Tags as Bridges between Domains: Improving Recommendation with Tag-Induced Cross-Domain Collaborative Filtering

Yue Shi, Martha Larson, and Alan Hanjalic

Multimedia Information Retrieval Lab, Delft University of Technology,
Mekelweg 4, 2628CD Delft, Netherlands
{y.shi,m.a.larson,a.hanjalic}@tudelft.nl

Abstract. Recommender systems generally face the challenge of making predictions using only the relatively few user ratings available for a given domain. Cross-domain collaborative filtering (CF) aims to alleviate the effects of this data sparseness by transferring knowledge from other domains. We propose a novel algorithm, Tag-induced Cross-Domain Collaborative Filtering (TagCDCF), which exploits user-contributed tags that are common to multiple domains in order to establish the cross-domain links necessary for successful cross-domain CF. TagCDCF extends the state-of-the-art matrix factorization by introducing a constraint involving tag-based similarities between pairs of users and pairs of items across domains. The method requires no common users or items across domains. Using two publicly available CF data sets as different domains, we experimentally demonstrate that TagCDCF substantially outperforms other state-of-the-art single domain CF and cross-domain CF approaches. Additional experiments show that TagCDCF addresses data sparseness and illustrate the influence of the number of tags used by users in both domains.

Keywords: Collaborative filtering, cross domain collaborative filtering, matrix factorization, tag, recommender systems.

1 Introduction

Collaborative filtering (CF) is one of the most successful techniques in recommender systems [1]. The idea of CF is to make use of the user-item rating matrix to predict items that individual users may like in future [4]. However, users usually rate a very limited number of items, giving rise to the widely-known data sparseness problem, which characterizes most recommender system tasks. Specifically to address the data sparseness problem, recent research has started to investigate cross-domain CF [6][7][9][16], which makes use of rating data from other domains to benefit a target domain. The key challenge in cross-domain CF is to discover linkage among domains (e.g., shared knowledge or common characteristics) that allows different domains to benefit each other effectively. Typically, domains are mutually exclusive, each involving a certain type of product (e.g., movies, music or books) and a set of users whose identities or identifiers are largely unique to the domain. As a result, it is difficult to directly extract common characteristics among users and items from different

Joseph A. Konstan et al. (Eds.): UMAP 2011, LNCS 6787, pp. 305–316, 2011.
© Springer-Verlag Berlin Heidelberg 2011

domains. Here, we turn to a novel source of information to link domains: user-generated tags. Many of today's recommender systems address tasks involving users who not only rate items, but also annotate items with tags denoting characteristics of items or their own personal preferences. Our approach is based on the insight that different users in different domains may use the same tags to describe items or express their opinions about items. We expect domains that share high-level characteristics, such as notions of plot and genre, to exhibit particularly useful overlap with respect to user-deployed tags. For example, tags such as "sci-fi", "fast-paced" and "romance" can be used by users to annotate items in both a movie domain and a book domain. Users with similar patterns of usage and preference can be assumed to have similar tagging patterns, which we anticipate to hold across domains. In sum, it is potentially beneficial to exploit tags shared between domains in order to transfer knowledge between those domains for the purpose of recommendation.

In this paper, we propose a novel tag-induced cross-domain collaborative filtering (TagCDCF) algorithm that exploits shared tags to link different domains, alleviating data sparseness in each individual domain and improving recommendation performance. TagCDCF uses an explicit encoding of the relationships between domains in the form of user-to-user and item-to-item similarity matrices based on tags shared between the domains to be linked. We formulate the TagCDCF from a probabilistic point of view, leading to an extended matrix factorization framework. The user-item matrices of different domains are factorized into domain-specific latent user features and latent item features. The latent features are linked across domains using the explicit tag-induced cross-domain similarities.

The contribution of this work is twofold: First, we present a novel cross-domain CF algorithm that is able to exploit explicit user-to-user and item-to-item similarities. Second, we show that using tags to bridge domains is able to address data sparseness and outperform state-of-the-art CF and cross-domain CF approaches.

The remainder of the paper is structured as follows. In the next section, we summarize related work and position our work in its context. We present the proposed TagCDCF algorithm in detail in Section 3, after which we evaluate its performance through a series of experiments. The last section sums up the key aspects of the proposed algorithm and briefly discusses directions for future work.

2 Related Work

Collaborative filtering. Collaborative filtering approaches can generally be categorized as memory-based or model-based. Memory-based approaches first compute similarities among users or among items based on a given user-item rating matrix, and then use these similarities to recommend items to a given user, either in terms of preferences of like-minded users (i.e., user-based CF (UBCF) [4]) or in terms of similarity to already rated items (i.e., item-based CF [11]). Model-based approaches first learn a prediction model based on a training set from the user-item rating matrix, and then apply this model to predict unknown preferences for users on items. Matrix factorization (MF) techniques [5] have become one of the most popular model-based CF approaches, due to the advantages of accuracy and scalability. Generally, MF techniques learn latent features of users and items from the observed ratings in a user-item matrix. The learned features are then further used to predict unobserved ratings.

