genre together, we use GWNMTF. referred to User Neighbor + Item Neighbor GWNMTF and User Demographic + Item Genre GWNMTF respectively. In GWNMTF, the regularization parameters λ and μ are set equal and they are tuned by searching the grid $\{0.1, 1, 10, 100\}$. The dimensionality of low dimensional user representation m and item representation d are set equal too. They are tuned by the grid $\{5, 10, 20\}$. The number of iterations is set to 1500. The other parameter settings are the same as that in GWNMF.

We randomly select 20%, 50% and 80% ratings as training set, and the rest as testing set. The random selection was carried out 10 times independently, and the average MAE is reported.

7.4 Results The experimental results are shown in Table 2 and Table 3 respectively.

From the two tables, we observe that:

1. Our methods which incorporate either internal or external information can improve WNMF greatly, which is a special case of our model. This verifies the effectiveness of our methods.

2. As to our methods, the smaller the number of training ratings (the sparse the user-item rating matrix) is, the more improvement can be achieved. This is because when the rating matrix is sparse, the factorization of the user-item matrix has many possible solutions. In this case, internal information (i.e. neighborhood information of users and items) and external information (i.e. user demographic information, item genre information, social trust network) can alleviate the problem of sparsity, and aid the matrix factorization to obtain more interpretable low-dimensional representations of users and items, which in turn benefit the recommendation accuracy.

3. User demographic information and item genre information is more useful than the neighborhood information. This is consistent with our ordinary intuitions since the external information is more informative than the internal one. The former plays the role as "class labels" in supervised learning, while the latter is "unsupervised". And supervised learning outperforms unsupervised learning in general.

4. Although social trust network benefits the recommendation accuracy, it is not as useful as the neighborhood information. This is probably because that the social trust network is not very reliable compared with the past ratings.

5. Using the neighborhood information of users and items together, or the user's demographic information and item's genre information together obtains comparable or better results than using these information respectively. We believe if we tune λ and μ more carefully, we can achieve much better results by using these information together.

6. In a certain range, the higher the dimensionality of the low-dimensional representation, the better the performance is. In our empirical study, when the dimensionality increase to 50, there is no significant improvement. On the other hand, it is easy to show that the higher the dimensionality, the more computational cost is needed. As a result, in our experiments, we set it around 10 to get a tradeoff.

7. When the number of training ratings is small, e.g. 20% (in other words, the user-item matrix is very sparse), PMF sometimes achieves better

Table 2: MAE comparison on MovieLens100K

Training Data Size	20%			50%			80%		
Dimensionality	d=5	d=10	d=20	d=5	d=10	d=20	d=5	d=10	d=20
PMF	0.6910	0.6720	0.6256	0.6637	0.6120	0.5400	0.6596	0.6019	0.5190
WNMF	0.7100	0.7185	0.6641	0.6597	0.6016	0.5374	0.6344	0.5827	0.5083
User Neighbor GWNMF	0.6924	0.6728	0.6218	0.6450	0.6015	0.5333	0.6337	0.5821	0.5080
Item Neighbor GWNMF	0.6944	0.6722	0.6198	0.6437	0.5987	0.5310	0.6336	0.5816	0.5058
User Demographic GWNMF	0.6872	0.6591	0.6208	0.6440	0.6006	0.5333	0.6343	0.5813	0.5085
Item Genre GWNMF	0.6848	0.6604	0.6181	0.6430	0.5988	0.5316	0.6330	0.5811	0.5069
User Neighbor + Item Neighbor GWNMTF	0.6840	0.6627	0.6348	0.6496	0.5949	0.5291	0.6305	0.5822	0.5016
User Demographic + Item Genre GWNMTF	0.6846	0.6610	0.6310	0.6523	0.5970	0.5349	0.6271	0.5751	0.5013

Table 3: MAE comparison on Epinions

Training Data Size	20%			50%			80%		
Dimensionality	d=5	d=10	d=20	d=5	d=10	d=20	d=5	d=10	d=20
PMF	0.8219	0.7894	0.7584	0.6602	0.5763	0.4839	0.6415	0.5415	0.4144
WNMF	0.8604	0.7659	0.6130	0.6606	0.5725	0.4205	0.6337	0.5290	0.3753
Item Neighbor GWNMF	0.7530	0.6534	0.5311	0.6544	0.5614	0.3960	0.6326	0.5251	0.3694
User Neighbor GWNMF	0.7426	0.6493	0.5208	0.6573	0.5547	0.3925	0.6324	0.5232	0.3655
Social Trust Network GWNMF	0.7937	0.6815	0.5272	0.6583	0.5601	0.3893	0.6332	0.5239	0.3630
User Neighbor + Item Neighbor GWNMTF	0.7271	0.6691	0.5579	0.6558	0.5332	0.3837	0.6325	0.5268	0.3704

results than WNMF. In other most cases, WNMF outperforms PMF.

