Joint User Modeling across Aligned Heterogeneous Sites

Xuezhi Cao , Yong Yu
Apex Data and Knowledge Management Lab
Dept. of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
cxz,yyu@apex.sjtu.edu.cn

ABSTRACT

An accurate and comprehensive user modeling technique is crucial for the quality of recommender systems. Traditionally, we model user preferences using only actions from the target site and may suffer from cold-start problem. As nowadays people normally engage in multiple online sites for various needs, we consider leveraging the cross-site actions to improve the user modeling accuracy. Specifically, in this paper we aim at achieving a more comprehensive and accurate user modeling by modeling user's actions in multiple aligned heterogeneous sites simultaneously. To do so, we propose a modularized probabilistic graphical model framework JUMA. We further integrate topic model and matrix factorization into JUMA for joint user modeling over text-based and item-based sites. We assemble and publish large-scale dataset for comprehensive analyzing and evaluation. Experimental results show that our framework JUMA out performs traditional within-site user modeling techniques, especially for cold-start scenarios. For cold-start users, we achieve relative improvements of 9.3% and 12.8% comparing to existing within-site approaches for recommendation in item-based and text-based sites respectively. Thus we draw the conclusion that aligning heterogeneous sites and modeling users jointly do help to improve the quality of online recommender systems.

Keywords

User Modeling, Recommender System, Graphical Model

1. INTRODUCTION

To improve user experience and to stimulate user actions, a great portion of online services include a native recommender system. Such recommender systems have brought great benefits for both users and service providers. Therefore, there exist plenty research works focusing on this topic. Experiments as well as online applications both prove the success of such systems.

RecSys '16, September 15-19, 2016, Boston , MA, USA

DOI: http://dx.doi.org/10.1145/2959100.2959155

An accurate and comprehensive user modeling technique is crucial for the quality of recommender systems. Traditionally, we model user's preferences using actions from the target site. For example, recommending movies based on one's previous movie rating logs. Although the performances of such approaches are satisfying, there are still problems and limitations with these approaches.

A major problem is that existing approaches may fail when dealing with new users due to insufficient historical data. This problem is referred to as the cold-start problem [26]. It widely exists and severely jeopardizes the user's first impression when exposed to the recommender systems. Several works aim at alleviating this problem using side-information [6, 33] or interview processes [8]. However, such approaches require extra efforts from users thus still jeopardize the users' first experiences.

Another limitation of existing techniques is the lack of comprehensiveness. When participating in specialized online services, users normally reveal only parts of his preferences. As traditional user models focus only on user actions in the target site, they can only capture the most revealed parts of the user preferences. Although these are the most important parts for future recommendations in the same site, we cannot claim other aspects are irrelevant. For example, it is hard to mine one's political preferences using only movie rating histories, but such preferences would be helpful when recommending political movies or documentaries.

On the other hand, as the usage of Internet develops, most people now engage in multiple sites for various needs. For example, we have Facebook for maintaining social relations, Twitter for microblogging, and IMDb for movie reviews. By aligning these sites and aggregating user's online actions, we can directly aim at modeling user's underlying general preferences instead of site-specific preferences. Following this direction, we can achieve comprehensive user preference modeling and actually solve the cold-start problem (users only encounter the cold-start problem once when joining his first online site).

The intuition behind this direction is: although the purposes and action types vary from site to site, the user's underlying preferences remains the same. Every natural person has an underlying general preferences that depend on his/her personality, hobbies and personal taste. Such preferences do not change with the site the user currently engages. We refer such preferences as user's universal preferences in the following of this paper. When participating in a specific site, the user conducts actions based on parts of his universal preferences. Existing works aim at capturing

the revealed parts of preferences in that specific site. By aligning the sites together and jointly model the users preferences in these sites simultaneously, we can directly capture the user's comprehensive universal preferences. When dealing with cold-start scenario where user has only few actions, the preferences learned from other aligned sites can also help to alleviate the problem.

Therefore, in this paper we target at joint user modeling over multiple aligned sites with heterogeneous actions. Extending within-site user modeling to cross-site joint modeling is not a trival task. The major challenge is the heterogeneity of actions from different sites. To tackle this, we propose a modularized probabilistic graphical model framework JUMA. The benefit of modularization is that we can easily plug the state-of-the-art techniques for different types of actions into the framework. The framework achieves great generality and expendability by integrating corresponding modules for different kinds of sites. We discuss how to deploy JUMA for modeling text-based and item-based sites in this paper. We target at these types because most online services can be categorized into these two, thus we can cover most use scenarios.

