Selective Knowledge Transfer for Cross-Domain Collaborative Recommendation

HONGWEI ZHANG\textsuperscript{1,2}, XIANGWEI KONG\textsuperscript{3,} (Member, IEEE), AND YUJIA ZHANG\textsuperscript{4}
\textsuperscript{1}School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China
\textsuperscript{2}School of Mathematics, Tonghua Normal University, Tonghua 134002, China
\textsuperscript{3}Department of Data Science and Engineering Management, Zhejiang University, Hangzhou 310058, China
\textsuperscript{4}Department of the School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA 19104-6391, USA

Corresponding author: Xiangwei Kong (kongxiangwei@zju.edu.cn)

This work was supported in part by the National Natural Science Foundation (NSFC) of China under Grant 61772111, and in part by the Foundation for Innovative Research Groups, NSFC, under Grant 71421001.

\textbf{ABSTRACT} Data sparsity is a major challenge for collaborative filtering recommender systems. A promising solution is to utilize feedback or ratings from multiple domains to improve the performance of recommendations in a collective way, known as the cross-domain recommendation. Cross-domain recommendation using heterogeneous feedback is a popular solution, which transfers knowledge from the more easily available auxiliary binary feedback to improve the prediction performance of the target domain. Most of the existing work focuses on the transfer of knowledge between different domains from the same website, where user behavior data in different domains can be fully shared. The existing work mainly assumes that data from different domains can be fully shared. However, due to the constraints of business privacy policies, it is difficult to directly share exactly the same user behavior data between different e-commerce websites. It results in that the user’s latent factors learned in the auxiliary domain cannot be directly transferred to the target domain, otherwise, it will cause a negative transfer issue. In this article, we consider that the auxiliary domain with numerical ratings and target domains with binary feedbacks only share overlapping items rather than users. We propose a Selective Knowledge Transfer for Cross-domain Collaborative Recommendation, called SKT. The proposed SKT framework not only transfers the item’s latent factors learned from the auxiliary domain to the target domain, but also selectively transfers the user’s latent factors learned from the auxiliary domain to the target domain. In addition, due to the introduction of co-graph regularization of user graphs and item graphs, SKT can maintain respective intrinsic geometric structure within each domain and thus avoid negative transfer issue. Extensive experiments conducted on two real-world datasets show that our SKT method is significantly better than all baseline methods at various density levels.

\textbf{INDEX TERMS} Transfer learning (TL), cross-domain recommendation, selective knowledge transfer, sparsity, heterogeneous feedbacks.

I. INTRODUCTION

With the explosive growth of online information, the recommender system has become an important tool to help users find the information they desired effectively [1], [2]. At its core is estimating how likely a user will adopt an item based on historical interactions such as ratings, purchases and clicks [3], [4]. Collaborative filtering (CF) addresses it by assuming that users with similar behaviors in history will also exhibit similar preference on items [5], [6]. Among the various existing collaborative filtering techniques, matrix factorization (MF) which can learn the latent factors for users and items is the most popular one [7]–[10]. However, the recommendation accuracy of MF is highly dependent on the rating matrix. Unfortunately, the rating matrix is usually very sparse in the real world, which forms a barrier to the widespread use of MF in realistic recommender systems. We have noticed that some websites often have a degree of homogeneity in their functionality and provided information. For example, there are many overlapped movies on IMDb\textsuperscript{1} and Douban,\textsuperscript{2} overlapped products on Amazon\textsuperscript{3} and Taobao,\textsuperscript{4}

\textsuperscript{1}https://www.imdb.com.
\textsuperscript{2}http://www.douban.com.
\textsuperscript{3}https://www.amazon.cn
\textsuperscript{4}http://www.taobao.com.

The associate editor coordinating the review of this manuscript and approving it for publication was Nilanjan Dey\textsuperscript{5}.

\begin{thebibliography}{10}
\bibitem{1} The first reference.
\bibitem{2} The second reference.
\end{thebibliography}
and overlapped music on last.fm\textsuperscript{5} and Yahoo Music,\textsuperscript{6} and overlapped videos on Iqiyi\textsuperscript{7} and Tencent Video.\textsuperscript{8} This provides us opportunities to improve the recommendations quality by enriching data, that is, using a rich user-item interaction data (such as purchases, clicks, and ratings) on one website to improve the quality of recommendations (such as ratings prediction) on another website. The task of using the auxiliary data from other domains to improve the recommendation quality of the target domain is called cross-domain recommendation [11]–[15].

Most of the existing cross-domain recommendation techniques focus on transferring the rating patterns [16]–[18] or latent factors [14], [19]–[21] learned from the auxiliary domain to the target domain as a priors or regularization to improve the accuracy of the recommendation. The rating pattern transfer generally requires the user feedback of auxiliary domain and target domain to be homogeneous. When the user rating feedback of the auxiliary and the target domain is heterogeneous, it is necessary to extract the partial rating pattern [22] or deeper latent knowledge [23] from the auxiliary data’s rating pattern for knowledge transfer. The latent factor transfer typically requires that the auxiliary domain and target domain share overlapping users or items. However, due to the constraints of the business privacy policy, it is often difficult to share user behavior data between different e-commerce websites [11]. Therefore, in cross-domain recommendations, it is often more realistic to adopt auxiliary domains that overlap with items in the target domain. Some cross-domain recommendation methods that only share overlapping items between domains have been proposed, and knowledge transfer is achieved by sharing the same item latent factors between domains. For example, the Collective Matrix Factorization (CMF) [24] is proposed for jointly factorizing the target rating matrix and an item-side content matrix with the constraints of sharing the same latent factor of item. Since the users in the auxiliary domain do not overlap with the users in the target domain, the CMF does not constrain the auxiliary domain to share any the latent factors of users with the target domain. However, there may be some users in the auxiliary domain that are similar to the users in the target domain, that is, they have similar preferences on the corresponding items. Selectively transfer the latent factors of users from the auxiliary domain to the target domain, which can further alleviate the sparsity of the target data and help improve the accuracy of user-item ratings prediction in the target domain.

In this work, we aim to help boost the prediction performance of the target domain with numerical ratings (e.g. 5-star rating) by using auxiliary domain with binary ratings (e.g. likes or dislikes) that overlap with the items in the target domain. To effectively utilize the latent factor of extracted from the auxiliary data, we propose a Selective Know Transfer for Cross-domain Collaborative Recommendation (SKT). First, SKT jointly factorizes the auxiliary rating matrix and the target rating matrix with the constraint of sharing the same latent factor of item and selectively sharing the latent factor of user. Second, to ensure positive transfer, we integrate the graph co-regularization of user graph and item graph into the proposed SKT model to maintain the respective intrinsic geometric property of the learned latent factor. The main contributions of this article are summarized as follows.

- We propose that in addition to the latent features of items, the latent features of users with similar preferences between domains can also be used as a bridge for knowledge transfer between domains when only overlapping items are shared between domains. As far as we know, this is the first work of exploring the establishment of user connections between domains with only overlapping items.

- We extend the CMF model [24] by selectively sharing user’s latent knowledge between domains and preserving the intrinsic geometry of entities within domains. In this way, more useful knowledge can be transferred to the target domain while avoiding negative transfer.