7.5 Impact of the Regularization Parameters

In this subsection, we will investigate the impact of the regularization parameters. We vary the value of λ or μ under different dimensionality of low dimensional representations and different numbers of training ratings, and plot the MAE on movieLens100K in Figure 1 and the MAE on Epinions in Figure 2. Each column corresponds to different number of training ratings, while each row corresponds to different dimensionality of low-dimensional representations.

We can see: when the number of training ratings is small, large regularization parameter (e.g. 10, 100) is preferred since the internal or external information is critical for matrix factorization. However, when the number of training ratings is large, small regularization parameter (e.g. 0.1, 1) is suitable since the factorization of rating matrix itself works well enough.

8 Conclusion and Future Works

In this paper, we propose a graph regularized nonnegative matrix factorization model for collaborative filtering. We construct two graphs on the item as well as user side, to utilize the internal and external information. Experiments on benchmark data sets demonstrate that the proposed methods outperform many state of the art collaborative filtering methods.

In our future work, we will investigate combining both internal information and external information, to aid matrix factorization. This corresponds to mixed graph regularization on users and items in our model. On the other hand, we will investigate incorporating internal and external information in PMF [14].

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No.60721003, No.60673106 and No.60573062) and the Specialized Research Fund for the Doctoral Program of Higher Education. We thank the anonymous reviewers for their helpful comments.

References

[1] Gediminas Adomavicius and Alexander Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, 2005.

[2] Marko Balabanović and Yoav Shoham, "Fab: content-based, collaborative recommendation," *Commun. ACM*, vol. 40, no. 3, pp. 66–72, 1997.

[3] Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers, and John Riedl, "An algorithmic framework for performing collaborative filtering," in *SIGIR*, 1999, pp. 230–237.

[4] Badrul M. Sarwar, George Karypis, Joseph A. Konstan, and John Riedl, "Item-based collaborative filtering recommendation algorithms," in *WWW*, 2001, pp. 285–295.

[5] Jun Wang, Arjen P. de Vries, and Marcel J. T. Reinders, "Unifying user-based and item-based collaborative filtering approaches by similarity fusion," in *SIGIR*, 2006, pp. 501–508.

[6] David M. Pennock, Eric Horvitz, Steve Lawrence, and C. Lee Giles, "Collaborative filtering by personality diagnosis: A hybrid memory and model-based approach," in *UAI*, 2000, pp. 473–480.

