# Multi-site User Behavior Modeling and Its Application in Video Recommendation

Chunfeng Yang
The Chinese University of Hong Kong
Hong Kong, China
yc012@ie.cuhk.edu.hk

Huan Yan
Tsinghua University
Beijing, China
yanh14@mails.tsinghua.edu.cn

Donghan Yu
Tsinghua University
Beijing, China
yudh14@mails.tsinghua.edu.cn

Yong Li
Tsinghua University
Beijing, China
liyong07@tsinghua.edu.cn

Dah Ming Chiu
The Chinese University of Hong Kong
Hong Kong, China
dmchiu@ie.cuhk.edu.hk

## ABSTRACT

As online video service continues to grow in popularity, video content providers compete hard for more eyeball engagement. Some users visit multiple video sites to enjoy videos of their interest while some visit exclusively one site. However, due to the isolation of data, mining and exploiting user behaviors in multiple video websites remain unexplored so far. In this work, we try to model user preferences in six popular video websites with user viewing records obtained from a large ISP in China. The empirical study shows that users exhibit both consistent cross-site interests as well as site-specific interests. To represent this dichotomous pattern of user preferences, we propose a generative model of Multi-site Probabilistic Factorization (**MPF**) to capture both the cross-site as well as site-specific preferences. Besides, we discuss the design principle of our model by analyzing the sources of the observed site-specific user preferences, namely, site peculiarity and data sparsity. Through conducting extensive recommendation validation, we show that our **MPF** model achieves the best results compared to several other state-of-the-art factorization models with significant improvements of F-measure by 12.96%, 8.24% and 6.88%, respectively. Our findings provide insights on the value of integrating user data from multiple sites, which stimulates collaboration between video service providers.

## CCS CONCEPTS

•**Information systems → Collaborative filtering; Personalization;** *Decision support systems;*

## KEYWORDS

Multi-site transfer learning, user modeling, recommender systems

## 1 INTRODUCTION

Online video consumption has become one of the most popular Internet activities worldwide. This trend has created a few very popular video content providers (sites). For example, YouTube and Netflix are among the top websites around the world. Similarly, most users in China also look for videos to watch from a few top video sites. These video content providers compete hard for user eyeballs. In the current market equilibrium, these top websites all tend to have their specialization in video content, e.g. some focus more on movies, while others focus on TV shows, or musical shows. Besides, there are often some overlaps of video catalogs among different sites. Some users visit multiple video sites to find videos of their interest, while some exclusively visit one site.

Normally, each video content provider only has access to its own user-video viewing records. Due to our collaboration with a major ISP in China, we are provided with access to a dataset that includes records of user accessing all video websites in a big city for a window of time. This opens up the possibility for us to investigate various interesting questions about multi-site user modeling: What can we learn about users from multi-site data that is not possible from a single site? How much value is the multi-site data to any video content provider? The answers to these questions allow us to understand more global patterns of online video service, and the nature of competition between video service providers, and whether it makes sense for them to share information.

In the systems studied, the first challenge for accurate user modeling is data sparsity [9], which affects many applications, including personalized recommendation, customer relationship management, targeted advertising, etc., and can be alleviated with more data. In our dataset, more than half of the users are multi-homed (in more than one site's data) users, who view more videos in general than exclusive users. In addition, while the majority of videos are exclusive videos, multi-homed videos are much more popular on average and contribute a considerable proportion of views to the system. Due to the existence of multi-homed users and multi-homed videos, an additional impact of multiple-site data on recommender systems is that one site may recommend to a user some videos that have been viewed in another site by this user. However, in our dataset, a user rarely views a video multiple times, which means knowing multiple-site data can avoid making potential duplicate recommendation. This issue is referred to as information completeness in this paper.

One important finding from our analysis is that there are both cross-site commonality and site peculiarity. Specifically, the multi-homed videos or common types of videos among different sites contribute to the consistent cross-site user preferences, while the exclusive videos affect users' choices in each site, leading to different observed interests in each site. This is evident from the result of principle component analysis in the empirical study section. Because of site peculiarity, merging the multiple-site data as if it is a single merged site does not exploit the most value from the multi-site data. Thus, we propose a generative model of Multi-site Probabilistic Factorization (**MPF**) to capture both the cross-site as well as site-specific aspects of user preferences. Moreover, we illustrate the design principle of our model by analyzing the origins of the observed site-specific user preferences, namely, site peculiarity and data sparsity that also affect how much information should be transferred between sites. The **MPF** model tries to seek a balance between exploiting multi-site data for a smoothed model and capturing site-specific information with each site's data.

We further compare the performance of different factorization models via a variety of recommendation experiments, including experimentation of all six sites, pairs of two sites, different number of latent features, and various sparsity levels. The results show that our model achieves the best recommendation performance by accurately capturing multi-site user preferences. Our findings provide insights for data sharing between video service providers, that is, sharing user viewing data is beneficial for improving recommendation. The main contributions of this work are as follows.

- We measure the user viewing records in six popular video websites, which shows potential value of multiple-site data (alleviating data sparsity and avoiding potential duplicate recommendation) and consistent cross-site as well as site-specific aspects of user preferences.
- We propose a generative model of Multi-site Probabilistic Factorization (**MPF**) to capture the cross-site commonality and site peculiarity of user preferences, which describes how users select videos in multiple video websites.
- We summarize the design principle of our model by analyzing the sources of observed discrepancy of user preferences in multiple sites, i.e., site peculiarity and data sparsity are shown to co-exist and affect the optimal degree of transferring we can apply.
- We conduct various experiments of video recommendation to show that our model outperforms several other state-of-the-art factorization models with significant improvements of F-measure by 12.96%, 8.24% and 6.88%, respectively, which provides insights for win-win data sharing between video service providers.

The rest of this paper is organized as follows. We introduce the data collection with basic statistics in Sec. 2. The empirical study on the data and the motivation of our model are shown in Sec. 3. In Sec. 4, we describe and analyze the proposed model of **MPF**. Then, experimental results are illustrated in Sec. 5. Finally, we present the related work in Sec. 6 and conclude the paper in Sec. 7.

## 2 DATA COLLECTION AND DESCRIPTION

### 2.1 Data Collection

The dataset in our study is an anonymized viewing log of 8,062,857 users, which is collected by a major ISP's fixed network from Shanghai, one of the major metropolitan in China, between November 1 and December 31, 2014. There are over 205 million viewing logs of 6 most popular video websites, including Youku (YK), Iqiyi (IQI), Sohu Video (SH), Kankan (KK), LETV (LE), and Tencent (TC). Each log contains user ID, timestamp and request URL. By crawling and parsing the video URLs, we obtain video title, type and viewed website from the respective content providers. Specifically, we classify videos into 6 common video types including show, TV, user generated content (UGC), movie, cartoon, and news. Meanwhile, we find multi-homed videos via matching the video titles. To be specific, since different content providers have their own naming conventions of videos titles, we first manually identify the rules. For example, we observe that the content provider' name is embedded at the beginning of the video titles for some video websites. Then, with the identified naming conventions, we can preprocess the video titles from our dataset and distinguish multi-homed videos accurately and effectively.