Note that this approach requires the sites to be fully or partially aligned, i.e. alignment between accounts indicating that they belong to the same natural person is known. Fortunately, now most sites allow users to login with cross-site accounts (*"Login with Facebook Account"*). Researchers also work on how to recover the alignment by analyzing user's profile, social relationship and user generated contents [20, 28]. The state-of-the-art approach achieves an accuracy over 80%. Therefore, the accessibility of such alignment should not be concerned.

The rest of the paper is organized as follows. We first discuss the related works in Section 2. We present the dataset as well as the preliminary analysis in Section 3. We propose out framework JUMA in Section 4, and then report the experimental results in Section 5. Finally, we draw conclusions and propose future works in Section 6.

2. RELATED WORK

2.1 User Modeling

User modeling is the core of recommender systems. There exist plenty research works in this direction. As the user generated contents (UGC) vary from site to site, researchers propose different user modeling techniques accordingly. Matrix factorization technique [15, 22] is widely used for this task, especially for item-based sites such as e-commercial and movie/music rating sites [13, 17]. Zhang et al. proposed an efficient bayesian hierarchical user modeling in [32]. Topic models, such as Latent Dirichlet Allocation [2], are employed for capturing user's topic distributions in text-based sites such as Twitter and Tumblr [7]. There are also works try to leverage social relationships in online social network for improving the quality of user modeling [12, 18].

However, most of existing works tackle each application separately, thus might suffer from cold-start problem and lack of comprehensive user preference due to data insufficiency. Therefore, in this paper we propose joint user modeling that model the user over multiple sites simultaneously and targets at improving the modeling precision as well as comprehensiveness.

2.2 Cold-Start Problem

The cold-start problem is that when giving recommendations to new users who have no or only few historic actions, his/her preferences are not yet revealed to the recommender system. Therefore, the quality of recommendations in these scenarios might be rather poor.

Such problem exists in almost all recommender systems, therefore is of great importance. Researchers propose various solutions to alleviate this problem [9, 16, 26]. One direction is to leverage additional side information to compensate the user/item's novelty. Gao et al. address the problem in location-based recommendation using social network as the side information [6]. Social tags are employed by Zhang et al. in [33]. Those approaches require specific types of side information for the given user, which may not be always available. Another direction is adding an interview process immediately after registration. This is the most widely adopted approach in commercial applications. The fundamental task in this direction is the form of the interview process and the questions asked during it. Golbandi et al. use decision trees to adaptively selecting the questions [8]. Functional matrix factorization is also employed for generating interview questions [35]. Nevertheless, these approaches do need the users to manually answer the questions or go through other types of interview process, thus have a negative impact on the user experiences.

2.3 Cross-Domain Recommendation

The motivation of this paper is to some extent similar with cross-domain recommendation, which is improving user modeling by leveraging historic actions from other sources [5]. However, existing cross-domain approaches mainly focus on homogeneous data. For example, transferring between movies and books [27]. There is also a good fraction of papers target at synthetic multi-domain data generated by subdividing a single-domain dataset [1, 25]. For those scenarios, different 'domains' actually share a lot in common (action type, user preference, behavior pattern, etc.).

Instead, we aim at user modeling across sites with heterogeneous actions, which is more challenging and general. There are also works aiming at heterogeneous data. McAuley et al. aim at understanding product ratings with review text in [21]. However, these approaches require aligned actions which is unavailable in the general setting. Also, such approaches can not directly extend to more than two sites.

2.4 Network Alignment

Although a large portion of online applications now enable users to login or associate with cross-site accounts (Facebook, Twitter, Google+, etc.), there still exist plenty unaligned users and isolated networks. Therefore, automatically aligning the fragmented online social networks is proposed as a new topic in recent years. Researchers propose various approaches to tackle this task by mining different kinds of information. Most works focus on mining the personal identifiable information in profile pages, including username, location, avatar etc. [19, 23, 24]. There are also works leveraging specific types of user generated contents in certain types of networks, for example social tags in tagging systems [10]. Liu et al. tackle the task by modeling users' heterogenous behaviors [20] and achieve a accuracy over 80%.