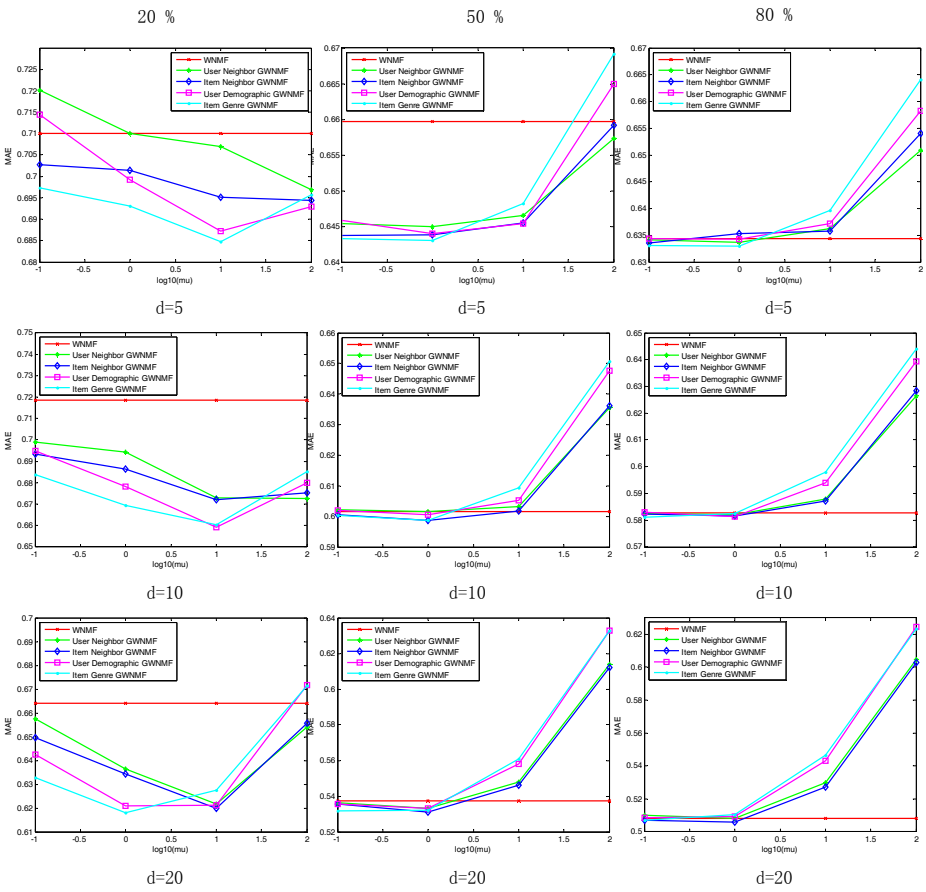

Figure 1: Impact of the Regularization Parameters on movieLens100K.

[7] Thomas Hofmann, "Latent semantic models for collaborative filtering," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 89–115, 2004.

[8] Luo Si and Rong Jin, "Flexible mixture model for collaborative filtering," in *ICML*, 2003, pp. 704–711.

[9] Nathan Srebro and Tommi Jaakkola, "Weighted low-rank approximations," in *ICML*, 2003, pp. 720–727.

[10] Nathan Srebro, Jason D. M. Rennie, and Tommi Jaakkola, "Maximum-margin matrix factorization," in *NIPS*, 2004.

[11] Fei Wang, Sheng Ma, Liuzhong Yang, and Tao Li, "Recommendation on item graphs," in *ICDM*, 2006, pp. 1119–1123.

[12] Sheng Zhang, Weihong Wang, James Ford, and Fillia Makedon, "Learning from incomplete ratings using non-negative matrix factorization," in *SDM*, 2006.

[13] Gang Chen, Fei Wang, and Changshui Zhang, "Collaborative filtering using orthogonal nonnegative matrix tri-factorization," in *ICDM Workshops*, 2007, pp. 303–308.

[14] Ruslan Salakhutdinov and Andriy Mnih, "Probabilistic matrix factorization," in *NIPS*, 2007.

[15] Alexandrin Popescul, Lyle H. Ungar, David M. Pennock, and Steve Lawrence, "Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments," in *UAI*, 2001, pp. 437–444.

[16] Prem Melville, Raymond J. Mooney, and Ramadass Nagarajan, "Content-boosted collaborative filtering for improved recommendations," in *AAAI/IAAI*, 2002, pp. 187–192.

[17] Justin Basilico and Thomas Hofmann, "Unifying collaborative and content-based filtering," in *ICML*, 2004.

[18] Alex J. Smola and Risi Imre Kondor, "Kernels and regularization on graphs," in *COLT*, 2003, pp. 144–158.

[19] Shuicheng Yan, Dong Xu, Benyu Zhang, HongJiang Zhang, Qiang Yang, and Stephen Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, 2007.

[20] Quanquan Gu and Jie Zhou, "Local relevance weighted maximum margin criterion for text classification," in *SDM*, 2009, pp. 1135–1146.

[21] Deng Cai, Xiaofei He, Xiaoyun Wu, and Jiawei Han, "Non-negative matrix factorization on manifold," in *ICDM*, 2008, pp. 63–72.

