Multi-Site User Behavior Modeling and Its Application in Video Recommendation

Huan Yan, Chunfeng Yang, Donghan Yu, Yong Li, *Senior Member, IEEE*, Depeng Jin, and Dah Ming Chiu

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 percent, respectively. Our findings provide insights on the value of integrating user data from multiple sites, which stimulates collaboration between video service providers.

Index Terms—Multi-site transfer learning, user modeling, recommender systems

1 Introduction

With the rapid development of network technologies, videos are much more prevailing [1], [2]. 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 Internet Service Provider (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 [3], 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,

- *H. Yan, D. Yu, Y. Li, and D. Jin are with Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. E-mail: {yanh14, yudh14}@mails.tsinghua.edu.cn, {liyong07, jindp}@tsinghua.edu.cn.*
- *C. Yang and D. Chiu are with Department of Information Engineering, Chinese University of Hong Kong, Hong Kong, China. E-mail: yangchunfeng6731700@gmail.com, dmchiu@ie.cuhk.edu.hk.*

Manuscript received 12 Apr. 2018; revised 5 Jan. 2019; accepted 12 June 2019. Date of publication 1 July 2019; date of current version 7 Dec. 2020. (Corresponding author: Yong Li.) Recommended for acceptance by J. Tang. Digital Object Identifier no. 10.1109/TKDE.2019.2926078

TABLE 1
General Statistics of Users, Videos and Views in 6 Sites

	YK	IQI	SH	KK	LE	TC
Total user count (10^3)	4,479	3,156	3,156	1,351	2,798	2,781
Total video count (10^3)	1,936	169	174	59	111	130
Total view count (10^3)	90,647	24,721	39,927	8,455	24,091	17,883
# multi-homed users (10^3)	3,302	2,488	3,100	1,112	2,347	2,445
# multi-homed videos	56,531	41,700	51,809	3,716	34,853	21,334

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 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 percent, 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 Section 2. The empirical study on the data and the motivation of our model are shown in Section 3. In Section 4, we describe and analyze the proposed model of *MPF*. Then, experimental results are illustrated in Section 5. Finally, we present the related work in Section 6 and conclude the paper in Section 7.

This paper is extended from our previous work [4] to give a more comprehensive analysis. First, we add the analysis of the multi-homed users and multi-homed videos based on the real data, which helps us to better understand the features of user interests across sites. Second, we add two new metrics called video gap and user viewing entropy to provide a more insightful analysis of our motivation of multi-site user behavior modeling. Importantly, we add two more experiments to further evaluate the effectiveness of our model.

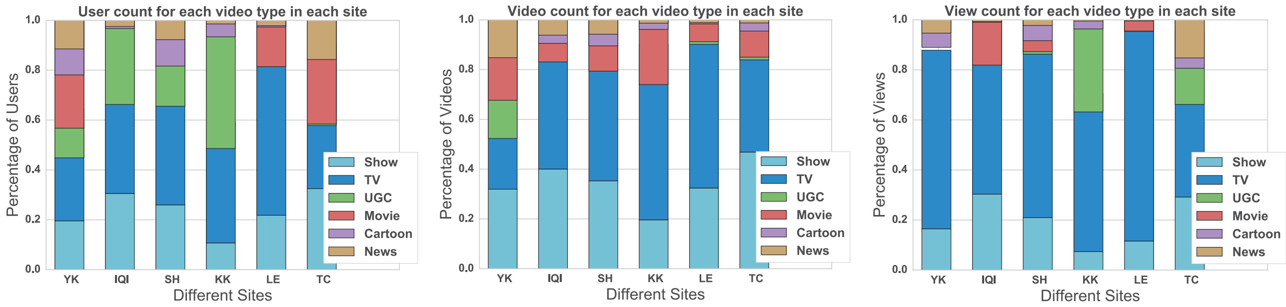
2 Data Collection and Description

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 identity (ID), timestamp and request Uniform Resource Locator (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 Fig. 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 Fig. 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.



(a) Users (b) Videos (c) Views

Fig. 1. Number of users, videos, and viewing records of different video types in each site.

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 percent of users are multi-homed users. Fig. 3 shows the distribution of multi-homed users over different number of sites. Among them, there are 20.3 percent in twos sites, 11.4 percent in three sites, 7.7 percent in four sites, 5.7 percent in five sites and 5.2 percent in all six sites. Compared with exclusive users, they view more videos in general (47 videos versus 5 videos on average). Meanwhile, we compute the number of views by two types of users, and plot the result in Fig. 4. We observe that multi-homed users have more views than exclusive ones. On the other hand, 97 percent of videos are exclusive videos. However, multi-homed videos are much more popular on average (956 views versus 62 views), and contribute more than 25 percent of views to the system. Their distribution over multiple sites is plotted in Fig. 5, where the number of videos decreases with the number of sites. Meanwhile, we calculate the number of views on two types of videos, and plot the result in Fig. 6. We can see that over 30 percent of multi-homed videos have more than 100 views but for exclusive ones the number only occupies 2 percent, which indicates that multi-homed videos attract much more views than exclusive ones.

3 Empirical Study and Motivation

In our data collection, the first challenging issue for accurate user modeling is data sparsity [3]. As shown in Fig. 2, around two thirds of the users have no more than 10 viewing records. Many applications are affected by this issue,

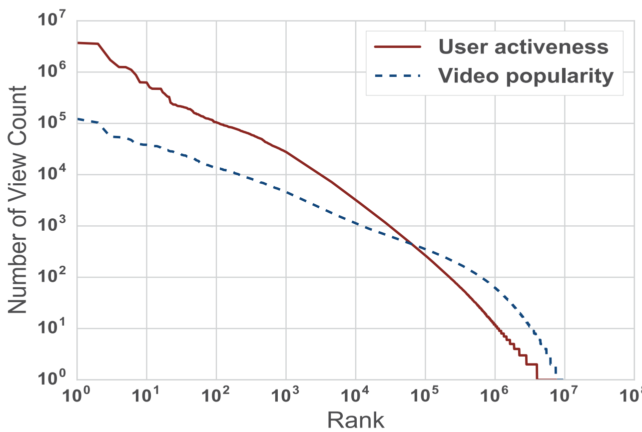


Fig. 2. Distribution of user activeness and video popularity.

such as personalized recommendation, customer relationship management, targeted advertising, etc.

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 Fig. 7, 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. To study how different multi-homed videos are viewed in each site, the view gap of multi-homed video v is calculated as follows:

$$G(v) = \frac{\max_{i\in\mathcal{I}_v} g_i(v) - \min_{i\in\mathcal{I}_v} g_i(v)}{\min_{i\in\mathcal{I}_v} g_i(v)}, \tag{1}$$

where $g_i(v)$ is the number of views of video v in site i, and $\mathcal{I}_v$ denotes the set of sites that have video v. Fig. 8 shows the corresponding cumulative distribution of view gap. We observe that there are nearly 30 percent of videos whose view gap is larger than 10, which indicates that multi-homed videos have great differences on views at different sites.

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.