- On two real-world datasets, we demonstrate the effectiveness of proposed SKT method at a variety of density levels of 0.01% ~ 1%, and the proposed method SKT shows better performance compared to several state-of-the-art baseline methods.

The organization of this article is as follows. We first review about some related work in Section II. We then formulate the problem and describe the proposed SKT method in Section III, and conduct extensive empirical studies of our SKT and the state-of-the-art methods in Section IV. Finally, we conclude this article in Section V. The notations used through this article are listed in Table 1.

II. RELATED WORK

In this section, related works about cross-domain recommendations are reviewed.

CF exploits user-item behavior interactions (e.g., ratings) only and therefore suffers from the data sparsity issue. To address this issue, one solution is to transfer knowledge from relevant domains, called cross-domain recommendation [8], [16], [19], [25]. According to the overlapping scenarios of entities between the target and auxiliary domains, existing cross-domain recommendation technologies mainly include the following two categories.

A. CROSS-DOMAIN RECOMMENDATION WITH NON-OVERLAPPING ENTITIES

This category assumes that the entities between the auxiliary and target domains do not overlap, and its core idea is to achieve knowledge transfer between the two domains by sharing group-level knowledge. The codebook transfer [16]
TABLE 1. Notations.

| Notation | Description |
|----------|-------------|
| $D_t, D_a$ | target/auxiliary domain |
| $\tau$ | domain index, $\tau \in \{ t, a \}$ |
| $U, V$ | cross-domain user/item sets |
| $|U|, |V|$ | #cross-domain user/item |
| $R_\tau$ | $|U| \times |V|$ rating matrix of $D_\tau$ |
| $Y_\tau$ | $|V| \times |V|$ indicator matrix of $D_\tau$ |
| $U_\tau$ | $|U| \times d$ latent feature matrix of users in $D_\tau$ |
| $V_\tau$ | $|V| \times d$ latent feature matrix of items in $D_\tau$ |
| $S^u_\tau, S^v_\tau$ | user/item similarity graph in $D_\tau$ |
| $L_\tau, \tau$ | graph Laplacian matrix of user graph in $D_\tau$ |
| $L_\tau, g_{\tau}$ | graph Laplacian matrix of item graph in $D_\tau$ |
| $S_\tau^p$ | diagonal sub-gradient matrix of matrix $P$ |
| $P$ | structured sparsity matrix |
| $d$ | #latent factors |
| $\lambda$ | tradeoff parameter |
| $\alpha_U, \alpha_V$ | graph regularization parameter in $D_t$ |
| $\beta_U, \beta_V$ | graph regularization parameter in $D_a$ |
| $\gamma_U, \gamma_V$ | regularization parameter of latent factors |
| $\gamma_p$ | penalty parameter |
| $K$ | #iterations |

assumes that two domains have similar group-level user behavior (i.e., user-item rating pattern, referred to as code-book), and then transfers the extracted codebooks in the auxiliary domain to the target domain to reconstruct the target domain’s rating matrix. An extension method of CBT is called the Rating Matrix Generation Model (RMGM) [17], which relaxes the hard membership constraint for user/item groups to soft membership. Gao et al. [26] relaxed the constraint of sharing the same cluster-level rating model between the auxiliary and target domains in the CBT method, but instead achieved the transfer of knowledge by sharing partial cluster-level rating pattern across multiple rating matrices. Since transferring the knowledge extracted from the auxiliary domain directly to the target domain may result in inconsistent knowledge, the CIT method [18] uses domain adaptation technology to map and adjust the potential groups of users and items in the two domains to maintain consistency in the transfer learning process.

Another solution to the scenario where entities between domains do not overlap at all is to exploit social tags as a bridge for knowledge transfer between domains to achieve cross-domain recommendation. Collective Factorization (CMF) [24] method, not only constraining the auxiliary domain and the target domain to share the same item latent factors, but also allowing interaction between the latent factors of users in the two domains. Transfer by Mixed Factorization (TMF) [21] introduces two interest profiles of user to model the user’s latent preference based on the iTCF [31] method. Although TCF, iTCF and TMF can handle heterogeneous user feedback, their assumptions are too strict, that is, users and items must have a one-to-one mapping between domains, which limits their application in practice. Embedding and Mapping framework for Cross-Domain Recommendation (EMCDR) [14] uses multi-layer perceptron to capture the nonlinear mapping function across domains, which provides high flexibility for learning the domain-specific features of entities in each domain. Since it is usually expensive to identify cross-domain entity correspondences in real-world scenarios. To this end, Zhao et al. [13] proposed an active transfer learning method for cross-domain recommendation, which constructed the entity correspondence between different domains through an entity selection strategy, and then used it as a bridge for knowledge transfer.

In addition to the two types of methods mentioned above, Zhang et al. [32] explored how to achieve cross-domain knowledge transfer when there are only partially overlapping entities between domains. In reality, due to the constraints of company policies, it is difficult to completely share behavior data of different users between websites [11]. To avoid leakage of user privacy, different websites usually only share overlapping items. For this reason, different from the above work, in this article, we focus on how to selectively transfer the latent knowledge of user from the auxiliary binary rating data to reduce the sparsity of the target numerical rating data when only the overlapping items are shared between the auxiliary and target domains.

III. SELECTIVE KNOWLEDGE TRANSFER FOR CROSS-DOMAIN COLLABORATIVE RECOMMENDATION

In this section, we first define the problem setting, then propose Selective Knowledge Transfer for Cross-domain
Collaborative Recommendation (SKT) framework. Lastly, we will show the optimization process of the proposed SKT method, and analyze the convergence and computational complexity of the optimization method.

A. PROBLEM FORMULATION

Suppose there is a user-item numerical rating matrix $R_t$ in the target domain, where $R_t$ is subject to $(R_t)_{ij} \in \{1, 2, 3, 4, 5, ? \}$ ("?" represents a missing value). $Y_t$ is an indicator matrix, $(Y_t)_{ij} \in \{0, 1\}$. $(Y_t)_{ij} = 1$ if user $i$ has rated item $j$, and $(Y_t)_{ij} = 0$ otherwise. Likewise, there is a user-item binary rating matrix $R_a$ in the auxiliary domain, where $R_a$ is subject to $(R_a)_{ij} \in \{0, 1, ?, \}$ ("?" represents a missing value). $Y_a$ is an indicator matrix, $(Y_a)_{ij} \in \{0, 1\}$. $(Y_a)_{ij} = 1$ if user $i$ has rated item $j$, and $(Y_a)_{ij} = 0$ otherwise. Here, we assume that $R_t$ and $R_a$ only share overlapping items, that is, the items in them are aligned, but do not share overlapping users. Let $U = \{U_1, U_2, \cdots \}$ and $V = \{V_1, V_2, \cdots \}$ denote the cross-domain user and item sets, respectively. Denote $D_t$ the target domain, $D_a$ the auxiliary domain and $\tau \in \{t, a\}$ the domain index. Our task is to predict the missing values of the extremely sparse rating matrix $R_t$ in the target domain $D_t$ by selectively transferring the rating knowledge of the dense rating matrix $R_a$ in the auxiliary domain $D_a$.