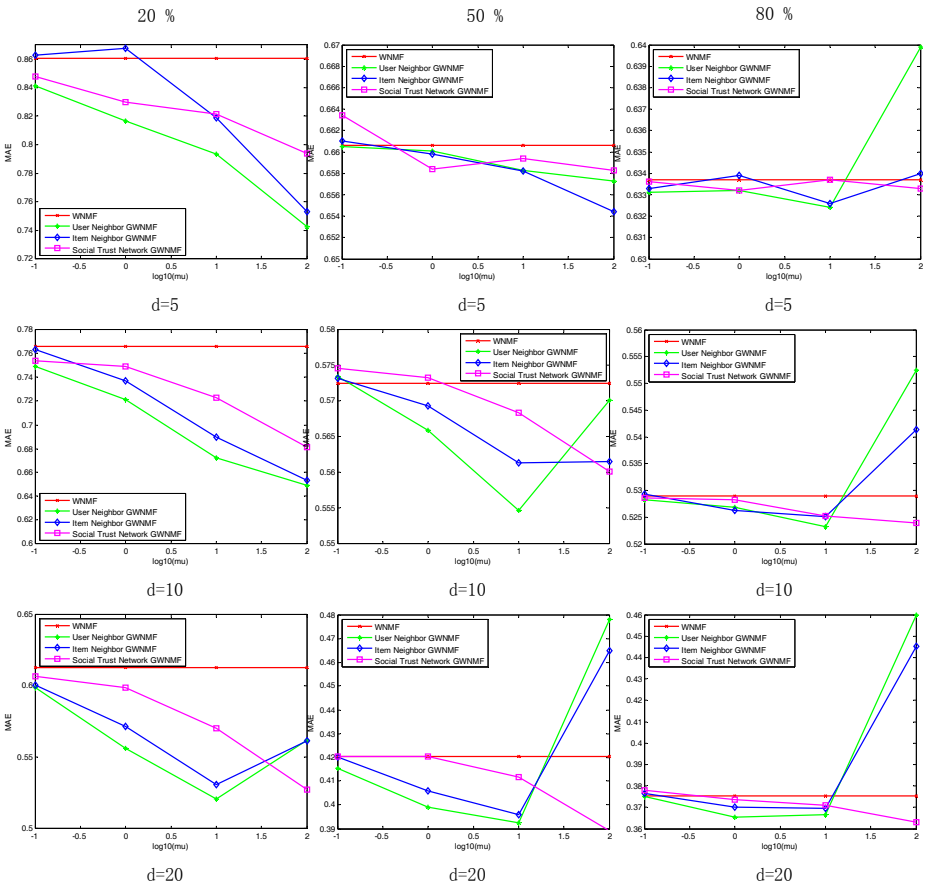

Figure 2: Impact of the Regularization Parameters on Epinions

[22] Quanquan Gu and Jie Zhou, "Local learning regularized nonnegative matrix factorization," in *IJCAI*, 2009, pp. 1046–1051.

[23] Quanquan Gu and Jie Zhou, "Co-clustering on manifolds," in *KDD*, 2009, pp. 359–368.

[24] Fei Wang, Tao Li, and Changshui Zhang, "Semi-supervised clustering via matrix factorization," in *SDM*, 2008, pp. 1–12.

[25] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf, "Learning with local and global consistency," in *NIPS*, 2003.

[26] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.

[27] Quanquan Gu and Jie Zhou, "Transductive classification via dual regularization," in *ECML/PKDD (1)*, 2009, pp. 439–454.

[28] Fan R. K. Chung, *Spectral Graph Theory (CBMS Regional Conference Series in Mathematics, No. 92) (Cbms Regional Conference Series in Mathematics)*, American Mathematical Society, February 1997.

[29] Rong Jin, Joyce Y. Chai, and Luo Si, "An automatic weighting scheme for collaborative filtering," in *SIGIR*, 2004, pp. 337–344.

[30] Punam Bedi, Harmeet Kaur, and Sudeep Marwaha, "Trust based recommender system for semantic web," in *IJCAI*, 2007, pp. 2677–2682.

[31] Hao Ma, Haixuan Yang, Michael R. Lyu, and Irwin King, "Sorec: social recommendation using probabilistic matrix factorization," in *CIKM*, 2008, pp. 931–940.

[32] Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf, "Learning from labeled and unlabeled data on a directed graph," in *ICML*, 2005, pp. 1036–1043.

[33] Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge University Press, Cambridge, 2004.

[34] Daniel D. Lee and H. Sebastian Seung, "Algorithms for non-negative matrix factorization," in *NIPS*, 2000, pp. 556–562.

[35] Chris H. Q. Ding, Tao Li, Wei Peng, and Haesun Park, "Orthogonal nonnegative matrix t-factorizations for clustering," in *KDD*, 2006, pp. 126–135.

[36] Chris H.Q. Ding, Tao Li, and Michael I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, no. 1, 2008.