### 2.2 General Statistics

The general statistics of our dataset, including the number of users, videos, viewing records in each site, etc., are shown in Table 1. Moreover, their distributions in terms of different video types are illustrated in Figure 1. We observe that TV dominates the video collections and contributes the largest part of the total eyeballs in each site. However, the view count of a video type is not necessarily proportional to their catalog volume. For example, in TC, a very small portion of news videos contribute considerable viewing records. We also compare the view count distributions of users and videos, as shown in Figure 2. Unlike the nearly power-law distribution of video popularity, the user activeness (view count per user) distribution is heavy-tailed due to that (1) most users are inactive; (2) individual users cannot view too many videos in a limited period of time.

We categorize users and videos into two types, i.e., exclusive (only in one site's data) and multi-homed (in more than one site's data), as per the number of associated sites. In our data, 50.3% of users are multi-homed users (to be specific, 20.3% in twos sites, 11.4% in three sites, 7.7% in four sites, 5.7% in five sites and 5.2% in all six sites), who view more videos (47 videos vs. 5 videos on average) in general than exclusive users. Moreover, while 97% of videos are exclusive videos, multi-homed videos are much more popular on average (956 views vs. 62 views) and contribute more than 25% of views to the system.

## 3 EMPIRICAL STUDY AND MOTIVATION

In our data collection, the first challenging issue for accurate user modeling is data sparsity [9]. As shown in Figure 2, around two thirds of the users have no more than 10 viewing records. Many applications are affected by this issue, such as personalized recommendation, customer relationship management, targeted advertising, etc.

**Table 1: General Statistics**

|                  | YK       | IQI      | SH       | KK       | LE       | TC       |
|------------------|----------|----------|----------|----------|----------|----------|
| Total user count | 4479428  | 3156249  | 3156351  | 1351640  | 2190856  | 2781976  |
| Total video count| 1936643  | 169889   | 174229   | 59041    | 111250   | 130240   |
| Total view count | 90647660 | 24721556 | 39927427 | 8455611  | 24091297 | 17883308 |



(a)                                         (b)                                         (c)
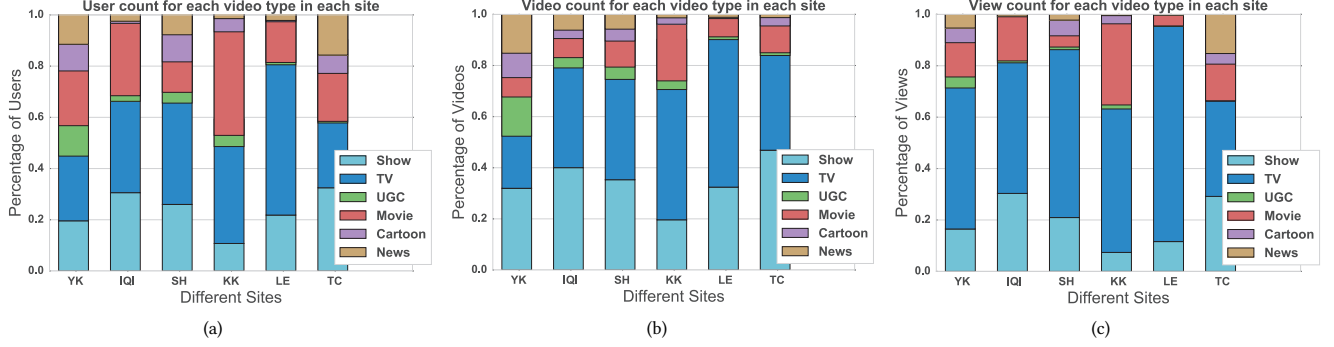
**Figure 1: Number of users, videos, and viewing records of different video types in each site.**
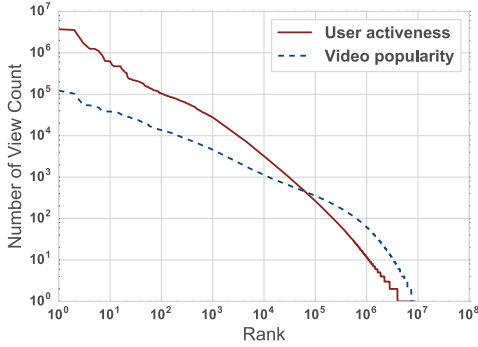


**Figure 2: Distribution of user activeness and video popularity.**
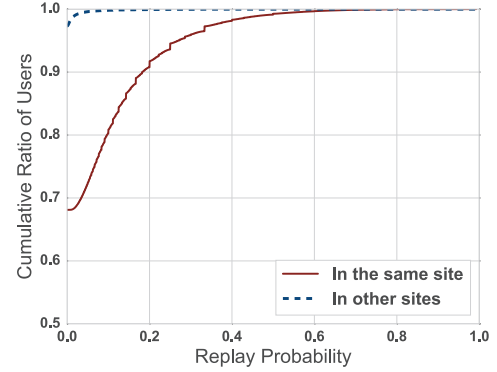


**Figure 3: Cumulative distributions of users' replay probability in the same site and other sites.**

Furthermore, since there are multi-homed users and multi-homed videos, it may happen that one site recommends to a user some videos that have been viewed in another site by this user. However, as shown in Figure 3, where replay probability of a user is defined as the ratio of views occurring on videos that have been viewed previously by her, users rarely replay videos watched before either in the same site[1] or in other sites, which means multiple-site data can help avoid making potential duplicate recommendation. This issue is referred to as information completeness. To show to what extent this issue exists, we define a potential duplicate recommendation for a certain site as that one of its multi-homed users viewed a video in another site while the same video also exists in the catalog

of this site. Then we compare the amount of potential duplicate recommendation with the total view count in each site.

From the results shown in Table 2, we can observe that for most sites (especially IQI, LE and TC), the total amount of potential duplicate recommendation is quite large, which reflects a proprietary aspect of the global information's value.

Merging the multiple-site data can mitigate the data sparsity problem and eliminate the information completeness issue, which, however, may ignore the discrepancy between different sites. Intuitively, the exclusive videos affect users' choices in each site, leading to difference in observed user interest. To check whether site peculiarity exists, we conduct empirical studies from the perspectives of videos and users, respectively.

*Video-based Discrepancy.* We calculate the Spearman's rank correlation coefficient [13] between the popularity of multi-homed

---

[1]The true value of intra-site replay probability may be even smaller because the case that users spend multiple time periods to finish watching a video was also counted as replaying.

**Table 2: Information Completeness**

| | Total potential duplicate recommendation | Total view count |
|---|---|---|
| YK | 9056541 | 90647660 |
| IQI | 19762540 | 24721556 |
| SH | 12172228 | 39927427 |
| KK | 1826602 | 8455611 |
| LE | 12404769 | 24091297 |
| TC | 11048972 | 17883308 |

videos in two sites, which measures the similarity of the popularity order of the multi-homed videos. The reason we use a rank correlation coefficient is to alleviate the impact of different eyeball volumes in these sites. The results in Table 3 show that videos that are more popular in one site are not necessarily for another site.