B. SKT METHOD

The proposed SKT framework jointly decomposes the auxiliary binary rating matrix and the target numerical rating matrix, with the constraints of sharing item-specific latent factors and selectively sharing user-specific latent factors. In addition, the co-graph regularizations of user and item graphs from two domains are integrated into the collective matrix factorization framework, so that the learned latent factors can preserve their intrinsic geometric property to avoid negative transfer issues. The graph model is shown in Figure 1.

1) WEIGHTED COLLECTIVE MATRIX FACTORIZATION

Given the nonnegative target numerical rating matrix $R_t \in \mathbb{R}^{\vert U \vert \times \vert V \vert}$ and auxiliary binary rating matrix $R_a \in \mathbb{R}^{\vert U \vert \times \vert V \vert}$, the latent factors of each rating matrix can be extracted by Weighted Nonnegative Matrix Factorization (WNMF) [33]. In WNMF, the nonnegative user-item rating matrix $R_t \in \mathbb{R}^{\vert U \vert \times \vert V \vert}$ can be decomposed into two low rank matrices $U_t \in \mathbb{R}^{\vert U \vert \times d}$ and $V_t \in \mathbb{R}^{\vert V \vert \times d}$, such that the reconstruction error of matrix $R_t$ is minimized. WNMF amounts to the following optimization problem:

$$
\min_{U_t, V_t, s.t. V_t \equiv V_a} \sum_{i=1}^{\vert U \vert} \sum_{j=1}^{\vert V \vert} (Y_t)_{ij}((R_t)_{ij} - (U_t V_t^T)_{ij})^2
$$

where $\circ$ denotes the element-wise product of matrices. $\| \cdot \|_F$ is the Frobenius norm, for a matrix $X \in \mathbb{R}^{m \times n}$, the Frobenius norm as $\|X\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij}^2}$, which provides a statistical metric of the magnitude of the matrix elements. $Y_t$ is the indicator matrix representing whether the rating in $R_t$ is observed or not. $U_t \in \mathbb{R}^{\vert U \vert \times d}$ is the latent factor matrix of users where the $i$th row $(U_t)_i$ represents $d$ latent preference of user $U_i$. $V_t \in \mathbb{R}^{\vert V \vert \times d}$ is the latent factor matrix of items where $j$th row $(V_t)_j$ represents $d$ latent features of item $V_j$. Since the target numerical rating matrix $R_t$ shares the same item as the auxiliary binary rating matrix $R_a$, the prediction accuracy can be improved by sharing the common latent factors of items underlying these two rating data. Similar to CMF [24], we extend basic WNMF to simultaneously factorize two relevant matrices, which leads to Weighted Collective Matrix Factorization (WCMF)

$$
\min_{U_t, U_a, V_t, V_a \geq 0} \| Y_t \circ (R_t - U_t V_t^T) \|_F^2 + \lambda \| Y_a \circ (R_a - U_a V_a^T) \|_F^2
$$

s.t. $V_t \equiv V_a \equiv V$.

(2)

where $\lambda$ is a trade-off parameter for balancing auxiliary and target data. The above optimization problem can be further expressed as

$$
\min_{U_t, U_a, V_t, V_a \geq 0} \| Y_t \circ (R_t - U_t V_t^T) \|_F^2 + \lambda \| Y_a \circ (R_a - U_a V_a^T) \|_F^2 + \gamma \| V_t - V_a \|_F^2,
$$

(3)

where the tradeoff parameters $\gamma$ represent the confidence on the auxiliary data. In formula (3), the item latent factors in the auxiliary domain are transferred to the target domain by regularization term $\| V_t - V_a \|_F^2$. However, when the target data is extremely sparse, it may not be sufficient to transfer only the latent factors of item learned from the auxiliary domain to the target domain.
2) SELECTIVE TRANSFER THE LATENT FACTOR OF USERS LEARNED FROM AUXILIARY DOMAIN

Although the auxiliary and target domains do not share overlapping users, some users in the auxiliary domain may have similar interactions with users in the target domain on the corresponding items, that is, they have similar preferences. These latent preference information is encoded into the user’s latent factor by collective matrix factorization. Adaptively selecting the latent factor of the user in the auxiliary domain that is useful to the target domain, and then transferring them to the target domain, which will further alleviate the sparsity issue of the target data and improve the predictive performance of cross-domain recommendation. To achieve this, we introduce transformation matrix \( P \), while constraining \( P \) with row-sparsity, so that transformation matrix \( P \) can adaptively select the latent factors of users in \( U_a \) and transfer them to the target domain. To this end, we integrate the constraint \( U_t = U_a P^T \) into framework (3). In addition, to introduce row-sparsity to the transformation matrix \( P \), we propose to impose the \( l_{2,1} \)-norm structured sparsity regularization on the transformation matrix \( P \). Therefore, we can obtain the following optimization objective

\[
\min_{U_t, U_a, V_t, V_a, P \succeq 0} \left\| Y_t \odot \left( R_t - U_t V_t^T \right) \right\|_F^2 + \lambda \left\| Y_a \odot \left( R_a - U_a V_a^T \right) \right\|_F^2 + \gamma_V \left\| V_t - V_a \right\|_F^2 + \gamma_P \left\| P \right\|_{2,1}^{2,1},
\]

where \( \left\| \cdot \right\|_F \) denotes the Frobenius norm, \( \left\| \cdot \right\|_2 \) is defined as follows

\[
\left\| \cdot \right\|_2 = \frac{1}{2} \sum_{i,j} (S^u_{ij})_{ij} \left\| u_i^T - u_j^T \right\|_2^2 = \text{tr}(U_t^T L_u^u U_t),
\]

\[
\left\| \cdot \right\|_2 = \frac{1}{2} \sum_{i,j} (S^v_{ij})_{ij} \left\| v_i^T - v_j^T \right\|_2^2 = \text{tr}(V_t^T L_v^v V_t),
\]

where \( (S^u_{ij})_{ij} \) and \( (S^v_{ij})_{ij} \) denote the cosine similarity between \( r_i^T \) and \( r_j^T \), and between \( r_i^T \) and \( r_j^T \), respectively. They are defined as

\[
(S^u_{ij})_{ij} = \frac{r_i^T r_j^T}{\left\| r_i^T \right\|_2 \left\| r_j^T \right\|_2}, \quad (S^v_{ij})_{ij} = \frac{r_i^T r_j^T}{\left\| r_i^T \right\|_2 \left\| r_j^T \right\|_2},
\]

where \( r_i^T \) and \( r_j^T \) denote the i-th row and the j-th column of the user-item rating matrix \( R_t \), respectively.

4) OPTIMIZATION FRAMEWORK

Integrate the graph regularization of into the framework (5), and get the final optimization framework as follows

\[
\min_{U_t, U_a, V_t, V_a, P \succeq 0} \frac{J}{J} = \left\| Y_t \odot \left( R_t - U_t V_t^T \right) \right\|_F^2 + \lambda \left\| Y_a \odot \left( R_a - U_a V_a^T \right) \right\|_F^2 + \gamma_V \left\| V_t - V_a \right\|_F^2 + \gamma_P \left\| P \right\|_{2,1}^{2,1} + \alpha_U \text{tr}(U_t^T L_u^u U_t) + \beta_V \text{tr}(V_t^T L_v^v V_t) + \alpha_U \text{tr}(U_t^T L_u^u U_t) + \beta_V \text{tr}(V_t^T L_v^v V_t),
\]

where \( \alpha_U \) and \( \alpha_V \) are the graph regularization parameters of users and items in the target domain, respectively. \( \beta_U \) and \( \beta_V \) are the graph regularization parameters of users and items in the auxiliary domain, respectively.