There are also works aiming at improving existing tasks

Table 1: Dictionaries for Multi-Modal Topic Model over Weibo & Douban

ID	Top Words	Top Movies
1	Food, Cat, Friend, Dog, Home, Like, Life	Hotaru no haka, The Pursuit of Happyness, Jeux d'enfants The Devil Wears Prada, Up, Tonari no Totoro, Ratatouille
2	China, Article, Book, Issue, America, Country, Society	Inception, Social Network, Source Code, Avatar, WALL·E V for Vendetta,The Lord of the Rings, Argo The Shawshank Redemption, The Bourne Identity, Titanic
3	We, Love, Myself, Life, Want, Time, Like, World	Amour, Love Letter, Amélie, Forrest Gump, Before Sunrise Before Sunset, Flipped, Love Actually, The Notebook
4	Design, Art, Photography, Artiest , Magazine	Amélie, Mallína, Les choristes, Léon, Roman Holiday Nuovo Cinema Paradiso, Black Swan, Pulp Fiction

using aligned networks. For example, Zhang et al. employ the aligned networks to improve the link prediction task [31].

3. DATA & PRELIMINARY ANALYSIS

In this section we first explain the data set used in this paper, and then conduct preliminary analysis for better understanding of the user preferences consistency across the aligned heterogeneous sites.

3.1 Data Set

We collect and publish a large-scale data set for analysis and evaluation. For data sources we use Douban[1] and Weibo[2], one of China's largest movie rating sites and microblogging sites respectively. We collect 141,614 randomly selected users who participate in both Weibo and Douban. The account alignments are retrieved from explicit information in Douban's profile pages. For item-based site Douban, we collect the rating histories for each user. On average each user has 123.49 rating logs. For text-based site Weibo, we collect up to 20 pages of microblogs (both original and re-tweet) for each user. On average each user has 343.78 microblogs. We also publish the dataset online for the research community[3]. Ideas and suggestions to extend the dataset are highly welcomed.

3.2 Preliminary Analysis

The underlying assumption of joint user modeling is that user's preferences are to some extent consistent across sites. Despite its intuitiveness, we analyze on real data to support the assumption.

As user preferences are not explicitly expressed, we can not directly measure the preferences consistency. Fortunately, user's action histories serve as explicit indicators of his preferences. Therefore, we show the user action similarity consistency instead.

Specifically, we analyze whether users with similar actions in one site are still tend to be similar in the other. We randomly select 5 million pairs of users and evaluate their pairwise similarity in the two sites according to their action histories. To model user similarity in text-based site Weibo, we first employ Latent Dirichlet Allocation (LDA [2]) to model the topic distribution of each user and then use L_1-distance to measure the dissimilarity between users. And for item-based site Douban, we use Jaccard similarity coefficient

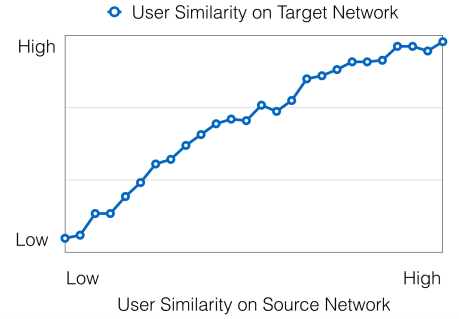


Figure 1: User Similarity Consistency

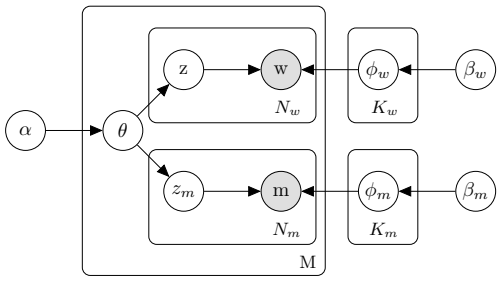


Figure 2: Multi-Modal Topic Model

upon the set of rated movies. Formally:

$$S_w(i,j) = -\sum_k |\theta_{ik} - \theta_{jk}|, \quad S_d(i,j) = \frac{|W_i \cap W_j|}{|W_i \cup W_j|} \qquad (1)$$

where θ_{i*} is the topic distribution of user i and W_i is the set of movies rated by user i. Then we group the pairs with close similarity in the source network (Weibo) to capture the general trend. We report the averaged similarity in the target network (Douban) and show the results in Figure 1.