**Table 3: Spearman Correlation of Multi-homed Videos' Popularity in Two Sites (Significant at the 0.01 Level)**

| Spearman Correlation | YK | IQI | SH | KK | LE | TC |
|---|---|---|---|---|---|---|
| TC | 0.460 | 0.543 | 0.353 | 0.242 | 0.656 | 1 |
| LE | 0.542 | 0.589 | 0.513 | 0.314 | 1 | |
| KK | 0.197 | 0.299 | 0.151 | 1 | | |
| SH | 0.515 | 0.524 | 1 | | | |
| IQI | 0.471 | 1 | | | | |
| YK | 1 | | | | | |

*User-based Discrepancy.* Considering the case of two sites, $u_i^{(1)}$ and $u_i^{(2)}$ are used to denote the latent interest factors in site 1 and site 2 for multi-homed user $i$. We use the Principal Component Analysis (PCA) to derive $u_i^{(1)}$ and $u_i^{(2)}$ for multi-homed users who have many viewing records[2] (more than 100 views), without distinguishing the sites of multi-homed videos. The number of principal components is set so that the proportion of explained variance is up to 95 percent. Then we measure the average Pearson product-moment correlation between $u_i^{(1)}$ and $u_i^{(2)}$ for those multi-homed users[3].

The results in Table 4 show that a user's interest in different sites is neither identical nor independent, validating the existence of both cross-site commonality and site peculiarity[4]. The discrepancy on the video basis as well as the user basis reflects site peculiarity from the aggregate level (videos) and the individual level (users), respectively. While the site peculiarity may stem from different sets of exclusive videos and featured videos (featured videos refer to those received more opportunities of impression or recommendation), the multi-homed videos or common types of videos among

---

[2]To obtain more accurate latent interest of users.

[3]Other distance measures, such as Maximum Mean Discrepancy (MMD) can also be used.

[4]The specific values of correlation are not completely accurate due to finite user-video records, but they can still provide insights for the conclusion.

different sites contribute to the cross-site commonality of user preferences. The co-existence of cross-site and site-specific aspects of user preferences motivates the design of our transfer learning model.

**Table 4: Average Pearson Correlation of Multi-homed Users' Latent Factors in Two Sites (Significant at the 0.01 Level)**

| Pearson Correlation | YK | IQI | SH | KK | LE | TC |
|---|---|---|---|---|---|---|
| TC | 0.295 | 0.391 | 0.365 | 0.390 | 0.361 | 1 |
| LE | 0.281 | 0.374 | 0.300 | 0.331 | 1 | |
| KK | 0.223 | 0.334 | 0.378 | 1 | | |
| SH | 0.190 | 0.396 | 1 | | | |
| IQI | 0.299 | 1 | | | | |
| YK | 1 | | | | | |
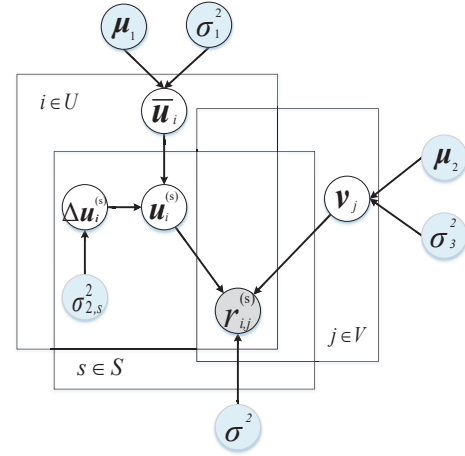


**Figure 4: Graphical representation of MPF.**

## 4 MODEL SPECIFICATION AND ANALYSIS

### 4.1 Model Specification

In this section, we propose a generative model of Multi-site Probabilistic Factorization (**MPF**) to capture the cross-site commonality and site peculiarity. The goal of the generative model is not just for solving the recommendation problem (predictive), but also for modeling multi-site user preferences that can be used to describe how users select videos so as to result roughly in the given distribution of users choosing the web sites as in the viewing records (descriptive).

We denote the set of sites in our study as $S$. The user set in all sites and in site $s$ is denoted as $U$ and $U^{(s)}$ with the size $|U|$ and $|U^{(s)}|$, respectively. The video set in all sites and in site $s$ is represented as $V$ and $V^{(s)}$ with the size $|V|$ and $|V^{(s)}|$, respectively. User $i$'s viewing video set in site $s$ is denoted as $V_i^{(s)}$. The set $S_i$ is the sites where user $i$ view videos. Thus, the multi-homed user

set is $\{i \in U \mid |S_i| > 1\}$. We use $R$ to represent the user-video viewing matrix in all sites. Let $R^{(s)}$ be the $|U^{(s)}| \times |V^{(s)}|$ viewing matrix for site $s$, whose element $r_{i,j}^{(s)} = 1$ if $j \in V_i^{(s)}$, and $r_{i,j}^{(s)} = 0$, otherwise. Note that, since we only have the implicit feedback data, i.e., the videos viewed by users, which take up a very small portion of all the videos, in order to solve the one-class problem, we adopt the sampling-based approach proposed in [14] to balance the positive and negative samples, which will be discussed in the Sec. 5. Multiple viewing of the same video from a user is only counted as one, which seldomly happens since the replay probability is very low.

The basic idea of **MPF** is to model cross-site user preferences and site-specific user preferences simultaneously, where the cross-site common part is learned over multiple-site data together and the site-specific part is learned with each site's data separately. With user-video viewing records data, we use matrix factorization techniques to find a common latent representation for users and videos in these sites. Let $\mathbf{U}^{(s)} \in \mathbb{R}^{K \times |U^{(s)}|}$ and $\mathbf{V}^{(s)} \in \mathbb{R}^{K \times |V^{(s)}|}$ be the latent user and video feature matrices in site $s$, with column vectors $\mathbf{u}_i^{(s)}$ and $\mathbf{v}_j$ representing the $K$-dimensional user-specific latent feature vector of users $i$ in site $s$ and video-specific latent feature vector of video $j$, respectively. Note that, we assume that interest shown by viewing a certain video is independent of the specific site, thus it holds one set of features for multi-homed videos[5].

We define the conditional distribution over the whole viewing matrix as

$$p(R|\mathbf{U}, \mathbf{V}, \sigma^2) = \prod_{s \in S} p(R^{(s)}|\mathbf{U}^{(s)}, \mathbf{V}^{(s)}, \sigma^2)$$
$$= \prod_{s \in S} \prod_{r_{i,j}^{(s)} \in R^{(s)}} \mathcal{N}(r_{i,j}^{(s)}|g\left((\mathbf{u}_i^{(s)})^T \mathbf{v}_j\right), \sigma^2), \quad (1)$$

where $\mathbf{U}$ and $\mathbf{V}$ are the set of user and video latent feature vectors in all sites, respectively. $\mathcal{N}(x|\mu, \sigma^2)$ is the probability density function of the Gaussian distribution with mean $\mu$ and variance $\sigma^2$. The function $g(x)$ is the logistic function $g(x) = \frac{1}{(1+e^{-x})}$ to bound the range of $(\mathbf{u}_i^{(s)})^T \mathbf{v}_j$ within $[0, 1]$, because our data is binary. For users in multiple sites, we decompose their latent feature vectors in each site into two parts, namely, common part and site-specific part: $\mathbf{u}_i^{(s)} = \overline{\mathbf{u}}_i + \Delta\mathbf{u}_i^{(s)}$. For exclusive user $i$ in site $s$, $\mathbf{u}_i^{(s)} = \overline{\mathbf{u}}_i + \mathbf{0}$. The latent feature vector $\mathbf{v}_j$ of a video $j$ is site-agnostic. We also place multivariate Gaussian priors on user and video feature vectors:

$$p(\mathbf{U}|\boldsymbol{\mu_1}, \sigma_1^2, \sigma_{2,*}^2) = \prod_{i \in U} \left( \mathcal{N}(\overline{\mathbf{u}}_i|\boldsymbol{\mu_1}, \sigma_1^2\mathbf{I}) \prod_{s \in S_i} \mathcal{N}(\Delta\mathbf{u}_i^{(s)}|\mathbf{0}, \sigma_{2,s}^2\mathbf{I}) \right),$$
$$p(\mathbf{V}|\boldsymbol{\mu_2}, \sigma_3^2) = \prod_{j \in V} \mathcal{N}(\mathbf{v}_j|\boldsymbol{\mu_2}, \sigma_3^2\mathbf{I}), \quad (2)$$

where each site has a different prior for the site-specific part of users' latent feature vectors. The graphical representation of **MPF** is illustrated in Figure 4. The generative process of the proposed **MPF** model is as follows:

- For each user $i$,
  - draw the vector $\overline{\mathbf{u}}_i \sim \mathcal{N}(\boldsymbol{\mu_1}, \sigma_1^2\mathbf{I})$;

---

- for each site $s \in S_i$, sample $\Delta\mathbf{u}_i^{(s)} \sim \mathcal{N}(\mathbf{0}, \sigma_{2,s}^2\mathbf{I})$, and set the user latent vector as $\mathbf{u}_i^{(s)} = \overline{\mathbf{u}}_i + \Delta\mathbf{u}_i^{(s)}$.
- For each video $j$, draw the vector $\mathbf{v}_j \sim \mathcal{N}(\boldsymbol{\mu_2}, \sigma_3^2\mathbf{I})$.
- For each user-video pair $(i, j)$ in site $s$, draw the value $r_{i,j}^{(s)} \sim \mathcal{N}(g\left((\mathbf{u}_i^{(s)})^T \mathbf{v}_j\right), \sigma^2)$.

Through a Bayesian inference, the posterior probability of the latent feature vector sets $\mathbf{U}$ and $\mathbf{V}$ can be obtained as follows:

$$p(\mathbf{U}, \mathbf{V}|R, \boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \sigma^2, \sigma_1^2, \sigma_{2,*}^2, \sigma_3^2)$$
$$\propto p(R|\mathbf{U}, \mathbf{V}, \sigma^2)p(\mathbf{U}|\boldsymbol{\mu_1}, \sigma_1^2, \sigma_{2,*}^2)p(\mathbf{V}|\boldsymbol{\mu_2}, \sigma_3^2)$$
$$= \prod_{s \in S} \prod_{r_{i,j}^{(s)} \in R^{(s)}} \mathcal{N}(r_{i,j}^{(s)}|g\left((\overline{\mathbf{u}}_i + \Delta\mathbf{u}_i^{(s)})^T \mathbf{v}_j\right), \sigma^2)$$
$$\times \prod_{i \in U} \left( \mathcal{N}(\overline{\mathbf{u}}_i|\boldsymbol{\mu_1}, \sigma_1^2\mathbf{I}) \prod_{s \in S_i} \mathcal{N}(\Delta\mathbf{u}_i^{(s)}|\mathbf{0}, \sigma_{2,s}^2\mathbf{I}) \right)$$
$$\times \prod_{j \in V} \mathcal{N}(\mathbf{v}_j|\boldsymbol{\mu_2}, \sigma_3^2\mathbf{I}). \quad (3)$$

The log of the posterior distribution over the user and video latent features is calculated as

$$\ln p(\mathbf{U}, \mathbf{V}|R, \boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \sigma^2, \sigma_1^2, \sigma_{2,*}^2, \sigma_3^2)$$
$$= -\frac{1}{2\sigma^2} \sum_{s \in S} \sum_{r_{i,j}^{(s)} \in R^{(s)}} \left[ r_{i,j}^{(s)} - g\left((\overline{\mathbf{u}}_i + \Delta\mathbf{u}_i^{(s)})^T \mathbf{v}_j\right) \right]^2$$
$$- \frac{1}{2\sigma_1^2} \sum_{i \in U} \|\overline{\mathbf{u}}_i - \boldsymbol{\mu_1}\|_F^2 - \sum_{i \in U} \sum_{s \in S_i} \frac{1}{2\sigma_{2,s}^2} \left\| \Delta\mathbf{u}_i^{(s)} \right\|_F^2$$
$$- \frac{1}{2\sigma_3^2} \sum_{j \in V} \|\mathbf{v}_j - \boldsymbol{\mu_2}\|_F^2$$
$$- \frac{1}{2} \sum_{s \in S} |U^{(s)}||V^{(s)}| \ln \sigma^2 + \frac{1}{2} K|U| \ln \sigma_1^2$$
$$+ \frac{1}{2} \sum_{s \in S} |U^{(s)}| \ln \sigma_{2,s}^2 + \frac{1}{2} K|V| \ln \sigma_3^2 + C \quad, \quad (4)$$

where $C$ is a constant that does not depend on the parameters. $\|\cdot\|_F^2$ denotes the Frobenius norm. Keeping the parameters, i.e., observation noise variance and prior variance, fixed, maximizing the log-posterior over latent features of users and videos is equivalent to minimizing the following objective function, which is a sum of squared errors with quadratic regularization terms:

$$\mathcal{L}(R, \mathbf{U}, \mathbf{V})$$
$$= \frac{1}{2} \sum_{s \in S} \sum_{r_{i,j}^{(s)} \in R^{(s)}} \left[ r_{i,j}^{(s)} - g\left((\overline{\mathbf{u}}_i + \Delta\mathbf{u}_i^{(s)})^T \mathbf{v}_j\right) \right]^2 + \frac{\lambda_1}{2} \|\overline{\mathbf{u}}_i - \boldsymbol{\mu_1}\|_F^2$$
$$+ \sum_{i \in U} \sum_{s \in S_i} \frac{\lambda_{2,s}}{2} \left\| \Delta\mathbf{u}_i^{(s)} \right\|_F^2 + \frac{\lambda_3}{2} \sum_{j \in V} \|\mathbf{v}_j - \boldsymbol{\mu_2}\|_F^2, \quad (5)$$

where $\lambda_1 = \frac{\sigma^2}{\sigma_1^2}, \lambda_{2,s} = \frac{\sigma^2}{\sigma_{2,s}^2}, \lambda_3 = \frac{\sigma^2}{\sigma_3^2}$. A local minimum of the objective function can be found by performing gradient descent on

---

[5]This is reasonable because videos are static and objective, while users are more dynamic and subjective.