C. LEARNING THE SKT

We can use an alternate minimization algorithm to optimize the proposed SKT framework. Specifically, we optimize a variable and calculate its update rules while fixing the remaining variables. Repeat the process until convergence.

1) LEARNING \( U_t \) AND \( U_a \)

Fixing other variables to solve \( U_t \), then the objective function in equation (10) can be expressed as

\[
\min_{U_t \succeq 0} J(U_t) = \left\| Y_t \odot \left( R_t - U_t V_t^T \right) \right\|_F^2 + \gamma_U \left\| U_t - U_a P^T \right\|_F^2 + \alpha_U \text{tr}(U_t^T L_u^u U_t).
\]

The derivative of \( J(U_t) \) with respect to \( U_t \) is

\[
\frac{\partial J(U_t)}{\partial U_t} = -2(Y_t \odot R_t) V_t + 2(Y_t \odot (U_t V_t^T)) V_t - 2Y_U U_a P^T + 2Y_U U_t + 2\alpha U L_u^u U_t.
\]
Using the Karush-Kuhn-Tucker complementary condition for the nonnegativity of $U_t$ and letting $\frac{\partial J(V_t)}{\partial U_t} = 0$, we can obtain

$$\begin{align*}
-(Y_t \circ R_t)V_t + (Y_t \circ (U_tV_t^T))V_t - \gamma V U_a P^T + \gamma V U_t + \\
+ \alpha U L^+ u_t U_t \circ V_t = 0.
\end{align*}$$

Since $L^u_t$ may take any signs, we replace it with $L^u_t = L^{u+}_t - L^{u-}_t$, where $L^{u+}_t = \frac{1}{2}(L^{u+}_t + L^{u-}_t)$, $L^{u-}_t = \frac{1}{2}(L^{u+}_t - L^{u-}_t)$, then

$$\begin{align*}
&-(Y_t \circ R_t)V_t + (Y_t \circ (U_tV_t^T))V_t - \gamma V U_a P^T + \gamma V U_t + \\
&+ \alpha U L^{u+}_t U_t - \alpha U L^{u-}_t U_t \circ V_t = 0.
\end{align*}$$

We obtain the following updating rule for learning $U_t$

$$U_t = U_t \circ \left[ (Y_t \circ R_t)V_t + \gamma V U_a P^T + \alpha U L^{u+}_t U_t \right],$$

where $\circ$ denotes element-wise division. Similarly, we can obtain the updating rules for learning $U_a$ as

$$U_a = U_a \circ \left[ \frac{\lambda(Y_a \circ R_a V_a + \gamma V U_a P + \beta V L^{u+}_a U_a)}{\lambda(Y_a \circ (U_a V_a^T))V_a + \gamma V U_a P^T + \beta V L^{u+}_a U_a} \right].$$

2) LEARNING $V_t$ AND $V_a$

Likewise, fixing other variables to solve $V_t$, then the objective function in equation (10) can be expressed as

$$\min_{V_t \geq 0} J(V_t) = \left\| Y_t \circ (R_t - U_t V_t^T) \right\|_F^2 + \gamma \parallel V_t \parallel_F^2 + \alpha \parallel r(V_t^T L^+_t V_t) \parallel_F^2.$$

The derivative of $J(V_t)$ with respect to $V_t$ is

$$\frac{\partial J(V_t)}{\partial V_t} = -2(Y_t \circ R_t)^T U_t + 2(Y_t \circ (U_t V_t^T))^T U_t - 2\gamma v U_a + 2\gamma v V_t + 2\alpha v L^+_t V_t.$$

Using the Karush-Kuhn-Tucker complementary condition for the nonnegativity of $V_t$ and letting $\frac{\partial J(V_t)}{\partial V_t} = 0$, we can obtain

$$\begin{align*}
&-(Y_t \circ R_t)^T U_t + (Y_t \circ (U_t V_t^T))^T U_t - \gamma \parallel V_t \parallel_F^2 + \\
&+ \gamma v V_t + \alpha v L^+_t V_t \circ V_t = 0.
\end{align*}$$

Since $L^+_t$ may take any signs, we replace it with $L^+_t = L^{+_+}_t - L^{+_+}_t$, where $L^{+_+}_t = \frac{1}{2}(L^{+_+}_t + L^{+_+}_t)$, $L^{+_+}_t = \frac{1}{2}(L^{+_+}_t - L^{+_+}_t)$, then

$$\begin{align*}
&-(Y_t \circ R_t)^T U_t + (Y_t \circ (U_t V_t^T))^T U_t - \gamma \parallel V_t \parallel_F^2 + \\
&+ \gamma v V_t + \alpha v L^{+_+}_t V_t - \alpha v L^{+_+}_t V_t \circ V_t = 0.
\end{align*}$$

We obtain the following updating rule for learning $V_t$

$$V_t = V_t \circ \left[ (Y_t \circ R_t)^T U_t + \gamma v V_t + \alpha v L^+_t V_t \right].$$

3) LEARNING $P$

Fixing other variables to solve $P$, then the objective function in equation (10) can be expressed as

$$\min_{P \geq 0} J(P) = \gamma U_a P^T + \gamma U_a V_t + \gamma U_a P^T + \gamma U_a L^{u+}_a U_a \parallel P \parallel_{2,1}.$$

The derivative of $J(P)$ with respect to $P$ is

$$\frac{\partial J(P)}{\partial P} = -2\gamma U_a V_t P + 2\gamma U_a P U_a T U_a + 2\gamma \parallel G_P \parallel.$$

Since $\parallel P \parallel_{2,1}$ is a non-smooth function at zero, we compute its sub-gradient as $\frac{\partial J(P)}{\partial P} = 2G_P P$ [37], [38], where $G_P$ is a diagonal sub-gradient matrix with $ith$ element equal to

$$G_P = \begin{cases} \frac{1}{2 \parallel P \parallel}, & \parallel P \parallel \neq 0 \\ 0, & \parallel P \parallel = 0. \end{cases}$$

Using the Karush-Kuhn-Tucker complementary condition for the nonnegativity of $P$ and letting $\frac{\partial J(P)}{\partial P} = 0$, we can obtain

$$\begin{align*}
&-\gamma U_a V_t^T U_a + 2\gamma U_a V_t P + \gamma U_a P U_a T U_a + 2\gamma \parallel G_P \parallel P \circ P = 0.
\end{align*}$$

Then we obtain the following updating rule for learning $P$

$$P = P \circ \left[ V_t U_a P + \gamma U_a P U_a T U_a + \gamma U_a P U_a T U_a \right].$$

D. CONVERGENCE ANALYSIS AND TIME COMPLEXITY

Based on the above updating rules for learning latent factors and structured sparse matrix, we can prove that the learning algorithm is convergent.