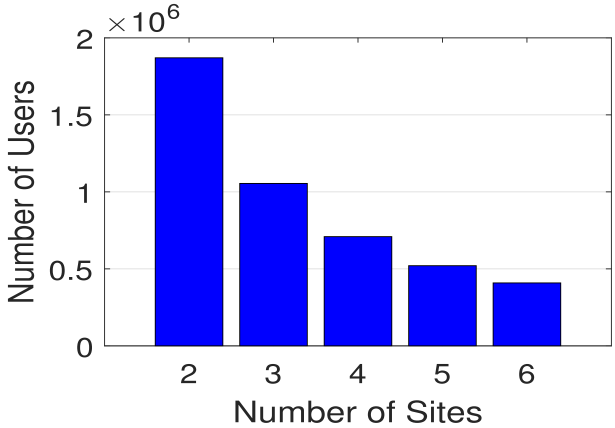


Fig. 3. Distribution of multi-homed users who have visited different number of sites.

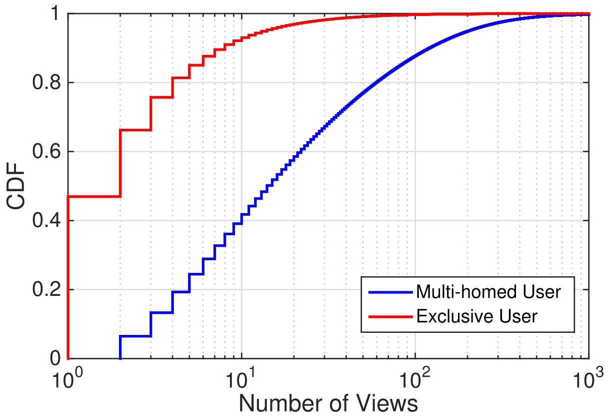


Fig. 4. Distribution of views by multi-homed and exclusive users.

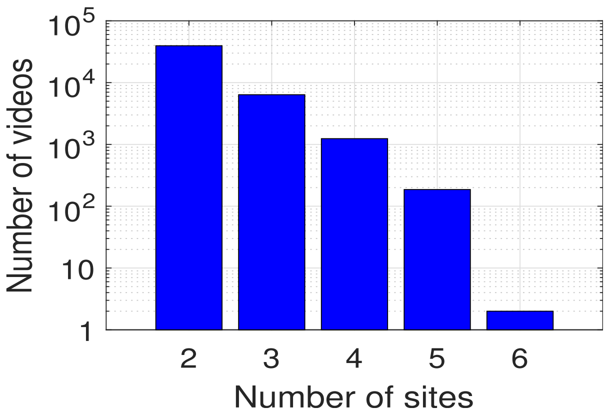


Fig. 5. Distribution of multi-homed videos over multiple sites.

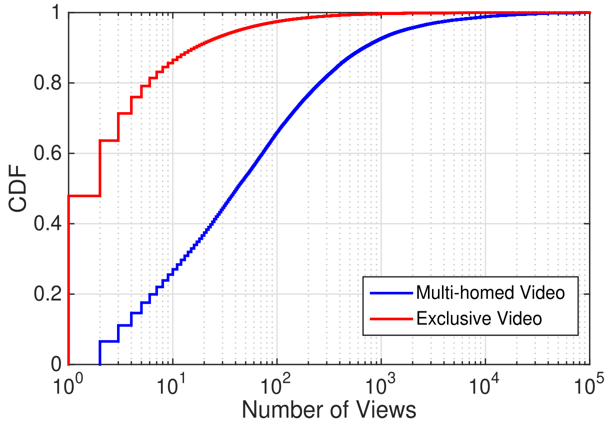


Fig. 6. Distribution of views on multi-homed and exclusive videos.

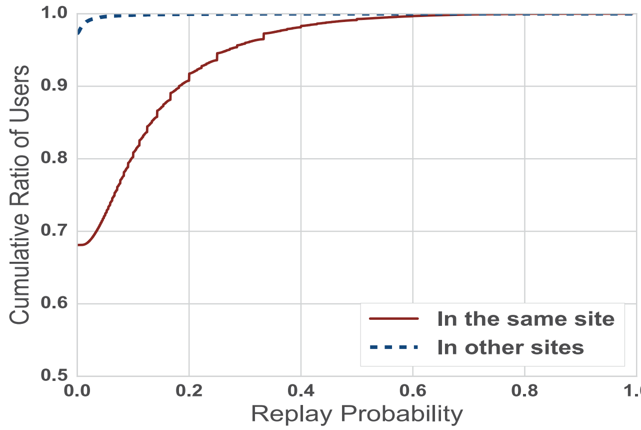


Fig. 7. Cumulative distributions of users' replay probability in the same site and other sites.

TABLE 2
Information Completeness: Comparison of the Number of Potential Duplicate Recommendation and Total View Count in Each Site

	Total potential duplicate recommendation	Total view count
YK	9,056,541	90,647,660
IQI	19,762,540	24,721,556
SH	12,172,228	39,927,427
KK	1,826,602	8,455,611
LE	12,404,769	24,091,297
TC	11,048,972	17,883,308

Moreover, we calculate the Spearman's rank correlation coefficient [5] between the popularity of multi-homed 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.

User-based Discrepancy. First, we use an entropy analysis approach to study how diverse multi-homed users watch different types of videos. Thus, a user viewing entropy $H_i(u)$ is defined as follows:

$$H_i(u) = -\frac{1}{log_2|L|}\sum_{l\in L}\frac{n_u^i(l)}{\sum_{l\in L} n_u^i(l)} log_2\left(\frac{n_u^i(l)}{\sum_{l\in L} n_u^i(l)}\right), \tag{2}$$

where $H_i(u)$ is the viewing entropy of user u at site i. L is the set of video types, and $|L|$ is the size of set L. $n_u^i(l)$ denotes the number of views of user u watching the videos belonging to type l at site i. A higher value of user viewing entropy indicates that user interests are more diversely distributed across different video types at the specified site. Fig. 9 shows the entropy distribution over different sites, where Fig. 9b considers multi-home users with more than 100 views. We have two following observations: (1) In the different sites, the distributions of entropy are obviously different. For example, there are 35 percent of users whose viewing entropy is higher

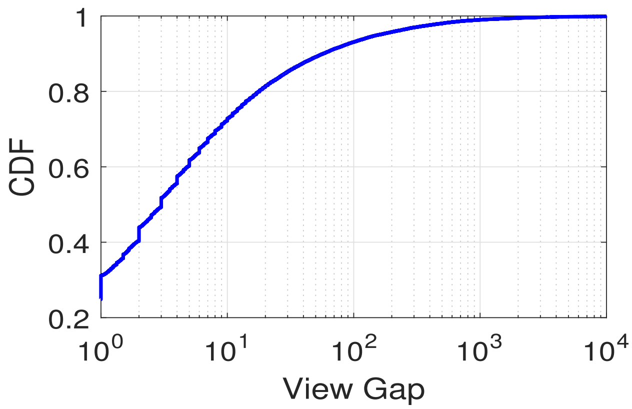


Fig. 8. Cumulative distribution of view gap for multi-homed videos.

than 0.4 in YK, while in LE and KK smaller than 10 percent of users satisfy the same condition; (2) In the same site, the distribution of entropy is neither uniform nor centralized. Even though a larger percentage of users in YK has high entropy, nearly 40 percent of users only watch a type of videos (the corresponding entropy equals to 0). These indicates that user interests exhibit great differences over different sites.