$\overline{\mathbf{u}}_i$, $\Delta\mathbf{u}_i^{(s)}$, $\mathbf{v}_j$, $\mu_1$ and $\mu_2$ for all users and videos.

$$\frac{\partial \mathcal{L}}{\partial \overline{\mathbf{u}}_i} = \sum_{s \in S} \sum_{r_{i,j}^{(s)} \in R^{(s)}} \left[ g\left( (\overline{\mathbf{u}}_i + \Delta\mathbf{u}_i^{(s)})^T \mathbf{v}_j \right) - r_{i,j}^{(s)} \right]$$
$$\times g'\left( (\overline{\mathbf{u}}_i + \Delta\mathbf{u}_i^{(s)})^T \mathbf{v}_j \right) \mathbf{v}_j + \lambda_1 (\overline{\mathbf{u}}_i - \mu_1) , \qquad (6)$$

$$\frac{\partial \mathcal{L}}{\partial \Delta\mathbf{u}_i^{(s)}} = \sum_{r_{i,j}^{(s)} \in R^{(s)}} \left[ g\left( (\overline{\mathbf{u}}_i + \Delta\mathbf{u}_i^{(s)})^T \mathbf{v}_j \right) - r_{i,j}^{(s)} \right]$$
$$\times g'\left( (\overline{\mathbf{u}}_i + \Delta\mathbf{u}_i^{(s)})^T \mathbf{v}_j \right) \mathbf{v}_j + \lambda_{2,s} \Delta\mathbf{u}_i^{(s)} , \qquad (7)$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{v}_j} = \sum_{s \in S} \sum_{r_{i,j}^{(s)} \in R^{(s)}} \left[ g\left( (\overline{\mathbf{u}}_i + \Delta\mathbf{u}_i^{(s)})^T \mathbf{v}_j \right) - r_{i,j}^{(s)} \right]$$
$$\times g'\left( (\overline{\mathbf{u}}_i + \Delta\mathbf{u}_i^{(s)})^T \mathbf{v}_j \right) (\overline{\mathbf{u}}_i + \Delta\mathbf{u}_i^{(s)})^T + \lambda_3 (\mathbf{v}_j - \mu_2) , \qquad (8)$$

$$\frac{\partial \mathcal{L}}{\partial \mu_1} = \frac{\sum_{i \in U} \overline{\mathbf{u}}_i}{|U|} , \qquad (9)$$

$$\frac{\partial \mathcal{L}}{\partial \mu_2} = \frac{\sum_{j \in V} \mathbf{v}_j}{|V|} , \qquad (10)$$

where $g'(x)$ is the derivative of the logistic function and $g'(x) = \frac{e^{-x}}{(1+e^{-x})^2}$.

To reduce the time complexity of model training, we adopted a mini-batch gradient descent approach to learn the parameters. With sampling, the cost of the gradient update no longer grows linearly in the number of matrix entries related to $\mathbf{u}_i^{(s)}$ and $\mathbf{v}_j$, but in the number of entries sampled. The hyper-parameters, i.e., the number of latent features, and regularization coefficients, are set by cross validation.

## 4.2 Model Analysis

On top of the multiple-site data, three other factorization models can be used. If no information is transferred between sites, the model is Separate Matrix Factorization (**SMF**) [19]. If we combine the user-video matrices of all sites directly and apply the matrix factorization to the merged data, the method is called Merged Matrix Factorization (**MMF**). The Collective Matrix Factorization (**CMF**) proposed in [20] assumes the latent features of multi-homed videos to be site-agnostic and learns the user preferences in each site independently. The difference of four factorization models is summarized in Table 5 and their graphical representations are illustrated in Figure 5.

**Table 5: Comparison of Four Factorization Models**

|  | SMF | MMF | CMF | MPF |
|---|---|---|---|---|
| Transfer user latent features? | No | Complete | No | Partially |
| Transfer video latent features? | No | Complete | Complete | Complete |

If we observe non-identical interests in different sites for multi-homed users[6], the disparity mainly results from two factors: (1) site

---

[6]The interest reflected by viewing a video is considered to be site-agnostic, thus we mainly analyze the discrepancy of users' interests in different sites.

peculiarity; (2) data sparsity. The first factor stems from different video catalogs, i.e., exclusive videos, and different featured types of videos, etc., in these sites. If we assume users' global preferences exist, the intrinsic preferences are shaped by site peculiarity, leading to different observed interest in each site. The latter factor means that, with a small sample size, interest is partially exposed, and data sparsity will exaggerate the observed discrepancy of cross-site user preferences.

From the perspective of a model's generalization ability, the generalization risk (a.k.a., testing error) of a model is mainly comprised of estimation error (due to randomness of training data, that is, finite sample size and noise) and approximation error (due to restriction of the model space). If data sparsity is the only reason, we can merge the data, and apply single-task learning method to solve the problem. In this case, the estimation error can be reduced with more data, and MMF is suitable for this case. If site peculiarity dominates, and in the extreme case user preferences are independent across different sites, multi-homed users should be modeled separately to avoid negative transfer between tasks that are less related. Using MMF in this case will obtain a smoothed model and enlarge the approximation error because it restricts the model space of learning heterogeneous patterns in each site. Therefore, CMF is appropriate in this case.

By intuition, users' preferences are unlikely to change drastically from site to site, which means even site peculiarity really exists, users' preferences still share commonality across different sites, as validated in Table 4.[7] Besides, there are many users with very few records. Therefore, these two factors, namely, site peculiarity and site data sparsity[8] co-exist and affect the degree of transferring we can apply. For example, we can transfer more between sites with sparser data and less peculiarity. The degree of transferring is reflected by the hyper-parameter $\lambda_1$ and $\lambda_{2,s}$, which can be set heuristically by cross validation.

Compared with MMF and CMF, our model aims to mitigate the limitations of these two approaches by accounting for both cross-site commonality and site peculiarity. In other words, **MPF** tries to seek a balance between reducing the estimation error by exploiting multi-site data for a smoothed model and capturing site-specific information with each site's data. Therefore, **MPF** is superior to MMF and CMF in modeling multi-site user preferences.

## 5 EXPERIMENTAL DESIGN AND RESULTS

### 5.1 Experimental Design

In this section, we conduct extensive experiments to compare the recommendation capabilities of our proposed **MPF** model with several state-of-the-art factorization models. Our experiments aim at answering the following questions:

(1) What results can be achieved by different algorithms for each site using multiple-site data?

(2) How much one site's recommendation can be improved with each of the other site's data?

---

[7]In the principle component analysis, we selected very active multi-homed users to reduce the impact of data sparsity.