The results indicate that users with similar actions in Weibo also tend to be similar in Douban on average, which support the assumption and indicate the existence of user preference consistency.

To gain further insight of preference consistency across sites, we employ multi-modal topic model to directly capture the relation between topics in Weibo and movies in Douban. Specifically, we consider movies as another set of 'words' and plug them into the original model. When user watches a movie, he/she first selects a topic according to his/her topic distribution and then chooses the movie according to the topic-movie distribution (corresponding to the topic-word dictionary ϕ in traditional model). We depict the graphical model for the multi-modal topic model in Figure 2.

[1]http://www.douban.com/
[2]http://www.weibo.com/
[3]http://dataset.apexlab.org/juma

We show the resulting word-dictionary (ϕ_w) and movie-dictionary (ϕ_m) for some example topics in Table 1. From the results we can notice the hidden correlation between movies and topics in these two sites. For example, first topic indicates users who tweet about pets are more likely to enjoy comedies, cartoons, and the second topic indicates ones interested in political news tend to prefer movies with depth or background stories.

4. APPROACH

In this section we propose JUMA, a probabilistic graphical model for joint user modeling over multiple aligned sites. We first describe the design of the model in general setting, and then discuss the design details for text-based and item-based sites. Finally we propose a hybrid learning process for the parameter learning.

4.1 JUMA in General

Probabilistic graphical model is a widely used technique for modeling user actions. For example, topic modeling [2], user preferences [29], social network modeling [11] and etc. The reason behind its popularity is its similarity towards the generation of user actions in real-world and its easiness for interpretation. Therefore, we employ probabilistic graphical model for the joint user modeling task.

The design of the general model mostly follows the assumption that user reveals part of his/her universal preferences when participating each specific site. We show the general graphical model in Figure 3 (a), which can be interpreted as follows: Each user i has an universal preference $U_i \in \mathbb{R}^{K_u}$. When participating in site q, the user's site-specific preference $P_i^q \in \mathbb{R}^{K_q}$ is transferred from universal preference U_i based on site-specific transferring model T^q. Finally, user conduct actions A_i^q base on site-specific preference P_i^q and site-specific item models $\{\phi_k^q\}$. $\alpha, \gamma^q, \beta^q$ are hyper-parameters for the prior. σ^q is the parameter that controls the coupling strength between the site-specific preferences and the universal preferences.

We design the universal preference U_i to have a multivariate Gaussian prior with mean 0 and variance α. For the site-specific variables, the model need to be designed accordingly. The item set varies from site to site, e.g. words for text-based sites and music/movie/product for item-based sites. Besides, the plate containing user actions A_i^q also varies according to different action types. Therefore, to plug a specific site into JUMA, we need to design the followings in details:

- **Preference Transferring.** The site-specific preference transferring parameters T^q along with the corresponding transferring model Ω^q in $P_i^q = \Omega^q(T^q, U_i)$.
- **Item Modeling.** Detailed modeling methodology (ϕ_k^q) for the site-specific items, e.g. latent factors for movies or topic-word distributions for topic model.
- **User Actions.** The graphical model for the generation process of user actions (the plate containing A_i^q).

4.2 JUMA for Item-Based & Text-Based Sites

Now we discuss the implementation details of JUMA in the context of modeling item-based and text-based sites. We depict the detailed graphical model of the extended version in Figure 3 (b), where we have two plates for the two sites respectively. We explain the notations we used in Table 2.