Further, we define the average entropy of multi-home user u as follows:

$$\overline{H}(u) = \sum_{i \in \mathcal{I}_u} H_i(u)/|\mathcal{I}_u|, \tag{3}$$

where $\mathcal{I}_u$ is the set of sites that user u has visited to watch videos, and $|\mathcal{I}_u|$ is the size of set $\mathcal{I}_u$. A lower value of the average entropy indicates that users prefer to some specified types of videos. We classify multi-home users based on the number of sites visited, and compute their average entropy. Fig. 10 shows the distribution of average entropy, where the more sites the user visits, the higher the average entropy is, which indicates the greater differences on user interests.

Next, 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. To extract them, we choose the Principal Component Analysis (PCA) since it can convert the possible correlated components into linearly uncorrelated ones. Moreover, it has the advantages of the dimensionality reduction and the sparsity alleviation. We construct the viewing matrix for each site, whose row and column of the matrix represent the multi-homed users and multi-homed videos, respectively. Note that we use a common set of multi-homed videos without distinguishing the sites.[2] By performing PCA, we derive $u_i^{(1)}$ and $u_i^{(2)}$ for multi-homed users who have many viewing records[3] (more than 100 views). 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,[4] which can reveal the similarities or differences of user interests in different sites.

2. Because the video latent features are common, it is comparable to measure the differences of user interests in different sites.
3. To obtain more accurate latent interest of users.
4. Other distance measures, such as Maximum Mean Discrepancy (MMD) can also be used.

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					

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.[5] 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 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.

4 Model Specification and Analysis

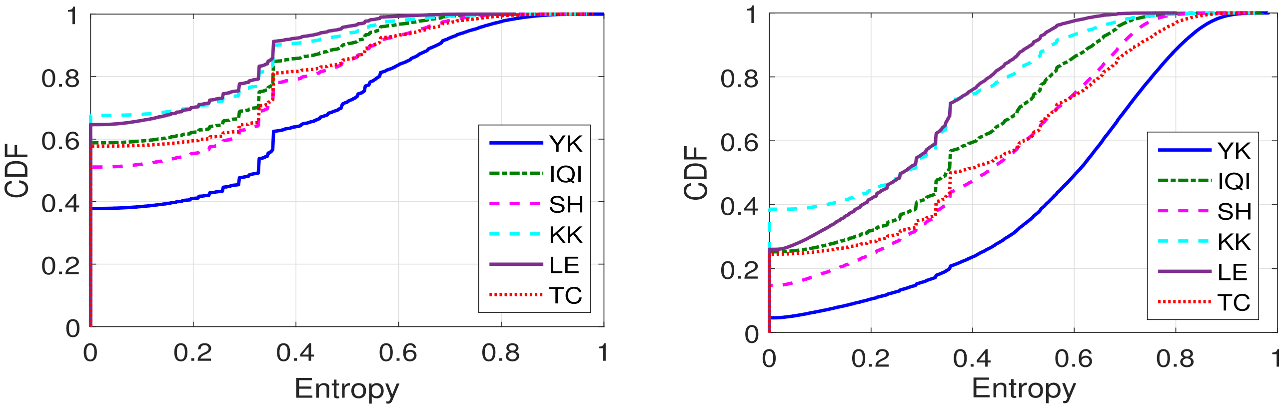
4.1 Model Specification

In this section, we propose a generative model of Multi-site Probabilistic Factorization 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 $\{i \in U \,|\, |S_i| > 1\}$. 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 [6] to balance the positive and negative samples, which will be discussed in the Section 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. The

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



(a) All users (b) Users who have 100+ views

Fig. 9. CDF of viewing entropy of multi-homed users in YK, IQI, SH, KK, LE and TC.

cross-site common part is relevant with multiple-site data; while the site-specific part that shows the site peculiarity of user interests is relevant 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. This is reasonable because videos are static and objective, while users are more dynamic and subjective.

We define the conditional distribution over the whole viewing matrix as

$$\begin{aligned} p(R|\mathbf{U},\mathbf{V},\sigma^2) &= \prod_{s\in S} p(R^{(s)}|\mathbf{U},\mathbf{V},\sigma^2) \\ &= \prod_{s\in S} p(R^{(s)}|\mathbf{U}^{(s)},\mathbf{V}^{(s)},\sigma^2) \\ &= \prod_{s\in S}\prod_{r_{i,j}^{(s)}\in R^{(s)}} \mathcal{N}\left(r_{i,j}^{(s)}|g\left((\mathbf{u}_i^{(s)})^T\mathbf{v}_j\right),\sigma^2\right), \end{aligned} \tag{4}$$

where $\mathbf{U}$ and $\mathbf{V}$ are the set of user and video latent feature vectors in all sites, respectively. Since the viewing matrices of different sites are independent, it holds the first equation. Further, the viewing matrix $R^{(s)}$ is determined by $\mathbf{U}^{(s)}$ and $\mathbf{V}^{(s)}$, thus we obtain the second equation. Similar with the probabilistic matrix factorization in [7], we adopt the linear Gaussian model to define the conditional distribution $p(R^{(s)}|\mathbf{U}^{(s)},\mathbf{V}^{(s)},\sigma^2)$. $\mathcal{N}(x|\mu,\sigma^2)$ is the probability density function of the Gaussian distribution with mean μ and variance σ^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:

$$\begin{aligned} &p(\mathbf{U}|\boldsymbol{\mu}_1,\sigma_1^2,\sigma_{2,*}^2) \\ &= \prod_{i\in U}\left(\mathcal{N}(\overline{\mathbf{u}}_i|\boldsymbol{\mu}_1,\sigma_1^2\mathbf{I})\prod_{s\in S_i}\mathcal{N}(\Delta\mathbf{u}_i^{(s)}|\mathbf{0},\sigma_{2,s}^2\mathbf{I})\right), \\ &p(\mathbf{V}|\boldsymbol{\mu}_2,\sigma_3^2) = \prod_{j\in V}\mathcal{N}(\mathbf{v}_j|\boldsymbol{\mu}_2,\sigma_3^2\mathbf{I}), \end{aligned} \tag{5}$$

where each site has a different prior for the site-specific part of users' latent feature vectors. $\boldsymbol{\mu}_1$ and σ_1 are the mean and variance of Gaussian distribution for common user latent feature vectors, respectively. $\sigma_{2,*}$ represents the variances in different sites. The graphical representation of *MPF* is illustrated in Fig. 11. The generative process of the proposed *MPF* model is as follows:

- For each user i,
  - draw the vector $\overline{\mathbf{u}}_i \sim \mathcal{N}(\boldsymbol{\mu}_1,\sigma_1^2\mathbf{I})$;

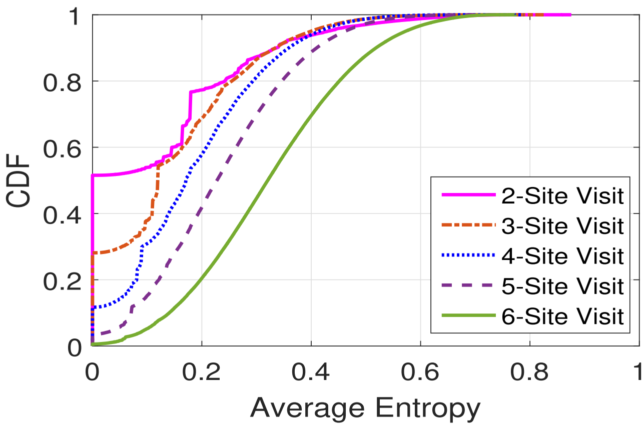


Fig. 10. CDF of average entropy of multi-homed users who visit different number of sites.