Collective matrix factorization [13] is proposed to factorize multiple matrices representing the same domain that share common latent user or item features. Probabilistic matrix factorization (PMF) [10] approaches factorization in a single domain from a probabilistic point of view. TagCDCF builds on the PMF concept, but goes beyond existing CF approaches by tackling the cross-domain CF problem rather than using only a single domain.

Exploiting Tags for CF. Recently, researchers in recommender systems community have started investigating the usefulness of user-generated tags in improving recommendation quality. Tags have been exploited to enhance item recommendations by several means, e.g., via tensor factorization for user-tag-item triplet data [14], via similarities between users or items in terms of associated tags [12][17], and via representing users and items with weighted tags that have been de-noised [8]. However, our work goes beyond the scope of the aforementioned works, since we make use of tags to bridge different domains in order to enable knowledge transfer from one domain to another.

Cross-Domain CF. Finally, we note that several cross-domain CF approaches have been proposed recently. Coordinate system transfer [9] is proposed to adapt learned latent features of users and items from auxiliary domains and use them to regularize the learning of latent features of users and items in the target domain. However, it requires that either users or items are shared between the domains, which is, as already mentioned above, a condition not commonly encountered in practical applications. Codebook transfer (CBT) [6] and the rating-matrix generative model [7] both learn an implicit cluster-level rating pattern that could be shared between different domains. The learned implicit rating pattern is then used to transfer knowledge between domains to alleviate data sparseness. Similarly, multi-domain CF is proposed to extend probabilistic matrix factorization in multiple domains together with learning an implicit correlation matrix [16], which is assumed to link different domains for knowledge transfer. Compared to all the aforementioned cross-domain CF approaches, the proposed TagCDCF is substantially different in that we exploit explicit common characteristics, i.e., common tags, between different domains for knowledge transfer, rather than learning implicit cross-domain relationships. It is expected that the explicit common characteristics could lead to a more reliable and effective cross-domain CF than the implicit ones.

3 Tag-Induced Cross-Domain Collaborative Filtering

We present Tag-induced cross-domain collaborative filtering (TagCDCF), first introducing cross-domain user-to-user similarity and item-to-item similarity and then presenting the central mechanism of TagCDCF, a matrix factorization approach that incorporates explicit cross-domain similarities to bridge different domains. Matrix factorization uses known data to estimate latent features representing users and items and is the key objective that determines the recommendation accuracy. The integration of tag-induced cross-domain similarities guides the factorization process and leads to improved recommendation performance. Although TagCDCF could be used to incorporate multiple domains, we concentrate in this paper on combining two domains.

3.1 Definition of Tag-Induced Cross-Domain Similarity

TagCDCF makes use of two types of tag-induced cross-domain similarities, item-to-item similarity and user-to-user similarity, which are defined over the tags that are shared between the domains to be combined. As mentioned before, different domains, e.g., a movie recommender system and a book recommender system may have completely different users and items. However, it is still possible that some users in different domains use the same tags to annotate items of interest, and that some items in different domains are tagged by same tags that encode their similar properties. For this reason, we can assume cross-domain user-to-user similarity in terms of common tags the users apply, and cross-domain item-to-item similarity in terms of that items are annotated with.

In the kth domain, we denote the set of users by M_k and the set of items by N_k. The totality of tags used by the users to tag the items is the tag set, $T^{(k)}$, consisting of L_k different tags. The user-tag indicator matrix in the kth domain, $\mathbf{A^{(k)}}$, is an $M_k \times L_k$ matrix, in which $A_{il}^{(k)} = 1$ if user i used tag l, and is otherwise 0. Similarly, the item-tag indicator matrix in the kth domain, $\mathbf{B^{(k)}}$, is a $N_k \times L_k$ matrix, in which $B_{jl}^{(k)} = 1$ if item j is tagged by tag l, an is otherwise 0.

In order to compute the tag-induced cross-domain similarities, we first extract a shared tag set CT that contains all tags in both domains, i.e., $CT=T^{(1)} \cap T^{(2)}$. Then, we define the cross-domain user-to-user similarity $S_{ip}^{(U)}$ (i.e., between user i in the first domain and user p in the second domain) and cross-domain item-to-item similarity $S_{jq}^{(V)}$ (i.e., between item j in the first domain and item q in the second domain) using the cosine similarity measure, as shown below:

$$S_{ip}^{(U)} = \frac{\sum_{t \in CT} \left(A_{ix(t)}^{(1)} A_{py(t)}^{(2)} \right)}{\sqrt{\sum_{t \in CT} \left(A_{ix(t)}^{(1)} \right)^2} \sqrt{\sum_{t \in CT} \left(A_{py(t)}^{(2)} \right)^2}} \quad , \quad S_{jq}^{(V)} = \frac{\sum_{t \in CT} \left(B_{jx(t)}^{(1)} B_{qy(t)}^{(2)} \right)}{\sqrt{\sum_{t \in CT} \left(B_{jx(t)}^{(1)} \right)^2} \sqrt{\sum_{t \in CT} \left(B_{qy(t)}^{(2)} \right)^2}} \tag{1}$$

Note that we use $x(t)$ to denote the index of tag t in the first domain, and $y(t)$ the index of tag t in the second domain.