Table 2: Notations & Interpretation of JUMA

Notation	Interpretation
U_i	Universal preference of user i
T^d, T^w	Site-specific preference transferring model
P_i^d	User i's movie preferences on Douban
A_{ik}^d	Indicator of whether user i rates movie k
ϕ_k^d	Latent feature vector for movie k
P_i^w	User i's topic distribution on Weibo
z_{ij}	Latent topic of user i's j^{th} word in Weibo
A_{ij}^w	j^th word posted by user i in Weibo
ϕ_k^w	Word distribution in topic k
α, γ, β	Hyper-parameters
σ^d, σ^w	Hyper-parameter for coupling strength

To model user actions in the item-based site Douban, we employ Matrix Factorization (MF) technique [15], the most widely used technique for modeling the item rating actions. It captures user preferences and item characteristics with latent factors. Specifically, we can formally define the MF modular for JUMA by:

- **Preference Transferring.** Item-based site's preference follows multivariate normal distribution with mean $T^d U_i$ and variance σ^d, where $T^d \in \mathbb{R}^{K_d \times K_u}$ is the transfer matrix.

$$P_i^d \sim \mathcal{N}(T^d U_i, \sigma^d), \quad T^d \sim \mathcal{N}(0, \lambda^d) \tag{2}$$

  Increasing σ^d leads to weaker coupling between item-based site's preference with the universal preference.

- **Item Modeling.** $\phi_k^d \in \mathbb{R}^{K_d}$ is the latent factor of movie k, with prior: $\phi_k^d \sim \mathcal{N}(0, \beta^d)$.

- **User Actions.** A_{ik}^d indicates whether user i watches movie k, which is modeled by sigmoid function applied on dot product of user's preference and movie's latent factor.

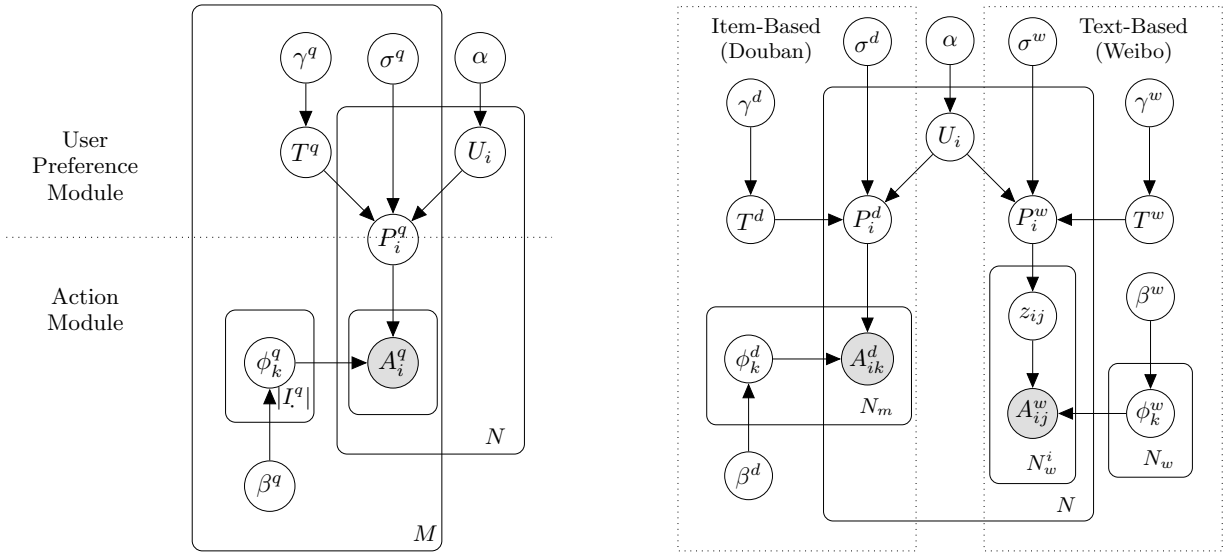
$$A_{ik}^d \sim \text{Bern}(\sigma(P_i^d \phi_k^d)), \quad \sigma(x) = \frac{1}{1+e^{-x}} \tag{3}$$

For text-based site Weibo, we use Latent Dirichlet allocation (LDA) [2] to model the user's topic distribution when writing or retweeting microblogs.

We first quickly go through the settings of traditional LDA model. For each user, we draw his topic distribution from a Dirichlet distribution with parameter α as the non-informative prior ($\theta_i \sim Dir(\alpha)$). For each word, we first draw the hidden topic z_{ij} from user's topic distribution ($z_{ij} \sim Multi(\theta_i)$), and then select the word w_{ij} according to topic-word dictionary $\phi_{z_{ij}}$. The topic-word dictionary ϕ_k also follows Dirichlet distribution with non-informative prior β.