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					

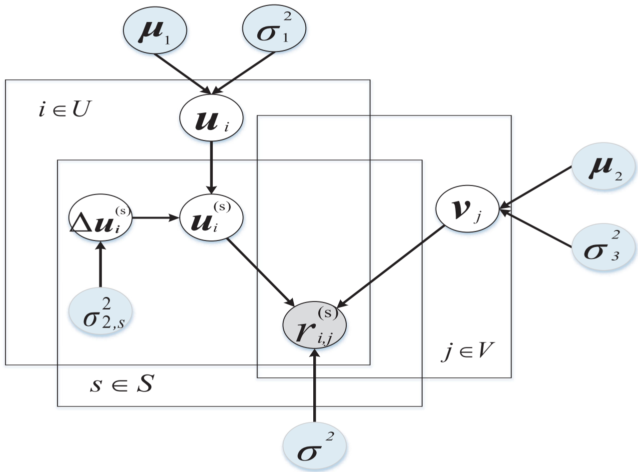


Fig. 11. Graphical representation of *MPF*. For each user i, the user latent vector $\mathbf{u}_i^{(s)}$ is decomposed into common part $\overline{\mathbf{u}}_i$ and site-specific part $\Delta\mathbf{u}_i^{(s)}$. Draw the vector $\overline{\mathbf{u}}_i \sim \mathcal{N}(\boldsymbol{\mu}_1, \sigma_1^2\mathbf{I})$, $\Delta\mathbf{u}_i^{(s)} \sim \mathcal{N}(\mathbf{0}, \sigma_{2,s}^2\mathbf{I})$, $\mathbf{v}_j \sim \mathcal{N}(\boldsymbol{\mu}_2, \sigma_3^2\mathbf{I})$, and $r_{i,j}^{(s)} \sim \mathcal{N}(g((\mathbf{u}_i^{(s)})^T\mathbf{v}_j), \sigma^2)$. The function $g(x)$ is the logistic function.

  - 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 $r_{i,j}^{(s)} \sim \mathcal{N}(g((\mathbf{u}_i^{(s)})^T\mathbf{v}_j), \sigma^2)$.

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

$$\begin{aligned}
&p(\mathbf{U}, \mathbf{V}|R, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \sigma^2, \sigma_1^2, \sigma_{2,*}^2, \sigma_3^2) \\
&\quad\propto p(R|\mathbf{U}, \mathbf{V}, \sigma^2)p(\mathbf{U}|\boldsymbol{\mu}_1, \sigma_1^2, \sigma_{2,*}^2)p(\mathbf{V}|\boldsymbol{\mu}_2, \sigma_3^2) \\
&\quad= \prod_{s\in S} \prod_{r_{i,j}^{(s)} \in R^{(s)}} \mathcal{N}\left(r_{i,j}^{(s)} \Big| g\Big((\overline{\mathbf{u}}_i + \Delta\mathbf{u}_i^{(s)})^T\mathbf{v}_j\Big), \sigma^2\right) \\
&\quad\times \prod_{i\in U}\left(\mathcal{N}(\overline{\mathbf{u}}_i|\boldsymbol{\mu}_1, \sigma_1^2\mathbf{I}) \prod_{s\in S_i} \mathcal{N}(\Delta\mathbf{u}_i^{(s)}|\mathbf{0}, \sigma_{2,s}^2\mathbf{I})\right) \\
&\quad\times \prod_{j\in V} \mathcal{N}(\mathbf{v}_j|\boldsymbol{\mu}_2, \sigma_3^2\mathbf{I}).
\end{aligned} \tag{6}$$

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

$$\begin{aligned}
&\ln p(\mathbf{U}, \mathbf{V}|R, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \sigma^2, \sigma_1^2, \sigma_{2,*}^2, \sigma_3^2) \\
&\quad= -\frac{1}{2\sigma^2}\sum_{s\in S}\sum_{r_{i,j}^{(s)}\in R^{(s)}}\left[r_{i,j}^{(s)} - g\Big((\overline{\mathbf{u}}_i + \Delta\mathbf{u}_i^{(s)})^T\mathbf{v}_j\Big)\right]^2 \\
&\quad - \frac{1}{2\sigma_1^2}\sum_{i\in U}\|\overline{\mathbf{u}}_i - \boldsymbol{\mu}_1\|_F^2 - \sum_{i\in U}\sum_{s\in S_i}\frac{1}{2\sigma_{2,s}^2}\left\|\Delta\mathbf{u}_i^{(s)}\right\|_F^2 \\
&\quad - \frac{1}{2\sigma_3^2}\sum_{j\in V}\|\mathbf{v}_j - \boldsymbol{\mu}_2\|_F^2 \\
&\quad - \frac{1}{2}\sum_{s\in S}|U^{(s)}||V^{(s)}|\ln\sigma^2 + \frac{1}{2}K|U|\ln\sigma_1^2 \\
&\quad + \frac{1}{2}\sum_{s\in S}|U^{(s)}|\ln\sigma_{2,s}^2 + \frac{1}{2}K|V|\ln\sigma_3^2 + C,
\end{aligned} \tag{7}$$

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

$$\begin{aligned}
&\mathcal{L}(R, \mathbf{U}, \mathbf{V}) \\
&= \frac{1}{2}\sum_{s\in S}\sum_{r_{i,j}^{(s)}\in R^{(s)}}\left[r_{i,j}^{(s)} - g\Big((\overline{\mathbf{u}}_i + \Delta\mathbf{u}_i^{(s)})^T\mathbf{v}_j\Big)\right]^2 \\
&\quad + \frac{\lambda_1}{2}\|\overline{\mathbf{u}}_i - \boldsymbol{\mu}_1\|_F^2 + \sum_{i\in U}\sum_{s\in S_i}\frac{\lambda_{2,s}}{2}\left\|\Delta\mathbf{u}_i^{(s)}\right\|_F^2 \\
&\quad + \frac{\lambda_3}{2}\sum_{j\in V}\|\mathbf{v}_j - \boldsymbol{\mu}_2\|_F^2,
\end{aligned} \tag{8}$$

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 $\overline{\mathbf{u}}_i$, $\Delta\mathbf{u}_i^{(s)}$, $\mathbf{v}_j$, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ for all users and videos.