3.2 Formulation of Tag-Induced Cross-Domain Collaborative Filtering

We denote user-item matrix in kth domain as $\mathbf{R^{(k)}}$, which is an $M_k \times N_k$ matrix containing ratings from M_k users on N_k items. Ratings in each domain are normalized to be within the range from 0 to 1. We adopt the convention of denoting the non-zero entries in a matrix $\mathbf{X}$ as $|\mathbf{X}|$. The latent user features in the kth domain are collected in $\mathbf{U^{(k)}}$, a $d \times M_k$ matrix, whose ith column indicates the d-dimensional latent feature vector for user i. Similarly, the latent item features in the kth domain are represented by $\mathbf{V^{(k)}}$. This is a $d \times N_k$ matrix**,** whose jth column indicates the d-dimensional latent feature vector for item j.

We first present TagCDCF in a graphical model that illustrates relationships among all variables, as shown in Fig.1. The latent features of users and items, i.e., $\mathbf{U^{(1)}}$,$\mathbf{V^{(1)}}$,$\mathbf{U^{(2)}}$,$\mathbf{V^{(2)}}$, are unknown variables that need to be estimated. As can be seen, the sub-graph that only involves $\mathbf{U^{(k)}}$,$\mathbf{V^{(k)}}$, $\mathbf{R^{(k)}}$ (k =1 or 2) is equivalent to PMF [10] in a single domain. The tag-induced cross-domain similarities $\mathbf{S^{(U)}}$ and $\mathbf{S^{(V)}}$ actually bring the two domains together, which can be seen as a key innovation in this paper.

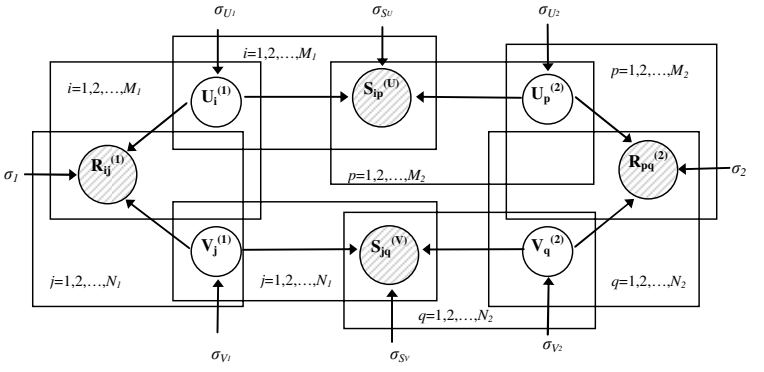


Fig. 1. Graphical model for the proposed TagCDCF

We can further interpret the graphical model from a probabilistic point of view. According to the graphical model theory [2], the joint distribution of all variables in the graph can be expressed as:

$$\begin{aligned} & p\left(U^{(1)},V^{(1)},U^{(2)},V^{(2)},R^{(1)},R^{(2)},S^{(U)},S^{(V)},\sigma_1,\sigma_2,\sigma_{U_1},\sigma_{V_1},\sigma_{U_2},\sigma_{V_2},\sigma_{S_U},\sigma_{S_V}\right) \\ &= p\left(R^{(1)} \mid U^{(1)},V^{(1)},\sigma_1\right) p\left(R^{(2)} \mid U^{(2)},V^{(2)},\sigma_2\right) p\left(S^{(U)} \mid U^{(1)},U^{(2)},\sigma_{S_U}\right) \\ & \quad p\left(S^{(V)} \mid V^{(1)},V^{(2)},\sigma_{S_V}\right) p\left(U^{(1)} \mid \sigma_{U_1}\right) p\left(V^{(1)} \mid \sigma_{V_1}\right) p\left(U^{(2)} \mid \sigma_{U_2}\right) p\left(V^{(2)} \mid \sigma_{V_2}\right) \\ & \quad p(\sigma_1)\, p(\sigma_2)\, p\left(\sigma_{U_1}\right) p\left(\sigma_{V_1}\right) p\left(\sigma_{U_2}\right) p\left(\sigma_{V_2}\right) p\left(\sigma_{S_U}\right) p\left(\sigma_{S_V}\right) \end{aligned} \tag{2}$$

After applying the product rule to the left side of Eq. (2), and neglecting the influence of constant prior probabilities, we have:

$$\begin{aligned} & p\left(U^{(1)},V^{(1)},U^{(2)},V^{(2)} \mid R^{(1)},R^{(2)},S^{(U)},S^{(V)},\sigma_1,\sigma_2,\sigma_{U_1},\sigma_{V_1},\sigma_{U_2},\sigma_{V_2},\sigma_{S_U},\sigma_{S_V}\right) \\ & \propto p\left(S^{(U)} \mid U^{(1)},U^{(2)},\sigma_{S_U}\right) p\left(S^{(V)} \mid V^{(1)},V^{(2)},\sigma_{S_V}\right) \prod_{k=1}^{2} p\left(R^{(k)} \mid U^{(k)},V^{(k)},\sigma_k\right) p\left(U^{(k)} \mid \sigma_{U_k}\right) p\left(V^{(k)} \mid \sigma_{V_k}\right) \end{aligned} \tag{3}$$