**Theorem 1:** Updating $U_t$, $U_a$, $V_t$, $V_a$ and $P$ sequentially by Equations (11) ~ (15) will monotonically decrease the objective function in Equation (10) until convergence.

**Theorem 1** can be proved by the auxiliary function method mentioned in [34], [39], [40].

We summarize the learning algorithm in Algorithm 1. The time complexity of SKT and other baseline method are listed in Table 3, where $q$ and $\bar{q}$ denote the number of observed ratings in the target and auxiliary rating matrix, respectively. $|P|$ and $|V|$ represent the average number of positive and negative feedbacks by a certain user in the target rating matrix, respectively. $K'$ denotes the total number of iterations of [32, Algorithms 1 and 2].

In addition, our proposed algorithm 1 is universal, it can not only deal with heterogeneous user feedback, but also with homogeneous user feedback.

IV. EXPERIMENTS

A. DATASETS AND EVALUATION METRICS

1) DATASETS

We adopt two real-world datasets to evaluate the proposed SKT method. The first dataset, Netflix-MovieLens, contains user and aligned movies from two public benchmark sets, namely the Netflix Prize and the MovieLens project. The Netflix\(^9\) rating data contains more than $10^8$ rating with value in

http://netflixprize.com/index.html.
Algorithm 1 SKT: Selective Knowledge Transfer

```
Input: Datasets $R_t, Y_t, \tau \in \{t, a\}$, parameters $\lambda, \gamma_U, \gamma_V, \gamma_P, \alpha_U, \alpha_V, \beta_U, \beta_V$.
Output: Latent factors $U_t, V_t, \tau \in \{t, a\}$ and transform matrix $P$.

Step 1. Scale ratings in $R_t ((R_t)_{ui} = ((R_t)_{ui} - 1)/4, (Y_t)_{ui} = 1, u = 1, 2, \ldots, |U|; i = 1, 2, \ldots, |V|)$.

Step 2. Construct similarity matrix $S^t_U$ and $S^t_V$ by Equations (8) and (9), $\tau \in \{t, a\}$.

Step 3. Randomly initialize $U_t, V_t, P, \tau \in \{t, a\}$.

Step 4. Update $U_t, V_t, P, \tau \in \{t, a\}$.

for $iter = 1$ to $K$
do

Step 4.1. Fix $V_t, U_a, V_a$ and $P$, update $U_t$ as show in Eq.(11).

Step 4.2. Fix $U_t, V_t, U_a$ and $P$, update $U_a$ as show in Eq.(12).

Step 4.3. Fix $U_t, U_a, V_a$ and $P$, update $V_t$ as show in Eq.(13).

Step 4.4. Fix $U_t, U_a, V_t$ and $P$, update $V_a$ as show in Eq.(14).

Step 4.5. Fix $U_t, U_a, V_t$ and $V_a$, update $P$ as show in Eq.(15).

end
```

\(1, 2, 3, 4, 5\), which are given by more than \(4.8 \times 10^5\) users on around 1.8 \(\times 10^4\) movies. The MovieLens 20M\(^{10}\) rating data contain 2.0 \(\times 10^7\) ratings with values in \(\{0.5, 1, 1.5, \ldots, 5\}\), which are given by more than 1.3 \(\times 10^5\) users on around 2.7 \(\times 10^4\) movies. We first randomly extract a 5000 \(\times 5000\) dense rating matrix $R_t$ from Netflix data, and then extract an item side auxiliary data $R_a$ of size 5000 \(\times 5000\) from the MovieLens data by identifying the movies appearing both in MovieLens 20M and Netflix. Clearly $R_t$ and $R_a$ share only common items but no users. Similar to [20], [41], we adopt a preprocessing approach on $R_a$ by relabeling ratings with value less than 4 in $X_a$ as 0 (dislike), and then ratings with value greater than or equal to 4 as 1 (like).

The second dataset was crawled from an online social network, i.e., Goodreads [42], where users give ratings to books. The Goodreads\(^{11}\) rating data contains more than 3.1 \(\times 10^7\) rating with values \(\{1, 2, 3, 4, 5\}\), which are given by more 3.0 \(\times 10^5\) users on around 1.9 \(\times 10^6\) movies. We randomly extract a 10000 \(\times 5000\) dense rating matrix $R$ from the Goodreads data, and take the sub-matrices $R_t = R_{1:5000, 1:5000}$ as the target rating matrix, and $R_a = R_{5001:10000, 1:5000}$ as the item side auxiliary data, so that $R_t$ and $R_a$ share only common items but not common users. The Goodread dataset also contains user-to-user relationships that are not used by us. Since $R_t$ and $R_a$ are equivalent to being randomly extracted from the entire dataset, which can ensure that users in both $R_t$ and $R_a$ do not have too many or too few user-to-user relationships compared to the distribution in the overall dataset. In other words, $R_t$ and $R_a$ randomly drawn from the Goodreads dataset will not introduce any skew regarding the user-to-user graph. To simulate heterogeneous auxiliary and target domain data, we adopt a preprocessing approach on $R_a$ by relabeling ratings with value less than 4 in $X_a$ as 0 (dislike), and then ratings with value greater than or equal to 4 as 1 (like).

In all of our experiments, the target domain rating set from $R_t$ is randomly split into training and test sets, $R_T$, $R_E$, with 50% ratings, respectively. $R_E$ is kept unchanged, while different number of observed ratings of 2500, 12500, 25000, 125000 and 250000 are randomly picked from $R_T$ for training, with different density levels of 0.01%, 0.05%, 0.1%, 0.5% and 1%. The final datasets are summarized in Table 2.

### Table 2. Description of target and auxiliary data, each of which contains 5000 users and 5000 items.

| Dataset     | Form                     | Density |
|-------------|--------------------------|---------|
| Goodreads   | target(training)          | \{1, 2, 3, 4, 5, ?\} | \leq 1\% |
| (subset)    | target(test)             | \{1, 2, 3, 4, 5, ?\} | 2.93\%  |
|             | auxiliary                | \{0, 1, ?\}   | 5.87\%  |
| Netflix-MovieLens | target(training)          | \{1, 2, 3, 4, 5, ?\} | \leq 1\% |
| (subset)    | target(test)             | \{1, 2, 3, 4, 5, ?\} | 11.27\% |
|             | auxiliary                | \{0, 1, ?\}   | 11.41\% |

### 2) EVALUATION METRICS

We adopt the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) as the evaluation metrics

\[
MAE = \sum_{(u, i, r_{ui}) \in R_E} |r_{ui} - \hat{r}_{ui}| / |R_E| \tag{16}
\]

\[
RMSE = \sqrt{\sum_{(u, i, r_{ui}) \in R_E} (r_{ui} - \hat{r}_{ui})^2 / |R_E|}, \tag{17}
\]

where $r_{ui}$ and $\hat{r}_{ui}$ denote the true and predicted rating, respectively. $|R_E|$ denotes the number of test ratings. We run 5 random trials when generating the required number of observed ratings from $R_T$, and the averaged results are reported.

### B. BASELINES AND PARAMETER SETTINGS

#### 1) BASELINES

We compare our SKT method with the following several related baseline algorithms:

- **WNMF**. Weight Nonnegative Matrix Factorization (WNMF) is a single-domain recommendation method. It decomposes the target rating matrix into the product of two low-dimensional non-negative matrices, which are then used to predict ratings and implement recommendations. The optimization objective of WNMF is shown in formula (1).