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\overline{\mathbf{u}}_i} &= \sum_{s\in S}\sum_{r_{i,j}^{(s)}\in R^{(s)}}\left[g\Big((\overline{\mathbf{u}}_i + \Delta\mathbf{u}_i^{(s)})^T\mathbf{v}_j\Big) - r_{i,j}^{(s)}\right] \\
&\quad \times g'\Big((\overline{\mathbf{u}}_i + \Delta\mathbf{u}_i^{(s)})^T\mathbf{v}_j\Big)\mathbf{v}_j + \lambda_1(\overline{\mathbf{u}}_i - \boldsymbol{\mu}_1)\ ,
\end{aligned} \tag{9}$$

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\Delta\mathbf{u}_i^{(s)}} &= \sum_{r_{i,j}^{(s)}\in R^{(s)}}\left[g\Big((\overline{\mathbf{u}}_i + \Delta\mathbf{u}_i^{(s)})^T\mathbf{v}_j\Big) - r_{i,j}^{(s)}\right] \\
&\quad \times g'\Big((\overline{\mathbf{u}}_i + \Delta\mathbf{u}_i^{(s)})^T\mathbf{v}_j\Big)\mathbf{v}_j + \lambda_{2,s}\Delta\mathbf{u}_i^{(s)},
\end{aligned} \tag{10}$$

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\mathbf{v}_j} &= \sum_{s\in S}\sum_{r_{i,j}^{(s)}\in R^{(s)}}\left[g\Big((\overline{\mathbf{u}}_i + \Delta\mathbf{u}_i^{(s)})^T\mathbf{v}_j\Big) - r_{i,j}^{(s)}\right] \\
&\quad \times g'\Big((\overline{\mathbf{u}}_i + \Delta\mathbf{u}_i^{(s)})^T\mathbf{v}_j\Big)(\overline{\mathbf{u}}_i + \Delta\mathbf{u}_i^{(s)})^T \\
&\quad + \lambda_3(\mathbf{v}_j - \boldsymbol{\mu}_2)\ ,
\end{aligned} \tag{11}$$

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\mu}_1} = \frac{\sum_{i\in U}\overline{\mathbf{u}}_i}{|U|}, \tag{12}$$

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\mu}_2} = \frac{\sum_{j\in V}\mathbf{v}_j}{|V|}, \tag{13}$$

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*) [7]. If

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

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 [8] 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 Fig. 12.

If we observe non-identical interests in different sites for multi-homed users,[6] the disparity mainly results from two factors: (1) site 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] coexist 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 λ_1 and $\lambda_{2,s}$, which can be set heuristically by cross validation.

MPF uses multi-site data to alleviate the data sparsity. Multi-site data contain the user viewing records over multiple sites, and provide more information about the user interaction with items. Through merging multi-site data, we can transfer the learned knowledges between sites for multi-homed users, and thus more accurately model user preferences.

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?
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.

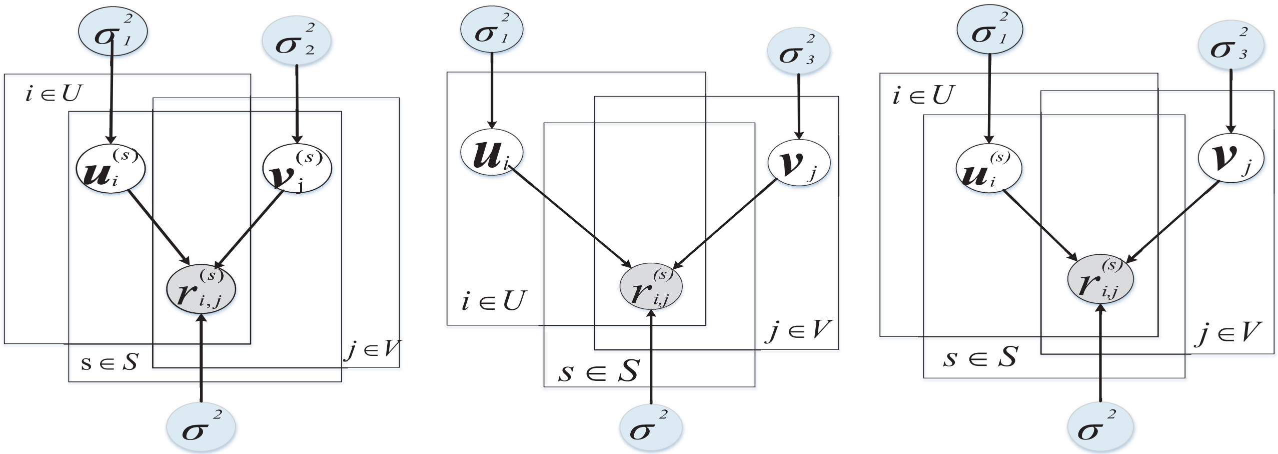
The generation of training set and test set is as follows. The viewing records data can be viewed as positive samples, which is split randomly into the positive training samples and the positive test samples with 70-30 proportions (this proportion holds for all experiments except for that on different 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

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.

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.



(a) SMF (b) MMF (c) CMF

Fig. 12. Graphical representation of SMF, MMF, and CMF. In SMF, the user-video viewing value $r_{i,j}^{(s)}$ is determined by $\mathbf{u}_i^{(s)}$ and $\mathbf{v}_j^{(s)}$, indicating that no information is transferred between sites. In MMF, it is determined by $\mathbf{u}_i$ and $\mathbf{v}_j$, and both of them contain transferred information between sites. In CMF, it is calculated by $\mathbf{u}_i^{(s)}$ and $\mathbf{v}_j$, and only video latent features are transferred.

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.

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.[10]

- *SMF* [7] It is a 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* [8] 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.

$$\mathrm{F-measure} = \frac{\sum_{i \in U^{\mathrm{test}}} \left| V_i^{\mathrm{test}} \cap V_i^{\mathrm{rec}} \right|}{\sum_{i \in U^{\mathrm{test}}} \left| V_i^{\mathrm{test}} \right|}, \tag{14}$$

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

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

10. Our baseline algorithms do not contain deep learning based methods since it is not comparable between them. The *MPF* model aims to characterize the cross-site commonality and site peculiarity based the linear matrix factorization method, while deep learning approaches focus on capturing both the linear and non-linear user-video interaction.

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.

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 percent 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 Section 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

TABLE 6
F-measure of Six Sites by SMF, MMF, CMF and PMF

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**

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%	\

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 percent in the table means *MPF* can improve 1.10 percent of F-measure for recommendation in site YK over SMF with the help of data from site IQI.

As observed in Table 7, the improved rate varies from one pair of sites to another. We further calculate the coefficient of multiple correlation [9] 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.[11] This also provides insights for tuning the hyper-parameters of regularization.

5.2.3 Performance Improvement with Different Number of Sites

Intuitively, if one site uses the data from more sites in the recommender system, the performance can be further improved. However, it is impossible to perform the data sharing of all the sites due to the fierce competition and data safety. Thus, we investigate how the data from different number of sites influence the performance improvement. To evaluate it, we calculate F-measure of one site by virtue of the data from different number of sites. Since small sites are more willing to attract more users by designing a good recommender system, we choose LE and TC in our experiment. Fig. 13 shows the results of LE and TC. In each step, when adding the data of a new site into the training set, we choose the site that can achieve the highest (corresponding to $Selection\ 1$) or lowest (corresponding to $Selection\ 2$) performance respectively.

In Fig. 13, for $Selection\ 1$, we observe that the improved rates of F-measure of LE and TC become slower as the number of sites increases. When using $Selection\ 2$, we find the same trend when the number of sites exceeds 3. This suggests that although multi-site data can improve the recommendation performance, their growth is not very significant after adding the data of a certain amount of sites.

5.2.4 Impact of K

Intuitively, increasing K, i.e., number of latent features, will add more flexibility to the model. The results in Fig. 14 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.

The result in Fig. 14 shows that the optimal value of K in this experiment is $\mathrm{MPF} > \mathrm{CMF} > \mathrm{MMF} > \mathrm{SMF}$. 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.