The conditional distribution over observed ratings in each domain can be defined as:

$$p\left(R^{(k)} \mid U^{(k)},V^{(k)},\sigma_k\right) = \prod_{i=1}^{M_k} \prod_{j=1}^{N_k} \left[\mathcal{N}\left(R_{ij}^{(k)} \mid U_i^{(k)T} V_j^{(k)},\sigma_k^2\right)\right]^{I_{ij}^{R^{(k)}}} \tag{4}$$

where $\mathcal{N}\left(x \mid \mu,\sigma^2\right)$ denotes the probability density function for a Gaussian distribution with mean μ and variance σ^2. For a matrix $\mathbf{X}$, the indicator function I_{ij}^{X} is equal to 1 if X_{ij}>0, and is otherwise 0.

We also define conditional distributions over observed cross-domain similarities as:

$$\begin{aligned} p\left(S^{(U)} \mid U^{(1)},U^{(2)},\sigma_{S_U}\right) &= \prod_{i=1}^{M_1} \prod_{p=1}^{M_2} \left[\mathcal{N}\left(S_{ip}^{(U)} \mid U_i^{(1)T} U_p^{(2)},\sigma_{S_U}^2\right)\right]^{I_{ip}^{S^{(U)}}}, \\ p\left(S^{(V)} \mid V^{(1)},V^{(2)},\sigma_{S_V}\right) &= \prod_{j=1}^{N_1} \prod_{q=1}^{N_2} \left[\mathcal{N}\left(S_{jq}^{(V)} \mid V_j^{(1)T} V_q^{(2)},\sigma_{S_V}^2\right)\right]^{I_{jq}^{S^{(V)}}} \end{aligned} \tag{5}$$

Finally, we use the zero-mean spherical Gaussian priors [15] to represent a latent user and movie features in each domain:

$$p\left(U^{(k)} \mid \sigma_{U_k}\right) = \prod_{i=1}^{M_k} \mathcal{N}\left(U_i^{(k)} \mid 0, \sigma_{U_k}^2 \mathrm{I}\right) \quad , \quad p\left(V^{(k)} \mid \sigma_{V_k}\right) = \prod_{j=1}^{N_k} \mathcal{N}\left(V_j^{(k)} \mid 0, \sigma_{V_k}^2 \mathrm{I}\right) \tag{6}$$

We substitute Eq. (4-6) into Eq. (3) and estimate latent user and item features in Eq. (3) by maximizing the posterior, which is equivalent to minimizing the negative log-posterior as shown below:

$$\begin{aligned} & -\ln p\left(U^{(1)}, V^{(1)}, U^{(2)}, V^{(2)} \mid R^{(1)}, R^{(2)}, S^{(U)}, S^{(V)}, \sigma_1, \sigma_2, \sigma_{U_1}, \sigma_{V_1}, \sigma_{U_2}, \sigma_{V_2}, \sigma_{S_U}, \sigma_{S_V}\right) \\ & = \sum_{k=1}^{2}\left(\frac{1}{2\sigma_k^2}\sum_{i=1}^{M_k}\sum_{j=1}^{N_k} I_{ij}^{R^{(k)}}\left(R_{ij}^{(k)} - U_i^{(k)T}V_j^{(k)}\right)^2\right) + \frac{1}{2\sigma_{S_V}^2}\sum_{j=1}^{N_1}\sum_{q=1}^{N_2} I_{jq}^{S^{(V)}}\left(S_{jq}^{(V)} - V_j^{(1)T}V_q^{(2)}\right)^2 \\ & + \frac{1}{2\sigma_{S_U}^2}\sum_{i=1}^{M_1}\sum_{p=1}^{M_2} I_{ip}^{S^{(U)}}\left(S_{ip}^{(U)} - U_i^{(1)T}U_p^{(2)}\right)^2 + \sum_{k=1}^{2}\left(\frac{1}{2\sigma_{U_k}^2}\sum_{i=1}^{M_k} U_i^{(k)T}U_i^{(k)} + \frac{1}{2\sigma_{V_k}^2}\sum_{j=1}^{N_k} V_j^{(k)T}V_j^{(k)}\right) + C \end{aligned} \tag{7}$$