- **PMF** [8]. Probabilistic Matrix Factorization (PMF) is a Matrix Factorization-based method that learns only in the target domain.

---

\(^{10}\)https://grouplens.org/datasets/movielens/20m/.

\(^{11}\)www.junminghuang.com/datasets/goodreads.tar.gz.
- GWNMF [36]. Graph Regularized Weighted Nonnegative Matrix Factorization (GWNMF) incorporates constructed user and item graphs into a framework of nonnegative matrix factorization to take advantage of internal and external information. It only learns in the target domain.

- iCMF [24]. Collective Matrix Factorization (CMF) method extends Matrix Factorization MF) by jointly learning the latent factors of item from the user-item rating matrix of multiple domains. The iTCF [31] method is an extension of the CMF method, which introduces richer interactions by sharing both the latent features of item and predictability in two heterogeneous data in a smooth manner. Here, we set the interaction parameter $\rho$ between the user-specific latent features in the iTCF method to be 1 as an approximation of the CMF method. We name this method iCMF.

- Item-CST [19]. Coordinate system transfer (CST) is a transfer learning method in collaborative filtering to transfer the coordinate system from two auxiliary binary rating matrices to a target rating matrix in an adaptive way. Since our problem has only one auxiliary binary rating matrix that overlaps with the items in the target domain, CST can be adapted to our problem by only preserving item-side regularization term in our task, and we named it Item-CST.

- Item-TMF [21]. Transfer by Mixed Factorization (TMF) is the state-of-the-art method for cross-domain recommendation using binary preferences as auxiliary data, which introduces two interest profiles of users to model the user’s latent preference based on the iTCF [31] method. Similar to the iTCF method, the TMF method also requires that auxiliary and target domain share overlapping users and items. We can adapt TMF method to our problem by setting the interaction parameter $\rho$ between the user-specific latent feature to be 1. We name this method Item-TMF.

- WNMF-TL. Zhao et al. proposed an Active Transfer Learning for Cross-System Recommendation [13] method, which can perform cross-domain entity correspondence, and then plug the actively constructed entity correspondence into a general matrix factorization model. In our problem formulation, the cross-domain items are fully one-to-one correspondence, hence we removed the active learning module in the originally proposed method. Moreover, to make a fairer comparison with our SKT method, we use WNMF as the matrix factorization model, and then exploit the items similarity learned from the auxiliary binary rating data as a prior to constrain the items similarity in the target domain. We named this method WNMF-TL.

- KerKT. Kernel-induced Knowledge Transfer (kerKT) uses domain adaptation technology to adjust the feature distribution of overlapping entities between domains, and then uses diffusion kernel completion to correlate non-overlapping entities between domains. Since the items between the domains are completely overlapped in our problem, we adapt KerKT to scenarios where the items between the domains are completely overlapped.

Our SKT method jointly factorizes auxiliary binary rating matrix and target numerical rating matrix with the constraints of sharing the same latent factor of item and selectively sharing the latent factor of user. In addition, we integrate the co-graph regularization of user and item graphs into proposed weighted collective matrix factorization framework to avoid negative transfer.

2) PARAMETER SETTINGS

For WNMF, PMF, GWNMF, Item-CST, iCMF, Item-TMF, WNMF-TL, KerKT and our proposed SKT, different numbers of latent factors $d \in \{5, 10, 15, 20\}$ are tried. For PMF, different trade-off parameters $\lambda_U = \lambda_V \in \{0.01, 0.1, 1\}$ are tried. For GWNMF, different trade-off parameters $\lambda = \mu \in \{0.1, 1, 10, 100\}$ are tried. For Item-CST, we set the trade-off parameter $\rho_u = 0$, and different trade-off parameters $\rho_v \in \{0.01, 0.1, 1, 10, 100\}$ are tried. For the iTCF method, we approximate it by setting $\rho$ to 1 in the iTCF [31] method. At the same time, we fixed $\lambda = 1$, and different trade-off parameters $\alpha_u = \alpha_v \in \{0.01, 0.1, 1\}$, $\beta_u = \beta_v \in \{0.01, 0.1, 1\}$ are tried. For Item-TMF, we fixed the trade-off parameter $\lambda = 1$, $\rho = 1$, $\alpha_u = \beta_u = 0.01$, $\omega_u = 1$, and different trade-off parameters $\alpha_v = \beta_v \in \{0.01, 0.1, 1, 10, 100\}$, $\delta_p = \delta_q \in \{0.1, 1\}$, $\omega_p \in \{1, 2, 3, 4, 5\}$ are tried. For WNMF-TL, different regularization parameters $\lambda_c \in \{0.01, 0.1, 1, 10, 100\}$ are tried. For KerKT, we fixed $\alpha = 0.5$, different trade-off parameter $\lambda_u = \lambda_v \in \{0.001, 0.01\}$, $\delta_p = \{0.001, 0.01\}$ and $\lambda_c = \lambda_v \in \{0.0001, 0.001\}$ are tried. For SKT, we fixed $\lambda = 1$, different trade-off parameters $\gamma \in \{0.01, 0.1, 1, 10, 50, 100\}$, $\gamma_p \in \{0.01, 0.05, 0.1, 0.5, 1, 5, 10\}$, $\gamma_p \in \{0.01, 0.1, 1, 10, 50, 100\}$, $\alpha_v = \beta_u \in \{0.01, 0.05, 0.1, 0.5, 1\}$ and $\alpha_u = \alpha_v \in \{0.01, 0.05, 0.1, 0.5, 1, 5\}$ are tried. In our experiment, the parameter settings of SKT under different density levels of the Netflix-MovieLens dataset are listed in Table 6.

C. EXPERIMENTAL RESULTS

The experimental results on Netflix-MovieLens and Goodreads are shown in Table 4 and Table 5 respectively. From these results, we can make the following observations:

1) For non-transfer learning methods, the GWNMF method shows superior performance compared to the WNMF method at all density levels. In addition, when the target rating matrix is denser (e.g. $\geq 0.1$ for Netflix-MovieLens and $\geq 0.5$ for Goodreads),
TABLE 3. The time complexity of SKT and other baseline methods.

| SKT and other baseline methods | Time Complexity |
|-------------------------------|-----------------|
| WNMF                          | \(O(Kd|V|d)\)     |
| PMF                           | \(O(Kqd + \text{max}(|U|, |V|)d^2)\) |
| GWNMF                         | \(O(|U||V|d + |U|^2d + |V|^2d + |U|^2|V| + |V|^2|U|)\) |
| Item-CST                      | \(O(Kd^3 + Kd^2)\) |
| iCMF                          | \(O(Kq + q)\)     |
| Item-TMF                      | \(O(kd|P|N|)\)    |
| WNMF-TL                       | \(O(|U||V|d + |U|^2d + |V|^2d + |U|^2|V| + |V|^2|U|)\) |
| KerKT                         | \(O(K'|U||V|d + K(|U||V|d + |U|^2d + |V|^2d + |U|^2|V| + |V|^2|U| + |V|^3)\) |
| SKT                           | \(O(Kd|U||V|d + |U|^2d + |V|^2d + |U|^2d + |V|^2d + |U|^2|V| + |V|^2|U|)\) |

GWNMF tends to perform better than PMF, while PMF beats GWNMF when the target rating matrix becomes sparser (e.g. \(\leq 0.05\) for Netflix-MovieLens and \(\leq 0.1\) for Goodreads). The reason is that for the GWNMF method, when the target rating matrix is denser, the neighborhood structure information obtained is more accurate, and the extracted latent factors are more refined. At last, we can see that when the rating matrix in the target domain is very sparse, the non-transfer methods fail to give good recommendations.