5.2.5 Different Levels of Sparsity

First, we investigate how one site can help another site at different levels of sparsity. Specifically, we use different ratios of training data (10, 30, 50, 70, and 90 percent) to test all the algorithms. Training data 90 percent, for example, means we randomly select 90 percent of the user-video data as the training data to predict the remaining 10 percent of data. Similar with Section 5.2.3, we choose LE (with TC data) and TC (with IQI data) to calculate their F-measure. Since TC data can hep LE to achieve better performance than the data of any other site, we use TC data to perform the evaluation. The same is for TC. As shown in Fig. 15, MPF achieves best performance among all the algorithms when the rating of training set exceeds 30 percent, and F-measure becomes higher when the training set is much denser.

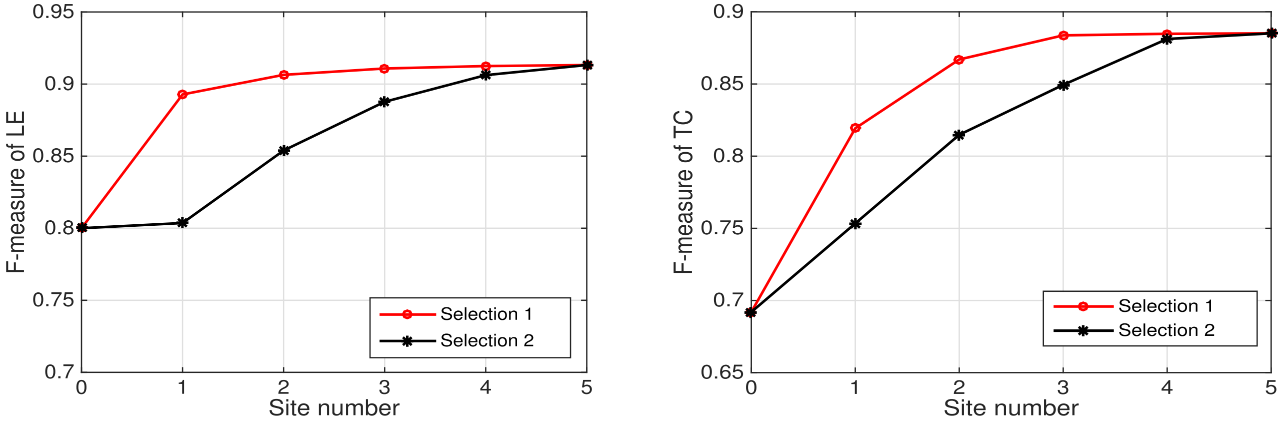
Next, we try to explore how multiple site data can help recover the user-video matrix when the data is at different sparsity levels. The corresponding sparsity is 99.983, 99.958, 99.930, 99.903 and 99.875 percent, respectively. The results are shown in Fig. 16.

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.

11. Other factors may include information completeness, data volume in each site, etc.



(a) LE (b) TC

Fig. 13. F-measure of LE and TC by virtue of the data from different number of sites.

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. Specifically, the third party provides the unified APIs for content providers to upload the viewing records in a safe way. Then, it utilizes the collected data from different sites to train the MPF model. After that, it distributes the trained model to each site to perform the video recommendation. In real recommender systems, it is practical to implement it since content providers do not need to share the data to other sites. Furthermore, they can simply utilize the trained model from the thirty parity to recommend liked videos for users.

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.

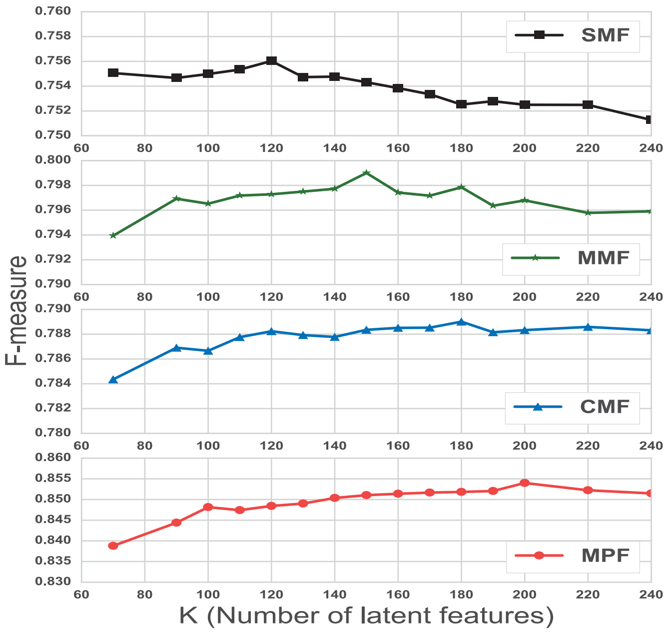


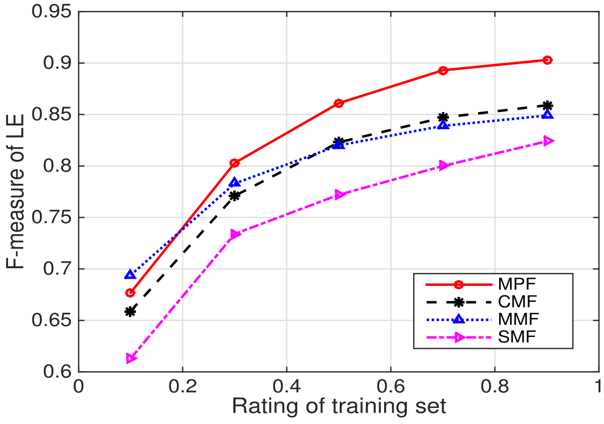
Fig. 14. F-measure of each algorithm with different K.

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 [10] 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) [11]. 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.

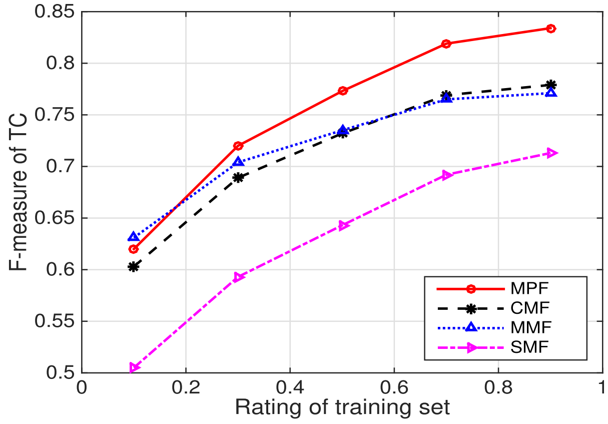
6 Related Work

In this section, we mainly review some related works on multi-site viewing behavior analysis as well as recommendation with implicit feedback and cross-domain collaborative filtering.

Different from most works [12], [13], [14], [15] that study user viewing behavior in a specified site, some [16], [17], [18], [19], [20] focus on multi-site viewing behavior. For example, Huan et al. study video viewing behaviors over multiple content providers in terms of the temporal, deivce-wise and spatial patterns [20]. Krishnan et al. have a quantitive analysis on how the quality of video streaming influences the user behavior in different sites [16]. Liu et al. use multi-site video records to characterize the cache behavior [17]. Authors in [19] explore how and why users switch from one site to another in video consumption. Unlike them, our work captures both cross-site and site-specific patterns of user interests, and models them to obtain better performance of video recommendation.