Note that C is a term containing rating variances, similarity variances and prior variances, which are independent of latent features. In order to simplify the model, we can assume that the variance of the user preferences in the two domains are comparable and that the user ratings (which have been normalized to the same scales in both domains) can be assumed to have equal variance, i.e., $\sigma_1^2 = \sigma_2^2$. We also assume that prior variances are the same across all the latent features, i.e., $\sigma_{U_1}^2 = \sigma_{V_1}^2 = \sigma_{U_2}^2 = \sigma_{V_2}^2$. Therefore, we can define the objective function $F(\mathbf{U^{(1)}}, \mathbf{V^{(1)}}, \mathbf{U^{(2)}}, \mathbf{V^{(2)}})$ as:

$$\begin{aligned} & F\left(U^{(1)}, V^{(1)}, U^{(2)}, V^{(2)}\right) \\ & = \frac{1}{2}\sum_{k=1}^{2}\sum_{i=1}^{M_k}\sum_{j=1}^{N_k} I_{ij}^{R^{(k)}}\left(R_{ij}^{(k)} - U_i^{(k)T}V_j^{(k)}\right)^2 + \frac{\alpha}{2}\sum_{j=1}^{N_1}\sum_{q=1}^{N_2} I_{jq}^{S^{(V)}}\left(S_{jq}^{(V)} - V_j^{(1)T}V_q^{(2)}\right)^2 \\ & + \frac{\beta}{2}\sum_{i=1}^{M_1}\sum_{p=1}^{M_2} I_{ip}^{S^{(U)}}\left(S_{ip}^{(U)} - U_i^{(1)T}U_p^{(2)}\right)^2 + \frac{\lambda}{2}\sum_{k=1}^{2}\left(\left\|U^{(k)}\right\|_{Fro}^2 + \left\|V^{(k)}\right\|_{Fro}^2\right) \end{aligned} \tag{8}$$

in which we have $\alpha = \sigma_1^2 / \sigma_{S_V}^2$, $\beta = \sigma_1^2 / \sigma_{S_U}^2$, and $\lambda = \sigma_1^2 / \sigma_{U_1}^2$. α and β are tradeoff parameters that control the influence of cross-domain item-to-item similarity and user-to-user similarity, respectively. λ is a regularization parameter that is usually used to penalize the complexity of latent feature matrices in order to alleviate over-fitting.

A local minimum solution for minimizing the objective function in Eq. (8) can be achieved by gradient descent with alternatively fixed $\mathbf{U^{(1)}}$, $\mathbf{V^{(1)}}$, $\mathbf{U^{(2)}}$, and $\mathbf{V^{(2)}}$. The gradients can be computed as below:

$$\begin{aligned} \frac{\partial F}{\partial U_i^{(1)}} &= \sum_{j=1}^{N_1} I_{ij}^{R^{(1)}}\left(U_i^{(1)T}V_j^{(1)} - R_{ij}^{(1)}\right)V_j^{(1)} + \beta\sum_{p=1}^{M_2} I_{ip}^{S^{(U)}}\left(U_i^{(1)T}U_p^{(2)} - S_{ip}^{(U)}\right)U_p^{(2)} + \lambda U_i^{(1)}, \\ \frac{\partial F}{\partial V_j^{(1)}} &= \sum_{i=1}^{M_1} I_{ij}^{R^{(1)}}\left(U_i^{(1)T}V_j^{(1)} - R_{ij}^{(1)}\right)U_i^{(1)} + \alpha\sum_{q=1}^{N_2} I_{jq}^{S^{(V)}}\left(V_j^{(1)T}V_q^{(2)} - S_{jq}^{(V)}\right)V_q^{(2)} + \lambda V_j^{(1)}, \\ \frac{\partial F}{\partial U_i^{(2)}} &= \sum_{j=1}^{N_2} I_{ij}^{R^{(2)}}\left(U_i^{(2)T}V_j^{(2)} - R_{ij}^{(2)}\right)V_j^{(2)} + \beta\sum_{p=1}^{M_1} I_{pi}^{S^{(U)}}\left(U_p^{(1)T}U_i^{(2)} - S_{pi}^{(U)}\right)U_p^{(1)} + \lambda U_i^{(2)}, \\ \frac{\partial F}{\partial V_j^{(2)}} &= \sum_{i=1}^{M_2} I_{ij}^{R^{(2)}}\left(U_i^{(2)T}V_j^{(2)} - R_{ij}^{(2)}\right)U_i^{(2)} + \alpha\sum_{q=1}^{N_1} I_{qj}^{S^{(V)}}\left(V_q^{(1)T}V_j^{(2)} - S_{qj}^{(V)}\right)V_q^{(1)} + \lambda V_j^{(2)} \end{aligned} \tag{9}$$

The learned latent features of users and items can be multiplied to compute the recommendation, i.e., predict unobserved ratings in each domain. By exploiting data sparseness, the complexity of TagCDCF is $O(d(|\mathbf{R}^{(1)}|+|\mathbf{R}^{(2)}|+|\mathbf{S}^{(U)}|+|\mathbf{S}^{(V)}|))$, which is linear with the total number of non-zeros in the user-item rating matrices and cross-domain similarity matrices. Such reasonable complexity reflects the ability of TagCDCF to scale up to use cases involving large sets of rating data.

4 Experimental Evaluation

We carry out a series of experiments to demonstrate that TagCDCF improves recommendation for both of the combined domains over state-of-the-art single domain CF and cross-domain CF approaches. Further experiments examine the ability of TagCDCF to address the data sparseness problem as faced in the case of users who have rated relatively few items. A final set of experiments, explores the dependence of TagCDCF performance on the number of tags shared between domains.