[8]We can even consider individual level data sparsity, i.e., number of user's viewing records. In our work, we did not use distinct regularization parameters for each user since it may cause overfitting.
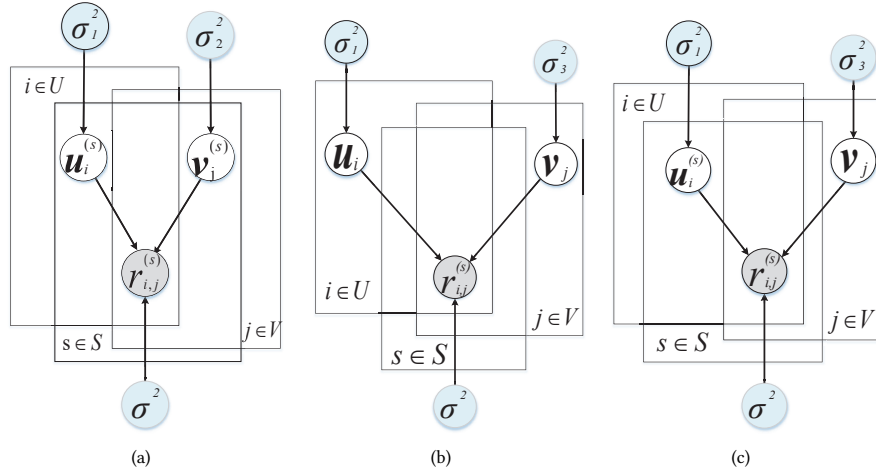
Figure 5: Graphical representation of SMF, MMF, and CMF.

(3) How does $K$ (number of latent features) affect the performance of different approaches?

(4) How do different factorization models perform under various sparsity levels?

To evaluate the **MPF** model that describes how users select videos in different sites, we focus on multi-homed users. We remove videos with too few views (less than 100 views). Since there is no explicit rating data available, we use the implicit feedback data and make top-$N$ recommendation instead of predicting ratings to validate the proposed model.

The generation of training set and test set is as follows. The viewing records data can be viewed as positive samples, which is split randomly into the positive training samples and the positive test samples with 70-30 proportions (this proportion holds for all experiments except for that on different sparsity levels). To balance the positive and negative samples in the training set, for each user, we draw the same number of negative samples as that of her positive training samples from the videos not viewed. Moreover, to avoid predicting all the user-video pairs in the evaluation, which is computationally challenging, we draw 10 times as many negative samples as the number of positive test samples for each user in the test set. Note that, in the test set we only keep users and videos that are also in the training set to make the prediction applicable.

In the model training, we perform 5-fold cross validation to set the hyper-parameters in our experiments. The top-$N$ recommendation generation process for all the factorization models is as follow. For each user i in the test set, we predict a preference score $\hat{r}_{i,j}^{(s)}$ for each video $j$ in the candidate set (union of positive test set and negative test set), where $\hat{r}_{i,j}^{(s)}$ is estimated as $(\mathbf{u}_i^{(s)})^T \mathbf{v}_j$ with different approaches to learn $\mathbf{u}_i^{(s)}$ and $\mathbf{v}_j$ in each factorization model[9]. The top-$N$ recommendation list is obtained by sorting $\hat{r}_{i,j}^{(s)}$ in a descending order and keeping the first $N$ videos.

---

[9] We did not use the logistic function since it's a monotonically increasing function which won't affect the ranking results.

In particular, since users differ in the number of viewed videos, we are not interested in the specific value of $N$. Therefore, we make $N$, namely, the number of videos recommended, equal to the number of her positive samples in the test set for each user. The ratio of negative samples and positive samples holds roughly the same for each user as described in the test set generation process.

*Comparative Algorithms.* To show the performance improvement of our **MPF** model, we use the three following state-of-the-art algorithms.

- **SMF** [19] This is the traditional matrix factorization approach that factorizes the user-video matrix of each site separately.

- **MMF** This approach combines the user-video matrix of all sites and applies the same factorization method as SMF to learn one set of user factors and video factors, irrespective of the specific site.

- **CMF** [20] This method uses one set of factors for multi-homed videos and learns the factors of users in each site separately.

*Evaluation Metrics.* With only binary ground truth data available, we adopt F-measure as the performance metrics. Since $N$ is equal to the number of positive test samples for each user, the F-measure value is the same as the precision value as well as the recall value.

$$\text{F-measure} = \frac{\sum_{i \in U^{\text{test}}} \left| V_i^{\text{test}} \cap V_i^{\text{rec}} \right|}{\sum_{i \in U^{\text{test}}} \left| V_i^{\text{test}} \right|}, \qquad (11)$$

where $U^{\text{test}}$ is the set of users in the test set; $V_i^{\text{test}}$ is the set positive test samples for user $i$; and $V_i^{\text{rec}}$ is the set of top-$N$ recommended videos for user $i$.

## 5.2 Experimental Results

*5.2.1 Overall Performance with Multiple-site Data.* The first question to investigate is how the overall recommendation performance of our approach compares with that of other state-of-the-art factorization models by virtue of multiple-site data. To achieve this, we

train each factorization model based on the data of six sites, and apply the learned models into top-$N$ recommendation of each site to test the performance.

**Table 6: Performance with Data of All Sites**

| F-measure | SMF | MMF | CMF | MPF |
|---|---|---|---|---|
| YK | 0.880 | 0.888 | 0.881 | **0.911** |
| IQI | 0.699 | 0.779 | 0.761 | **0.829** |
| SH | 0.771 | 0.808 | 0.806 | **0.868** |
| KK | 0.694 | 0.717 | 0.706 | **0.788** |
| LE | 0.800 | 0.837 | 0.833 | **0.892** |
| TC | 0.692 | 0.769 | 0.770 | **0.834** |
| Overall | 0.756 | 0.799 | 0.789 | **0.854** |

From Table 6, we can observe that our **MPF** model outperforms the other approaches. On average, it improves the F-measure by 12.96%, 8.24%, and 6.88% compared to SMF, CMF and MMF, respectively. The significant improvements show the promising benefit of our probabilistic factorization approach. The results demonstrate that multi-site data is valuable, however, blindly merging the data cannot make the best of the data. Moreover, all the experimental results in Sec. 5 are statistically significant.

*5.2.2 Site-site Performance Improvement.* We conduct a pairwise experiment to investigate how much the recommendation in one site can be improved by virtue of the data of another one. Specifically, we calculate the improved rate of F-measure achieved by **MPF** over that by SMF for each site when collaborating with one of the other sites. The results are shown in Table 7. Note when using one site's data, **MPF** degenerates to SMF, thus, this experiment is to justify the value of more data. The entry 1.10% in the table means **MPF** can improve 1.10% of F-measure for recommendation in site YK over SMF with the help of data from site IQI.

**Table 7: Site-site Improved Rate of F-measure by MPF over SMF**

| Improved Rate | YK | IQI | SH | KK | LE | TC |
|---|---|---|---|---|---|---|
| YK | \ | 1.10% | 4.35% | 0.97% | 11.3% | 7.86% |
| IQI | 9.70% | \ | 11.8% | 13.0% | 15.2% | 23.7% |
| SH | 1.93% | 8.83% | \ | 9.35% | 6.09% | 17.2% |
| KK | 3.18% | 10.9% | 9.50% | \ | 6.15% | 13.2% |
| LE | 0.45% | 9.68% | 6.08% | 4.96% | \ | 11.6% |
| TC | 9.01% | 18.4% | 13.1% | 10.5% | 11.9% | \ |

As observed in Table 7, the improved rate varies from one pair of sites to another. We further calculate the coefficient of multiple correlation [1] between the improved rate and two predictor variables, namely, site-site correlation of multi-homed users' latent factors shown in Table 4 and data sparsity of each site (i.e., proportion of unobserved entries in the user-video matrix). The multiple correlation coefficient is 0.716, which validates the analysis that data sparsity level and site peculiarity are two of the key factors

affecting the optimal degree of transferring[10]. This also provides insights for tuning the hyper-parameters of regularization.