2) The prediction performance of Item-CST is not always better than the non-transfer baseline methods at all tasks, especially when the target data becomes sparser, Item-CST performs worse than PMF and GWNMF. This indicates that when the target data is seriously sparse, Item-CST is prone to suffer from negative transfer, therefore it is unstable. A reasonable explanation is that when the target data becomes sparser, the divergence of data distribution between the two domains is greater. The latent factor extracted from auxiliary data is directly adapted to the target data, which will easily encounter negative transfer.

3) The iCMF method performs better than the non-transfer baseline methods and Item-CST at all density levels. Unlike the adaptive knowledge transfer method adopted by Item-CST, iCMF belongs to a collective knowledge transfer method, which is a bi-directed knowledge transfer method with richer interactions. iCMF extracts latent factors by joint matrix factorization, through which the data distribution divergence between the two domains is reduced. Therefore, iCMF can cope with the negative transfer issue better than Item-CST. However, when the target data is extremely sparse, the iCMF method has the limitation of insufficient knowledge transfer.

4) In addition to our proposed SKT method, the Item-TMF method performs best. Item-TMF incorporates virtual user profiles into prediction rules, which can model user preferences more accurately, thereby improving recommendation performance.

5) We can see that in all cases, WNMF-TL is significantly better than non-transfer learning method of WNMF, which shows that using the items similarity learned in the auxiliary data to constrain the items similarity in the target data can help improve the prediction performance of the model. In addition, WNMF-TL is better than non-transfer method of GWNMF when the density is lower (e.g. \(\leq 0.1\%\) for Netflix-MovieLens and Goodreads), while GWNMF beats WNMF-TL when the target rating matrix becomes denser. This is because when the target data becomes denser, the item similarity obtained from the original target rating data is more accurate than the item similarity learned from the auxiliary data.

6) KerKT performs best when the density is 0.1%, except for SKT. However, kerKT performs worse than other transfer learning methods when the density is lower than 0.1%. The possible reason is that KerKT encountered under transfer or negative transfer when the density is lower than 0.1%.

7) The proposed SKT can achieve significantly better prediction performance than all the other baseline methods in all cases. Especially when the target rating matrix is extremely sparse, SKT can achieve a greater performance improvement than other baseline methods. Unlike the other three cross-domain recommendation methods, the SKT method can selectively transfer users latent factors from auxiliary domain that do not overlap with users in the target domain. In addition, SKT integrates the intradomain entity similarity information from the target and auxiliary domains, through which positive transfer can be guaranteed.

D. PARAMETER ANALYSIS

In this section, we will describe the hyper-parameter tuning process in controlling their contributions. There are 9 hyper-parameters in the proposed SKT: \(\lambda\), \(d\), \(\gamma V\), \(\gamma U\), \(\gamma P\), \(\alpha U\), \(\alpha V\), \(\beta U\) and \(\beta V\). Using different trade-off parameters \(\lambda\), the performance of the SKT method is relatively stable. In addition, for the latent factor \(d\), setting it too large will
increase the time complexity of SKT, and setting it too small will result in too few latent factors of user that can be selected, which will reduce the performance of SKT. To simplify the problem, we fixed the parameter $\lambda = 1, d = 20$ and let $\alpha_U = \beta_U = \theta, \alpha_V = \beta_V = \delta$, focusing only on how the parameters $\gamma_V, \gamma_U, \gamma_p, \theta$ and $\delta$ affect the performance of SKT. For the sake of simplicity, we have only presented the result for the subset of Netflix-MovieLens. We use datasets with five density levels to test the five parameters. MAE and RMSE are used as evaluation metrics. Since MAE and RMSE are similar, only RMSE results are shown. The results are presented in Figure 2.
For the parameters $\gamma_V$ and $\gamma_p$, we search for the best parameter values by searching the grid \{0.01, 0.1, 1, 10, 50, 100\}. For the parameters $\gamma_U$, we search for the best parameter values by searching the grid \{0, 0.01, 0.05, 0.1, 0.5, 1, 5, 10\}. For the parameters $\theta$ and $\delta$, we search for the best parameter values by searching the grid \{0.01, 0.05, 0.1, 0.5, 1\}. The best parameter values of these five parameters at different density levels are listed in Table 6.

### Table 6. For the parameters $\gamma_U$, $\gamma_p$, $\gamma_V$, $\theta$ and $\delta$, their best parameter settings at different density levels.

| Dataset          | Parameters | Density levels |
|------------------|------------|----------------|
|                  | $\gamma_V$ | 0.01 0.05 0.1 0.5 1 |
| Netflix-MovieLens| $\gamma_U$ | 10 10 10 50 50 |
|                  | $\gamma_p$ | 1 1 1 1 1 |
|                  | $\theta$   | 1 1 1 0.1 0.1 |
|                  | $\delta$   | 1 0.1 0.1 0.01 0.01 |

To analyze the parameter $\gamma_V$, we set $\gamma_U$, $\gamma_p$, $\theta$ and $\delta$ to the best parameter values for different density levels, as shown in Table 6. From Figure 2(a), we can see that RMSE changes with different settings for $\gamma_V$. The parameter reflects the influence of sharing the latent factors of items across domain on matrix factorization. When the density level is lower than or equal to 0.1%, $\gamma_V$ does have a significant influence on RMSE, while when the density level is higher than or equal to 0.5%, $\gamma_V$ doesn’t influence significantly on RMSE. In our experiment, to achieve the best prediction performance, when the density level is equal to 0.01%, we set $\gamma_V = 100$; when the density level is higher than or equal to 0.05% and lower or equal to 0.1, we set $\gamma_V = 50$; and when the density level is higher than or equal to 0.5%, we set $\gamma_V = 1$. Similarly, to analyze the parameter $\gamma_U$, we fix the remaining four parameters, as shown in Table 6. The parameter $\gamma_U$ reflects the influence of selectively transferring the latent factors of users learned from auxiliary domain on matrix factorization. From Figure 2(b), we can also see that the more sparse the target data, the more significant the influence of $\gamma_U$ on RMSE. In our experiment, to obtain better prediction results, we set $\gamma_U = 0.1$ when the density is equal to 1%, while when the density is lower than or equal to 0.5%, we set $\gamma_U = 1$. We use the same method to analyze the parameters $\gamma_p$, $\theta$ and $\delta$. The parameter $\gamma_p$ reflects the influence of applying structured sparsity constraints to $P$ on matrix factorization. In Figure 2(c), we can see that when the density level is equal to 0.01%, the influence of $\gamma_p$ on RMSE is significant, and when the density level is greater than or equal to 0.05%, the influence of $\gamma_p$ on RMSE is not significant. In our experiment, when the density level is lower than or equal to 0.1%, we choose $\gamma_p = 10$, and when the density level is greater than or equal to 0.5%, we choose $\gamma_p = 50$. From Figure 2(d) or Figure 2(e), we can see that the parameters $\theta$ and $\delta$ have significant influence on RMSE. These two parameters reflect the influence of the similarity

FIGURE 2. Result of RMSE with different parameter settings on the subset of Netflix-MovieLens.
between the entities in the auxiliary and target domains on the matrix factorization. In our experiment, when the density level is equal to 0.01%, we set \( \theta = 1, \delta = 1 \); when the density level is higher than or equal to 0.05% and lower than or equal to 0.1%, we set \( \theta = 0.1, \delta = 0.1 \); and when the density level is higher than or equal to 0.5%, we set \( \theta = 0.1, \delta = 0.01 \).