4.1 Experimental Framework

Data sets. We evaluate the proposed TagCDCF algorithm on two different domains represented by two publicly available data sets: the MovieLens data set (http://www.grouplens.org/node/73) [4], with ca. 10 million ratings, and the LibraryThing data set (http://homepage.tudelft.nl/5q88p/LT) [3], with ca. 750 thousand ratings. Both of the data sets have 5-star rating scale, with half star increments. In addition, the MovieLens data set contains 16529 unique tags with 95580 tag assignments from users to movies, and the LibraryThing data set contains 10559 unique tags with ca. 2 million tag assignments from users to books. There are in total 2277 tags common to the two domains.

Our experiments are conducted on the first 5000 users chosen according to the identifiers in the original data sets from the 71567 MovieLens users and the 7279 LibraryThing users and on the first 5000 items from the 10681 MovieLens movies and the 37232 LibraryThing books. This selection of subsets was necessary in order to implement a full-range of baselines for comparison, including the computationally expensive UBCF and CBT. We avoided random selection to facilitate future reproducibility. The experimental subsets are denoted here as ML (from MovieLens) and LT (from LibraryThing). We note that their size is comparable to that of the largest data set used to date for cross-domain CF [9]. The rating data sparseness is 97.7% in ML, and 99.3% in LT.

Experimental Protocol. For each data set, we generate a data partition by randomly selecting 80% ratings as training set, and using the remaining 20% ratings as the test set. In this way, for each data set we generated six data partitions, one of which is randomly selected for validation, i.e., for tuning the parameters, and the other five for testing, i.e., for reporting the performance of the proposed algorithm and comparing it with other approaches. We set the dimensionality d of latent features to 10 for the TagCDCF. Experimental investigation revealed that the performance did not substantially change when further increasing d, due to which 10 is a good choice in terms of model complexity. The regularization parameter is tuned to 0.01, which is the same value as used for the baseline approach PMF.

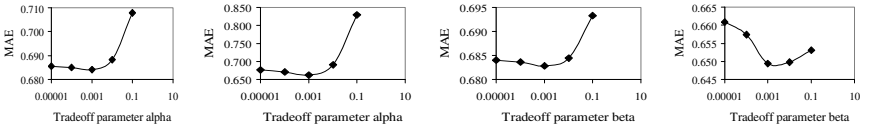
Evaluation Metric. To be consistent with recent studies [6][7][9][16] on this topic, we also use mean absolute error (MAE) as the evaluation metric, defined as:

$$MAE = \sum_{(i,j)\in Ts} \left| \hat{R}_{ij} - R_{ij} \right| \Big/ |Ts| \tag{10}$$

where Ts denotes the set of user-item pairs whose ratings need to be predicted in the test set. We denote by $|Ts|$ the size of the set Ts. $\hat{R}_{ij}$ denotes the predicted rating for user i on item j. Note that the lower MAE means better recommendation performance.

4.2 Experimental Results

Impact of Tradeoff Parameters. Our first experiment investigates the impact of tradeoff parameters in the proposed TagCDCF. This experiment is conducted on the validation partition in each domain. We first set the tradeoff parameter β=0 and investigate the impact of the tradeoff parameter α. The change in MAE caused by varying the value of α in each data set is shown in Fig.2(a) and (b). The influence of α on MAE confirms that exploiting the cross-domain item-to-item similarity could introduce performance gain in both domains. Then, we fix the tradeoff parameter α with the optimal value as 0.001, and investigate the impact of β, shown in Fig.2(c) and (d). The influence of β (with the optimal value as 0.001) confirms that additional improvement can be achieved in both domains by introducing the cross-domain user-to-user similarity. Summarizing, we find both domains stand to benefit by exploitation of tag-induced cross-domain similarities via TagCDCF.



(a) (b) (c) (d)

Fig. 2. Impact of tradeoff parameters on TagCDCF: (a) α on ML, (b) α on LT, (c) β on ML, and (d) β on LT

Performance Comparison. Next, we compare the performance of the TagCDCF with other baseline approaches, using the five test partitions in each data set. Note that the tradeoff parameters used are the optimal ones, as tuned using the validation partition. The baseline approaches are listed below:

UBCF: User-based collaborative filtering [4] is used as a representative of a conventional memory-based CF approach. The neighborhood size is set to 50, determined to be the best-performing neighborhood size using the validation partition.

PMF: Probabilistic matrix factorization [10] is used as a representative of a state-of-the art model-based CF approach. In a single domain, PMF is equivalent to the TagCDCF when both tradeoff parameters are set to 0. The regularization parameter is tuned to 0.01, which achieved optimal performance on the validation partition. Note that both UBCF and PMF are CF approaches that use only a single domain.

CBT: Codebook transfer [6] is used to represent a state-of-the-art cross-domain CF approach. Here, one domain is the auxiliary domain, which is used to construct the codebook, and the other is the target domain in which the predictions are carried out. We use ML as the source of the auxiliary domain for LT predictions and LT as the source of the auxiliary domain for ML predictions. In each case, we follow the protocol used in [6] and select the 500 users and the 500 items with most ratings to constitute the auxiliary domain, while setting the number of user and item clusters to 50.

The performance of the TagCDCF and the baseline approaches are shown in Table 1. Note that the MAE is averaged across all the five test partitions.