*5.2.3 Impact of $K$.* Intuitively, increasing $K$, i.e., number of latent features, will add more flexibility to the model. The results in Figure 6 for different values of $K$ show that increasing $K$ from the beginning improves the results. However, after reaching the peak, further increasing $K$ lowers the performance, which may be caused by overfitting with redundant parameters. We obtain the overall performance of different algorithms using data from all sites by changing the value of $K$.
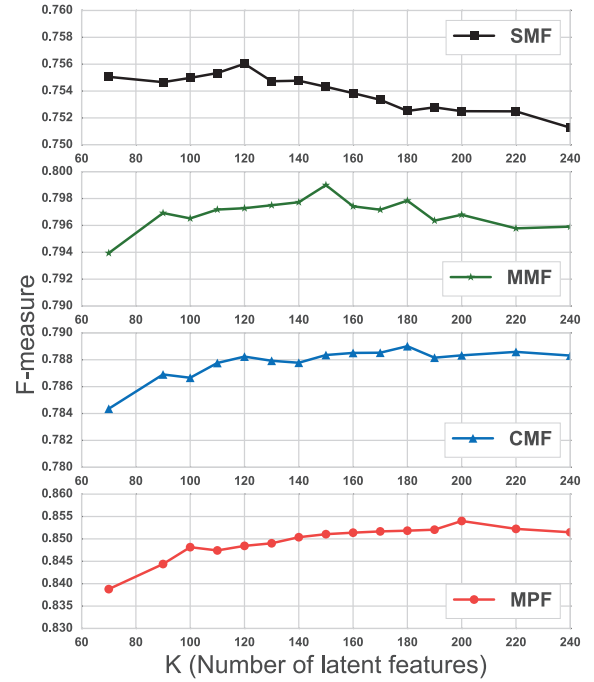


**Figure 6: F-measure of each algorithm with different $K$.**

The result in Figure 6 shows that the optimal value of $K$ in this experiment is MPF > CMF > MMF > SMF. SMF fits users and videos in each site independently. Thus, the optimal number of latent features will not outnumber the largest number of topics in any of the sites. MMF fits users and videos in all sites as a whole and obtained smoothed results (unpopular topics may be overwhelmed). CMF fits videos in all sites, while user preferences are learned separately. **MPF** fits all sites jointly and can learn both the common topics and the site-specific topics. Compared with MMF and MPF, CMF does not need to smooth user preferences. However, with less data in each single site, some weak signals may be neglected. These results also imply that our model captures more of users' interest.

*5.2.4 Different Levels of Sparsity.* We try to explore how multiple site data can help recover the user-video matrix when the data is at different sparsity levels. Specifically, we use different ratios of training data (10%, 30%, 50%, 70%, and 90%) to test all the

---

[10]Other factors may include information completeness, data volume in each site, etc.

algorithms. Training data 90%, for example, means we randomly select 90% of the user-video data as the training data to predict the remaining 10% of data. The corresponding sparsity is 99.983%, 99.958%, 99.930%, 99.903% and 99.875%, respectively. The results are shown in Figure 7.
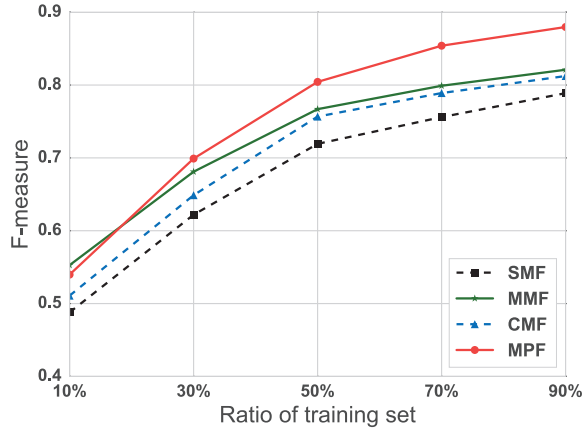


**Figure 7: F-measure under different levels of data sparsity.**

In summary, we have the following important observations:

- With the lowest ratio of training set, MMF performs best, while SMF is worst, where site data sparsity is a dominant factor.
- As data density increases, **MPF** outperforms the other three algorithms, which indicates that **MPF** needs moderate data to capture the site-specific user preferences.
- The gap between SMF and MMF reduces as the training data becomes denser, which means the impact of data sparsity decreases. Note that even when the ratio is 0.9, the data is still quite sparse.

In conclusion, when data sparsity is the main issue for cross-site user behavior modeling, multiple-site information is valuable and combining the user-video matrices of all sites directly (MMF) works well enough. When both data sparsity and site peculiarity matter, **MPF** performs best.

### 5.3 Discussion

The experimental results provide an insight that data sharing of multiple sites is win-win for each other. One concern is how to share the data safely and avoid malicious competition, which is not our focus. A potential strategy is to authorize a trusted third party to gather the multiple-site data and train the **MPF** model, and then distribute only the trained model rather than share the data to each of the sites.

The recommendation performance improvement we obtain from multiple-site data stems from three aspects: data sparsity, information completeness, and personalization. While the first two can be tackled by merging the multiple-site data directly, the last aspect requires the consideration of site peculiarity. **MPF** can control the degree of transferring which is affected by site peculiarity and data sparsity. Site peculiarity and data sparsity, on the other hand, are

two major factors for the observed discrepancy of user preferences in multiple sites.

Instead of optimizing the recommendation performance for a certain site, we focus on modeling the multi-site user behaviors and show the benefit to each site with a unified model. However, it is not difficult to adjust our model into an adaptive transfer learning model [7] by adding weights $W_s$ to the user-video matrix of each site $s$, and learning the optimal weights by cross validation to optimize the recommendation accuracy for a specified site. Another approach for adjustment is iteratively reweighting data samples in both the non-target sites and the target site, that is, increasing the weights of the misclassified data in the target site and reducing the weights of the misclassified data in the non-target sites, which is similar to the idea of TrAdaBoost (i.e., Transfer Adaboost) [5]. To conclude, we use a single model to investigate the problem that whether video websites can benefit from each other from the multi-task learning perspective. Additionally, adaptive transfer learning is more suitable for the case that data in the auxiliary domain is much more than that in the target domain.

## 6 RELATED WORK

In this section, we mainly review some related works on recommendation with implicit feedback and cross-domain collaborative filtering.

Explicit feedbacks, such as numerical ratings, are often unavailable in reality. Thus, recommendation is often based vast amounts of implicit user behaviors, such as views, clicks, purchases [2, 4]. Existing methods handling implicit feedback mainly fall into two categories [3], i.e., imputation-based methods and Bayesian personalized ranking. The basic idea of imputation-based approaches [8] is to convert the one-class data to a balanced dataset by artificially assigning values to some unobserved preferences. Bayesian Personalized Ranking (BPR) [18] utilizes Bayesian inference to train a pairwise ranking model from only positive feedback, based on the assumption that any observed action of a user on an item is an indication that the user should prefer this item to any other item on which she has not performed an action. In addition, with implicit feedback, several matrix factorization methods [10, 21]have been proposed for collaborative filtering, where user and item features are learned through low-rank approximations.