(a) LE



(b) TC

Fig. 15. F-measure of LE and TC by virtue of the data of one site under different levels of sparsity.

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 [21], [22], [23]. Existing methods handling implicit feedback mainly fall into two categories [24], i.e., imputation-based methods and Bayesian personalized ranking. The basic idea of imputation-based approaches [25] is to convert the one-class data to a balanced dataset by artificially assigning values to some unobserved preferences. Bayesian Personalized Ranking (BPR) [26] 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. Visual BPR (VBPR) considers the influence of visual appearance of items on user preferences. Tang et al. studies the vulnerability of VBPR, and proposed an Adversarial Multimedia Recommendation (AMR) method that performs adversarial learning to improve the robustness of multimedia recommender systems [27]. In addition, with implicit feedback, several matrix factorization methods [28], [29], [30], [31] have been proposed for collaborative filtering (CF), where user and item features are learned through low-rank approximations. Authors in [32] consider the heterogeneity between user and item latent spaces, and design a cross-space tensor model for affinity learning.

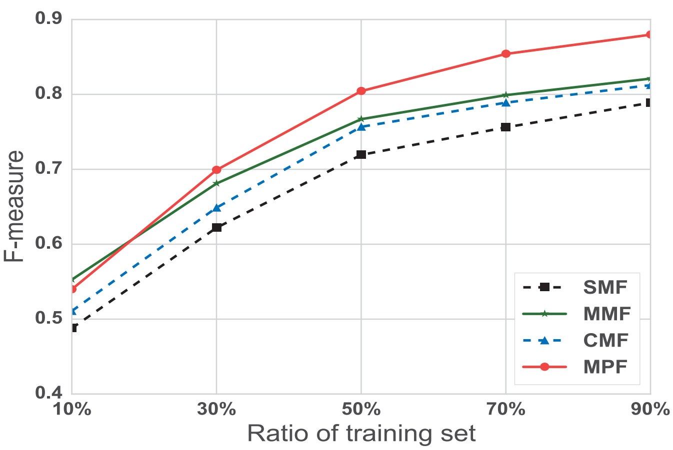


Fig. 16. F-measure under different levels of data sparsity.

Cross-domain collaborative filtering 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 [33] 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) [34] 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 [35] 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) [36] 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) [37]. 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) [8] which explores co-factorizing multiple matrices in the context of relational learning. Both MMF, as described in Section 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 [38] and [39]. However, the focus of [38] is multi-task SVM rather than multi-site collaborative filtering. Moreover, the focused matrix factorization model (FMF) proposed in [39] 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.

Recently, many works use the deep learning based methods in recommender systems [40], [41], [42]. For example, Authors in [40] propose a deep collaborative filtering model to capture the complicated interaction between users and videos. Gao et al. present a dynamic recurrent neural network to model users' dynamic preferences over time [42]. However, they focus on exploring the user viewing behaviors in a specified site. Unlike them, our model can capture multi-site viewing behaviors of multi-homed users. As our future work, we would design the neural network for multi-site viewing behavior mining. One potential approach is to use residual fashion to effectively capture user interests through the neural network.

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.

Acknowledgments

This work was supported in part by The National Key Research and Development Program of China under grant SQ2018YFB180012, the National Nature Science Foundation of China under 61971267, 61972223, 61861136003, and 61621091, Beijing Natural Science Foundation under L182038, Beijing National Research Center for Information Science and Technology under 20031887521, and research fund of Tsinghua University–Tencent Joint Laboratory for Internet Innovation Technology.

References

[1] J. Tang, X. Hua, M. Wang, Z. Gu, G. Qi, and X. Wu, "Correlative linear neighborhood propagation for video annotation," *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)*, vol. 39, no. 2, pp. 409–416, Apr. 2009.

[2] J. Tang, G.-J. Qi, M. Wang, and X.-S. Hua, "Video semantic analysis based on structure-sensitive anisotropic manifold ranking," *Signal Process.*, vol. 89, no. 12, pp. 2313–2323, 2009.

[3] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, *Recommender Systems: An Introduction*. Cambridge, UK.: Cambridge Univ. Press, 2010.

[4] C. Yang, H. Yan, D. Yu, Y. Li, and D. M. Chiu, "Multi-site user behavior modeling and its application in video recommendation," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2017, pp. 175–184.

[5] J. L. Myers, A. Well, and R. F. Lorch, *Research Design and Statistical Analysis*. Abingdon, U.K.: Routledge, 2010.

[6] R. Pan, Y. Zhou, B. Cao, N. N. Liu, R. Lukose, M. Scholz, and Q. Yang, "One-class collaborative filtering," in *Proc. 8th IEEE Int. Conf. Data Mining*, 2008, pp. 502–511.

[7] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2011, vol. 20, pp. 1–8.

[8] A. P. Singh and G. J. Gordon, "Relational learning via collective matrix factorization," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 650–658.

[9] H. Abdi, "Multiple correlation coefficient," *Encyclopedia of measurement and statistics*, The University of Texas at Dallas, pp. 648–65, 2007.

[10] I. Fernández-Tobías, I. Cantador, M. Kaminskas, and F. Ricci, "Cross-domain recommender systems: A survey of the state of the art," in *Proc. Spanish Conf. Inf. Retrieval*, 2012, pp. 187–198.

[11] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 193–200.

[12] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system," in *Proc. 7th ACM SIGCOMM Conf. Internet Meas.*, 2007, pp. 1–14.

[13] V. Gopalakrishnan, R. Jana, K. Ramakrishnan, D. F. Swayne, and V. A. Vaishampayan, "Understanding couch potatoes: Measurement and modeling of interactive usage of IPTV at large scale," in *Proc. ACM SIGCOMM Conf. Internet Meas. Conf.*, 2011, pp. 225–242.

[14] Z. Li, J. Lin, M.-I. Akodjenou, G. Xie, M. A. Kaafar, Y. Jin, and G. Peng, "Watching videos from everywhere: A study of the PPTV mobile VoD system," in *Proc. ACM Internet Meas. Conf.*, 2012, pp. 185–198.

[15] F. Figueiredo, F. Benevenuto, and J. M. Almeida, "The tube over time: Characterizing popularity growth of YouTube videos," in *Proc. 4th ACM Int. Conf. Web Search Data Mining*, 2011, pp. 745–754.

[16] S. S. Krishnan and R. K. Sitaraman, "Video stream quality impacts viewer behavior: Inferring causality using quasi-experimental designs," in *Proc. Internet Meas. Conf.*, 2012, pp. 211–224.

[17] J. Liu, H. Yan, Y. Li, D. Wu, S. Li, and D. Jin, "Cache behavior characterization and validation over large-scale video data," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 3, pp. 734–745, Mar. 2018.

[18] H. Yan, T. Lin, G. Wang, Y. Li, H. Zheng, D. Jin, and B. Zhao, "A first look at user switching behaviors over multiple video content providers," in *Proc. 11th Int. AAAI Conf. Web Social Media*, 2017, pp. 700–703.

[19] H. Yan, T.-H. Lin, G. Wang, Y. Li, H. Zheng, D. Jin, and B. Y. Zhao, "On migratory behavior in video consumption," in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2017, pp. 1109–1118.