Table 1. Performance comparison between TagCDCF and baseline approaches (MAE ± std.)

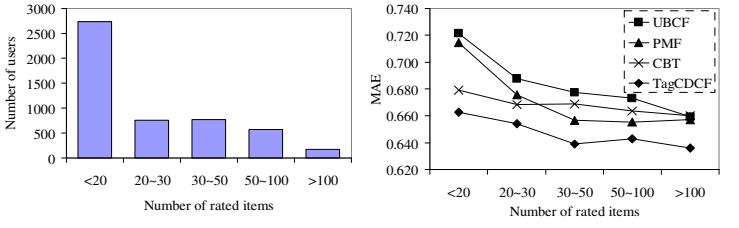
Data set	UBCF	PMF	CBT	TagCDCF
ML	0.691 ±0.002	0.686 ±0.001	0.688 ±0.002	**0.684 ±0.001**
LT	0.682 ±0.002	0.677 ±0.004	0.671 ±0.003	**0.653 ±0.004**

The results demonstrate that TagCDCF significantly outperforms other approaches in both data sets—improvements in Table 1 are statistically significant according to the Wilcoxon signed rank significance test with $p<0.01$. TagCDCF achieved an improvement over CBT on both data sets, indicating that the explicit tag-induced relationships between two domains could be more effective than the hypothesized implicit relationships solely learned from rating data. We notice that the improvement achieved by TagCDCF on the LT data set is substantial, i.e., up to 4.4%, and is larger than the improvement on the ML data set, i.e., up to 1% improvement. The difference reflects the smaller number of ratings (i.e., the higher sparseness) mentioned in Section 4.1 in the LT data set, meaning that there is a greater potential for TagCDCF to introduce improvement. Note that the performance of CBT is even worse than PMF on the ML data set, meaning that solely relying on rating data for linking domains is not effective in the case that the auxiliary domain (i.e., LT in this case) is sparser in rating data than the target domain (i.e., ML).

Performance for Different Users. We further investigate the performance of TagCDCF for users with different characteristics. Our investigation is focused on the LT data set, which we take to be representative of a case that derives particular benefits from TagCDCF, as suggested by the results above. Although we report the results obtained on one randomly selected partition, the same trend can be observed on the other partitions. Our first goal is to understand the ability of TagCDCF in alleviating data sparseness. For this reason, we analyze the performance of the TagCDCF and other baseline approaches for users with different number of rated items (cf. Fig. 3).

As can be seen from Fig. 3(a), most users rated limited number of items, i.e., <20 items, while much fewer users rated a lot of items, e.g., >100 items, a common situation in recommender systems. The "<20" group of users usually contains those who are most likely to suffer from the data sparseness problem. For this group of users, the TagCDCF achieves over 8% improvement, compared to the single domain CF approaches, i.e., UBCF and PMF, as shown in Fig. 3(b). These results indicate that TagCDCF could be particularly helpful for users with sparse rating profiles. The other cross-domain CF approach, CBT, also shows a more modest benefit than TagCDCF,

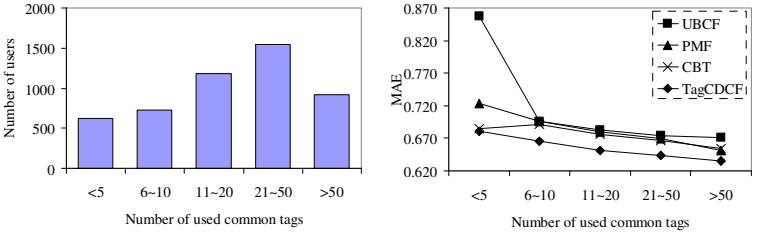
but still can be seen to perform particularly well for users with sparse rating profiles. This similarity confirms that TagCDCF shares the same advantage as other cross-domain CF approaches in specifically addressing the data-sparseness problem. In addition, we can see that for users with relatively more rated items, CBT fails to outperform single domain CF approaches. In contrast, the improvement introduced by TagCDCF consistently outperforms single domain CF approaches across the board, indicating that the improvement introduced when tag-induced cross-domain similarities are exploited as a source of information to link domains could be robust for users with different rating profiles.



(a) (b)

Fig. 3. (a) User distribution of the number of rated items. (b) Performance for users with different number of rated items.

Impact of number of tags shared between domains. Our final goal is to investigate the impact of the number of tags shared between domains on the performance of TagCDCF. Here, we again report experimental results on a randomly selected partition from the LT data set. We analyze the performance of TagCDCF for users who use a different number of tags from the set of tags shared between domains (cf. Fig.4).



(a) (b)

Fig. 4. (a) User distribution of the number of used common tags. (b) Performance for users with different number of used common tags.