V. CONCLUSION AND FUTURE WORK

In this article, we present a novel cross-domain recommendation method with heterogeneous feedbacks for knowledge transfer, called SKT. Specifically, SKT can not only directly transfer the latent features of the items from the auxiliary domain, which only shares overlapping items with the target domain, but also selectively transfer the latent preferences of users from the ones. Furthermore, to avoid negative transfer, we integrate the similarity between entities of intradomain from target and auxiliary domain into SKT. Experimental results show that the proposed SKT method achieves best performance compared to seven non-transfer learning and cross-domain recommendation methods.

For future work, we plan to extend our proposed method to scenarios where there are only a few cross-domain entity correspondences or no cross-domain entity correspondences. In addition, there are still some interesting problems to be explored. For example, how does the sparsity of auxiliary data influence the prediction performance? And when there are several auxiliary domains available, how to choose the best auxiliary domain?

REFERENCES

[1] H. Shen, D. Wang, C. Song, and A.-L. Barabási, “Modeling and predicting popularity dynamics via reinforced Poisson processes,” in Proc. 28th AAAI Conf. Artif. Intell., Quebec City, QC, Canada, 2014, pp. 291–297.
[2] J. Lu, D. Wu, M. Mao, W. Wang, and G. Zhang, “Recommender system application developments: A survey,” Decis. Support Syst., vol. 74, pp. 12–32, Jun. 2015.
[3] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutierrez, “Recommender systems survey,” Knowl.-Based Syst., vol. 46, pp. 109–132, Jul. 2013.
[4] R. Yera and L. Martinez, “Fuzzy tools in recommender systems: A survey,” Int. J. Comput. Intell. Syst., vol. 10, no. 1, pp. 776–803, 2017.
[5] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl, “Item-based collaborative filtering recommendation algorithms,” in Proc. 10th Int. Conf. World Wide Web (WWW), Hong Kong, vol. 1, May 2001, pp. 285–295.
[6] M. Zhang, X. Guo, and G. Chen, “Prediction uncertainty in collaborative filtering: Enhancing personalized online product ranking,” Decis. Support Syst., vol. 83, pp. 10–21, Mar. 2016.
[7] Y. Koren, R. Bell, and C. Volinsky, “Matrix factorization techniques for recommender systems,” Computer, vol. 42, no. 8, pp. 30–37, Aug. 2009.
[8] R. Salakhutdinov and A. Mnih, “Probabilistic matrix factorization,” in Proc. 20th Ann. Conf. Neural Inf. Process. Syst., Vancouver, BC, Canada, Dec. 2008, pp. 1257–1264.
[9] G. Chen, F. Wang, and C. Zhang, “Collaborative filtering using orthogonal nonnegative matrix tri-factorization,” Inf. Process. Manage., vol. 45, no. 3, pp. 368–379, May 2009.
[10] A. Hernando, J. Bobadilla, and F. Ortega, “A non negative matrix factorization for collaborative filtering recommender systems based on a Bayesian probabilistic model,” Knowl.-Based Syst., vol. 97, pp. 188–202, Apr. 2016.
[11] C. Gao, X. Chen, F. Feng, K. Zhao, X. He, Y. Li, and D. Jin, “Cross-domain recommendation without sharing user-relevant data,” in Proc. World Wide Web Conf. (WWW), San Francisco, CA, USA, 2019, pp. 491–502.
[12] D. Yang, J. He, H. Qin, Y. Xiao, and W. Wang, “A graph-based recommendation across heterogeneous domains,” in Proc. 24th ACM Int. Conf. Inf. Knowl. Manage., Melbourne, VIC, Australia, Oct. 2015, pp. 463–472.
[36] Q. Gu, J. Zhou, and C. Ding, “Collaborative filtering: Weighted non-negative matrix factorization incorporating user and item graphs,” in Proc. SIAM Int. Conf. Data Mining, Columbus, OH, USA, Apr. 2010, pp. 199–210.

[37] R. He, W.-S. Zheng, and B.-G. Hu, “Maximum correntropy criterion for robust face recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 8, pp. 1561–1576, Aug. 2011.

[38] K. Wang, R. He, L. Wang, W. Wang, and T. Tan, “Joint feature selection and subspace learning for cross-modal retrieval,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 10, pp. 2010–2023, Oct. 2016.

[39] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in Proc. 15th Adv. Neural Inf. Process. Syst., Vancouver, BC, Canada, Dec. 2001, pp. 556–562.

[40] D. Cai, X. He, J. Han, and T. S. Huang, “Graph regularized nonnegative matrix factorization for data representation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 8, pp. 1548–1560, Aug. 2011.

[41] V. Sindhwani, S. Bucak, J. Hu, and A. Mojsilovic, “A family of non-negative matrix factorizations for one-class collaborative filtering problems,” in Proc. 3rd ACM Conf. Rec. Syst. (RecSys), New York, NY, USA, Oct. 2009, pp. 1–8.

[42] J. Huang, X.-Q. Cheng, H.-W. Shen, T. Zhou, and X. Jin, “Exploring social influence via posterior effect of word-of-mouth recommendations,” in Proc. 5th ACM Int. Conf. Web Search Data Mining (WSDM), Seattle, WA, USA, 2012, pp. 573–582.

HONGWEI ZHANG received the master’s degree from Liaoning Normal University, China, in 2010. He is currently pursuing the Ph.D. degree with the School of Information and Communication Engineering, Dalian University of Technology, China. Since 2010, he has been a Lecturer with the School of Mathematics, Tonghua Normal University, China. His research interests include transfer learning, recommender systems, and machine learning.

XIANGWEI KONG (Member, IEEE) received the Ph.D. degree in management science and engineering from the Dalian University of Technology, China, in 2003. From 2006 to 2007, she was a Visiting Researcher with the Department of Computer Science, Purdue University, USA. From 2014 to 2015, she was a Senior Research Scientist with the Department of Computer Science, New York University, USA. She is currently a Professor with the School of Management, Zhejiang University, China. Her research interests include digital image processing and recognition, multimedia information security, digital media forensics, image retrieval and mining, multisource information fusion, knowledge management, and business intelligence.

YUJIA ZHANG is currently pursuing the master’s degree in computer and information technology with the School of Engineering and Applied Science, University of Pennsylvania. Her research interests include recommender systems, business analytics, and artificial intelligence.

***