[20] H. Yan, T.-H. Lin, C. Gao, Y. Li, and D. Jin, "On the understanding of video streaming viewing behaviors across different content providers," *IEEE Trans. Netw. Serv. Manage.*, vol. 15, no. 1, pp. 444–457, Mar. 2018.

[21] P. Cremonesi, Y. Koren, and R. Turrin, "Performance of recommender algorithms on top-n recommendation tasks," in *Proc. 4th ACM Conf. Recommender Syst.*, 2010, pp. 39–46.

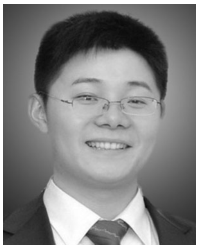
[22] N. Barbieri and G. Manco, "An analysis of probabilistic methods for top-n recommendation in collaborative filtering," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2011, pp. 172–187.

[23] V. C. Ostuni, T. Di Noia, E. Di Sciascio, and R. Mirizzi, "Top-n recommendations from implicit feedback leveraging linked open data," in *Proc. 7th ACM Conf. Recommender Syst.*, 2013, pp. 85–92.
[24] H. Chen, D. Niu, K. Lai, Y. Xu, and M. Ardakani, "Separating-plane factorization models: Scalable recommendation from one-class implicit feedback," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, 2016, pp. 669–678.
[25] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *Proc. 8th IEEE Int. Conf. Data Mining*, 2008, pp. 263–272.
[26] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," in *Proc. 25th Conf. Uncertainty Artif. Intell.*, 2009, pp. 452–461.
[27] J. Tang, X. He, X. Du, F. Yuan, Q. Tian, and T.-S. Chua, "Adversarial training towards robust multimedia recommender system," *IEEE Trans. Knowl. Data Eng.*, to be published, doi: 10.1109/TKDE.2019.2893638.
[28] M. Jamali and M. Ester, "A transitivity aware matrix factorization model for recommendation in social networks," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, vol. 11, pp. 2644–2649.
[29] Q. Yuan, L. Chen, and S. Zhao, "Factorization vs. regularization: Fusing heterogeneous social relationships in top-n recommendation," in *Proc. 5th ACM Conf. Recommender Syst.*, 2011, pp. 245–252.
[30] X. Xin, C.-Y. Lin, X.-C. Wei, and H.-Y. Huang, "When factorization meets heterogeneous latent topics: An interpretable cross-site recommendation framework," *J. Comput. Sci. Technol.*, vol. 30, no. 4, pp. 917–932, 2015.
[31] D. Liang, J. Altosaar, L. Charlin, and D. M. Blei, "Factorization meets the item embedding: Regularizing matrix factorization with item co-occurrence," in *Proc. 10th ACM Conf. Recommender Syst.*, 2016, pp. 59–66.
[32] J. Tang, G. Qi, L. Zhang, and C. Xu, "Cross-space affinity learning with its application to movie recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 7, pp. 1510–1519, Jul. 2013.
[33] W. Pan, "A survey of transfer learning for collaborative recommendation with auxiliary data," *Neurocomput.*, vol. 177, pp. 447–453, 2016.
[34] B. Li, Q. Yang, and X. Xue, "Can movies and books collaborate? Cross-domain collaborative filtering for sparsity reduction," in *Proc. 21st Int. Joint Conf. Artif. Intell.*, 2009, vol. 9, pp. 2052–2057.
[35] B. Li, Q. Yang, and X. Xue, "Transfer learning for collaborative filtering via a rating-matrix generative model," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 617–624.
[36] W. Pan, E. W. Xiang, N. N. Liu, and Q. Yang, "Transfer learning in collaborative filtering for sparsity reduction," in *Proc. 24th AAAI Conf. Artif. Intell.*, 2010, vol. 10, pp. 230–235.
[37] W. Pan, N. N. Liu, E. W. Xiang, and Q. Yang, "Transfer learning to predict missing ratings via heterogeneous user feedbacks," in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, vol. 22, Art. no. 2318.
[38] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 109–117.
[39] B. Kanagal, A. Ahmed, S. Pandey, V. Josifovski, L. Garcia-Pueyo, and J. Yuan, "Focused matrix factorization for audience selection in display advertising," in *Proc. IEEE 29th Int. Conf. Data Eng.*, 2013, pp. 386–397.
[40] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for youtube recommendations," in *Proc. 10th ACM Conf. Recommender Syst.*, 2016, pp. 191–198.
[41] C.-Y. Wu, A. Ahmed, A. Beutel, A. J. Smola, and H. Jing, "Recurrent recommender networks," in *Proc. 10th ACM Int. Conf. Web Search Data Mining*, 2017, pp. 495–503.
[42] J. Gao, T. Zhang, and C. Xu, "A unified personalized video recommendation via dynamic recurrent neural networks," in *Proc. ACM Int. Conf. Multimedia*, 2017, pp. 1–9.

Huan Yan received the BS degree in electronic information engineering from Nanchang University, Nanchang, China, in 2009, and the MS degree in communication engineering from Hangzhou Dianzi University, Hangzhou, China, in 2012. He is currently working toward the PhD degree in the Department of Electronic Engineering, Tsinghua University, Beijing, China. His current research interests include network big data and software-defined network.

Chunfeng Yang received the BS degree from the Huazhong University of Science and Technology, Wuhan, China, and is currently working toward the PhD degree in information engineering from the Chinese University of Hong Kong, Hong Kong. His current research interest includes data analysis, P2P networks, online video, and video recommendation.

Donghan Yu is working toward the BS degree in the Electronic Engineering Department, Tsinghua University, Beijing, China. His research interests include mobile big data mining and recommender system.



Yong Li (M'09–SM'16) received the BS degree in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2007 and the PhD degree in electronic engineering from Tsinghua University, Beijing, China, in 2012. He is currently a faculty member of the Department of Electronic Engineering, Tsinghua University. He has served as general chair, TPC Chair, and SPC/TPC member for several international workshops and conferences, and he is on the editorial board of two IEEE journals. His papers have total citations more than 6,900. Among them, 10 are ESI Highly Cited Papers in Computer Science, and four received conference Best Paper (run-up) Awards. He received the following awards: IEEE 2016 ComSoc Asia-Pacific Outstanding Young Researchers, Young Talent Program of China Association for Science and Technology, and the National Youth Talent Support Program.

Depeng Jin received the BS and PhD degrees from Tsinghua University, Beijing, China, in 1995 and 1999 respectively both in electronics engineering. He is a professor at Tsinghua University and the chair of Department of Electronic Engineering. He was awarded National Scientific and Technological Innovation Prize (Second Class), in 2002. His research fields include telecommunications, high-speed networks, ASIC design and future Internet architecture.

Dah Ming Chiu (A'86–M'03–SM'03–F'08) received the BS degree from Imperial College London, London, United Kingdom, and the PhD degree from Harvard University, Cambridge, MA. He is currently the Department Chairman of Information Engineering with the Chinese University of Hong Kong, Hong Kong. He previously was with Sun Labs, DEC, and AT&T Bell Labs. His current research interest includes P2P networks, network measurement, architecture and engineering, network economics, and social networks. He is a fellow of the IEEE.