We notice that the MAE decreases when users use more tags that are shared between domains. This result indicates that the greater the number of tags used by the user that are common across the domains, the more benefits could be introduced by exploiting tag-induced cross-domain similarities. UBCF demonstrates a marked

performance deterioration on users using very limited number of shared tags, i.e., <5. This spike suggests that users using fewer shared tags use less tags in general, and are perhaps also not very active with respect to assigning ratings. Lack of adequate numbers of assigned ratings would make these users challenging to the UBCF approach, which depends on ratings to reliably calculate user neighborhoods. We also observe that when the user only used very limited number of common tags, i.e., <5 common tags on the LT data set, the performance of TagCDCF is close to the CBT, indicating that the benefits introduced by tags are limited. However, as can be seen in Fig. 4(a), the number of users making use of a limited number of shared tags is not the majority. Most users use more than 5 common tags and are able, as shown in Fig. 4(b), to benefit more from TagCDCF than from competing approaches.

5 Conclusions and Future Work

We have presented TagCDCF, a novel algorithm that is able to improve recommendation performance in multiple domains by linking them via user-assigned tags. TagCDCF extends a matrix factorization approach to collaborative filtering by making use of tags as a source of explicit information that connects users and items across domains. Cross-domain similarities calculated on the basis of user-assigned tags are used to constrain matrix factorization, resulting in improved recommendation performance. Experimental investigation demonstrated that TagCDCF improved the performance in both domains being linked, with more dramatic performance improvements observed in the domain with greater data sparseness. TagCDCF was shown to outperform baselines representative of conventional single-domain CF approaches as well as a state-of-the-art cross-domain CF approach. The relative size of improvement achieved for users with few rated items demonstrates the ability of TagCDCF to address the sparse-data problem. A final set of experiments showed that the improvement offered by TagCDCF is related to the number of tags used by a user that are common to the domains being linked. TagCDCF was observed to outperform other approaches once a user made use of a relatively small number of shared tags.

Our future work will involve further investigation of specific characteristics of recommendation domains. Here, we have seen that the number of tags shared between domains makes an important contribution to TagCDCF performance. We are also interested in discovering additional properties that make two domains particularly suited to benefit each other via TagCDCF. As formulated here, TagCDCF exploits tags to calculate independent item-to-item and user-to-user similarities. We intend to explore whether integrating information on user-tag co-occurrences can be used to refine these similarities. Finally, although TagCDCF was designed and developed to exploit user-contributed tags in the recommender systems domain, the framework is also potentially applicable to cases in which other explicit comparisons can be made between users and between items. We will investigate the effectiveness of alternate information sources, e.g., information derived from content analysis, as the basis for cross-domain linkage within the TagCDCF framework.

Acknowledgements. The research leading to these results was carried out within the PetaMedia Network of Excellence and has received funding from the European Commission's 7th Framework Program under grant agreement n° 216444.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE TKDE 17(6), 734–749 (2005)
2. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2006)
3. Clements, M., de Vries, A.P., Reinders, M.J.T.: The influence of personalization on tag query length in social media search. Inf. Process. Manage. 46(4), 403–412 (2010)
4. Herlocker, J., Konstan, J., Borchers, A., Riedl, J.: An algorithmic framework for performing collaborative filtering. In: SIGIR 1999, pp. 230–237 (1999)
5. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. IEEE Computer 42(8), 30–37 (2009)
6. Li, B., Yang, Q., Xue, X.: Can movies and books collaborate?: cross-domain collaborative filtering for sparsity reduction. In: IJCAI 2009, pp. 2052–2057 (2009)
7. Li, B., Yang, Q., Xue, X.: Transfer learning for collaborative filtering via a rating-matrix generative model. In: ICML 2009, pp. 617–624 (2009)
8. Liang, H., Xu, Y., Li, Y., Nayak, R., Tao, X.: Connecting users and items with weighted tags for personalized item recommendations. In: HT 2010, pp. 51–60 (2010)
9. Pan, W., Xiang, E., Liu, N., Yang, Q.: Transfer learning in collaborative filtering for sparsity reduction. In: AAAI 2010, pp. 230–235 (2010)
10. Salakhutdinov, R., Mnih, A.: Probabilistic matrix factorization. In: NIPS 2008, p. 20 (2008)
11. Sarwar, B., Karypis, G., Konstan, J., Reidl, J.: Item-based collaborative filtering recommendation algorithms. In: WWW 2001, pp. 285–295 (2001)
12. Sen, S., Vig, J., Riedl, J.: Tagommenders: connecting users to items through tags. In: WWW 2009, pp. 671–680 (2009)
13. Singh, A.P., Gordon, G.J.: Relational learning via collective matrix factorization. In: KDD 2008, pp. 650–658 (2008)
14. Symeonidis, P., Nanopoulos, A., Manolopoulos, Y.: A Unified Framework for Providing Recommendations in Social Tagging Systems Based on Ternary Semantic Analysis. IEEE TKDE 22(2), 179–192 (2009)
15. Tipping, M.E., Bishop, C.M.: Probabilistic principal component analysis. Technical Report NCRG/97/010, Neural Computing Research Group, Aston University (1997)
16. Zhang, Y., Cao, B., Yeung, D.-Y.: Multi-domain collaborative filtering. In: UAI 2010 (2010)
17. Zhen, Y., Li, W.-J., Yeung, D.-Y.: TagiCoFi: tag informed collaborative filtering. In: RecSys 2009, pp. 69–76 (2009)