Cross-domain collaborative filtering (CF) is an emerging research topic in recommender systems. It aims to alleviate the sparsity problem in individual CF domains by transferring knowledge among related domains. Matrix factorization based methods are also the state-of-the-art in cross-domain collaborative recommendation because they are able to digest the sparse data well via learning latent variables and are also flexible to incorporate different types of auxiliary data. Some representative cross-domain collaborative filtering methods can categorized into adaptive knowledge transfer and collective knowledge transfer [15] where the former is a directional way exploiting information from a source domain to make recommendations in a target domain, while collective knowledge transfer jointly learns the shared knowledge and unshared effects of multiple domains simultaneously, similar to multi-task learning algorithms. CodeBook Transfer (CBT) [11] studies knowledge transfer ability

between two distinct data domains in an adaptive style. Specifically, it transfers knowledge of cluster-level rating behavior from auxiliary data of movies to target data of books. Rating-Matrix Generative Model [12] extended this idea with a probabilistic model to solve collective transfer learning problems. The drawback of RMGM comes from its inability to handle binary feedback data for transferring. Coordinate System Transfer (CST) [17] is transfer learning method of collaborative filtering to transfer the latent features from auxiliary implicit feedbacks of browsing records to target explicit feedbacks of ratings in an adaptive way. The counterpart of CST with respect to collective knowledge transfer is Transfer by Collective Factorization (TCF) [16]. The defect of CST is that it can not be applied to the multiple domains (more than two). Moreover, the heterogeneity between different domains in our study does not lie in the rating scale. Another popular method for collective transfer of latent features is Collective Matrix Factorization (CMF) [20] which explores co-factorizing multiple matrices in the context of relational learning. Both MMF, as described in Sec. 4.2, and CMF merely capture one aspect of the dichotomous pattern of user preferences.

The method in this paper is a bit similar to that in [6] and [10]. However, the focus of [6] is multi-task SVM rather than multi-site collaborative filtering. Moreover, the focused matrix factorization model (FMF) proposed in [10] is an adaptive knowledge transfer approach that leverages information from non-targeted campaigns into targeted campaigns that are usually smaller, whereas our model is a generative probabilistic model that collectively learns user preferences in multiple sites and shows win-win benefits of data sharing between video service providers.

The main novelty of this work is summarized as follows: 1) No prior work focuses on mining user behaviors in multiple video websites due to the data privacy of different online video service providers; 2) There are multi-homed and exclusive users and videos in the studied system, and information completeness presents one unique value of multiple-site data; 3) We observe the cross-site commonality and site peculiarity and model multiple-homed users' preferences as the integration of the common part and site-specific part; 4) The proposed model can be applied to two or more sites to improve their recommendation performance, acting as the incentive of win-win data sharing for multiple sites.

## 7 CONCLUSION

In this work, with user viewing records obtained from a large ISP of China, we model user preferences in six popular video websites, which is unexplored before. Through real data analysis, we observe the dichotomous pattern of user preferences comprising both consistent cross-site interests as well as site-specific interests. To represent this pattern, we propose a generative model **MPF** to capture both the cross-site as well site-specific aspects of user preferences. The model parameters can be derived via matrix factorization based on our multi-site user video consumption data. The proposed **MPF** model helps address the data sparsity problem, information completeness issue, and personalization of user modeling, which exist in the systems we studied. We conduct a variety of top-N video recommendation experiments to validate

the performance of our model in different scenarios of recommendation. The results show that **MPF** model performs best compared with three state-of-the-art factorization models. Our findings provides insights on the value of combining user data from multiple sites, which motivates win-win data sharing between video service providers.

## REFERENCES

[1] Herve Abdi. 2007. Multiple correlation coefficient. *The University of Texas at Dallas* (2007).

[2] Nicola Barbieri and Giuseppe Manco. 2011. An analysis of probabilistic methods for top-n recommendation in collaborative filtering. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 172–187.

[3] Haolan Chen, Di Niu, Kunfeng Lai, Yu Xu, and Masoud Ardakani. 2016. Separating-Plane Factorization Models: Scalable Recommendation from One-Class Implicit Feedback. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 669–678.

[4] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. 2010. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 39–46.

[5] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. 2007. Boosting for transfer learning. In *Proceedings of the 24th international conference on Machine learning*. ACM, 193–200.

[6] Theodoros Evgeniou and Massimiliano Pontil. 2004. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 109–117.

[7] Ignacio Fernández-Tobías, Iván Cantador, Marius Kaminskas, and Francesco Ricci. 2012. Cross-domain recommender systems: A survey of the state of the art. In *Spanish Conference on Information Retrieval*.

[8] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*. Ieee, 263–272.

[9] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. 2010. *Recommender systems: an introduction*. Cambridge University Press.

[10] Bhargav Kanagal, Amr Ahmed, Sandeep Pandey, Vanja Josifovski, Lluis Garcia-Pueyo, and Jeff Yuan. 2013. Focused matrix factorization for audience selection in display advertising. In *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*. IEEE, 386–397.

[11] Bin Li, Qiang Yang, and Xiangyang Xue. 2009. Can Movies and Books Collaborate? Cross-Domain Collaborative Filtering for Sparsity Reduction.. In *IJCAI*, Vol. 9. 2052–2057.

[12] Bin Li, Qiang Yang, and Xiangyang Xue. 2009. Transfer learning for collaborative filtering via a rating-matrix generative model. In *Proceedings of the 26th annual international conference on machine learning*. ACM, 617–624.

[13] Jerome L Myers, Arnold Well, and Robert Frederick Lorch. 2010. *Research design and statistical analysis*. Routledge.

[14] Rong Pan, Yunhong Zhou, Bin Cao, Nathan N Liu, Rajan Lukose, Martin Scholz, and Qiang Yang. 2008. One-class collaborative filtering. In *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 502–511.

[15] Weike Pan. 2016. A survey of transfer learning for collaborative recommendation with auxiliary data. *Neurocomputing* 177 (2016), 447–453.

[16] Weike Pan, Nathan N Liu, Evan W Xiang, and Qiang Yang. 2011. Transfer learning to predict missing ratings via heterogeneous user feedbacks. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, Vol. 22. 2318.

[17] Weike Pan, Evan Wei Xiang, Nathan Nan Liu, and Qiang Yang. 2010. Transfer Learning in Collaborative Filtering for Sparsity Reduction.. In *AAAI*, Vol. 10. 230–235.

[18] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. AUAI Press, 452–461.

[19] Ruslan Salakhutdinov and Andriy Mnih. 2011. Probabilistic matrix factorization. In *NIPS*, Vol. 20. 1–8.

[20] Ajit P Singh and Geoffrey J Gordon. 2008. Relational learning via collective matrix factorization. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 650–658.

[21] Quan Yuan, Li Chen, and Shiwan Zhao. 2011. Factorization vs. regularization: fusing heterogeneous social relationships in top-n recommendation. In *Proceedings of the fifth ACM conference on Recommender systems*. ACM, 245–252.