When integrating LDA with JUMA, instead of using non-informative prior for the user's topic distribution, we can now borrow user's universal preference. To keep the conjugate prior, we have:

$$P_i^w \sim Dir(\sigma^w T^w U_i), \quad T^w \sim \mathcal{N}(0, \lambda^w) \tag{4}$$



(a) Joint User Modeling in General (b) Joint User Modeling over Douban & Weibo

Figure 3: JUMA : A Probabilistic Graphical Model for Joint User Modeling via Aggregating Multi-Site Actions

where σ^w is a scale parameter for tuning the coupling strength (prior strength). Item modeling as well as user action modeling are kept the same with LDA.

$$\phi_k^w \sim Dir(\beta^w), \;\; z_{ij} \sim Multi(P_i^w), \;\; A_{ij}^w \sim Multi(\phi_{z_{ij}}^w) \quad (5)$$

4.3 Hybrid Learning Process

In this section we discuss the learning of JUMA. We propose a hybrid learning process to estimate the parameters. The learning is mainly based on Gibbs sampling technique [30], a Markov chain Monte Carlo algorithm for approximating multivariate probability distribution.

Specifically, we partition the graphical model into two sub-modules: user preferences module and site-specific action module (separated by dashed line in Figure 3 (a)). As showed in the figure, the two sub-modules are linked only through site-specific user preferences P_i^q. Viewing each module as a hyper parameter and applying Gibbs sampling over the modules, we can learn the model by iteratively updating the modules separately. In the following subsections we discuss how we update the parameters in each modules.

4.3.1 Site-Specific Learning

Viewing from site's action module, the difference with original models is the distribution of user preferences. Such change can be viewed as a change of prior knowledge. In traditional models we use non-informative prior distribution for user preferences. Instead, when integrating with JUMA, we take advantage of the informative prior transferred from universal preference.

Text-Based Sites. The topic distribution P_i^w now has a prior distribution of $Dir(\sigma^w T^w U_i)$ instead of $Dir(\alpha)$. As we design the model by following the conjugate distributions as in traditional design, only the parameter is shifted while the form of the distribution is kept the same as in the original model. Therefore, the inferences of the parameters are mostly the same as in original model thus we skip the details here. By plugging the new prior parameter, the posterior distribution of the latent topic variable z_{ij} is:

$$\begin{aligned} &P(z_{ij} = k | z_{\neg ij}, A^w, U, T^w, \cdot) \\ &\propto \frac{n_{ik}^{\neg ij} + \sigma^w U_i T_k^w}{\sum_{k'}(n_{ik'}^{\neg ij} + \sigma^w U_i T_{k'}^w)} \cdot \frac{m_{kw_{ij}}^{\neg ij} + \beta_{w_{ij}}^w}{\sum_{w'} m_{kw'}^{\neg ij} + \beta_{w'}^w} \end{aligned} \quad (6)$$

where n_{ik} is the number of times topic k being assigned to user i ($\#z_{i*} = k$) and m_{kj} is the number of times word j being assigned to topic k. For simplicity, we abbreviate the conditioned variables by '$\cdot$'. After sufficient sampling iterations, the parameters P^w, ϕ^w can be estimated by the following:

$$\hat{P}_{ik}^w = \frac{n_{ik} + \sigma^w U_i T_k^w}{\sum_{k'} n_{ik'} + \sigma^w U_i T_{k'}^w}, \;\; \hat{\phi}_{kj}^w = \frac{m_{kj} + \beta_j^w}{\sum_{j'} m_{kj'} + \beta_{j'}^w} \quad (7)$$

Item-Based Sites. User preference P_i^d now has a prior of $\mathcal{N}(T^d U_i, \sigma^d)$ instead of $\mathcal{N}(0, \sigma^d)$. Still, only parameters are shifted. The loss function is now:

$$\begin{aligned} \mathcal{L}(P^d, \phi^d | U, T^d, \cdot) = &\prod_i \mathcal{N}(P_i^d | T^d U_i, \sigma^d) \\ &\times \prod_{A_{ik}^d = 1} \sigma(P_i^d \phi_k^d) \prod_{A_{ik}^d = 0} (1 - \sigma(P_i^d \phi_k^d)) \end{aligned} \quad (8)$$

We employ gradient descend technique, a widely used method for matrix factorization, to maximize the log likelihood. The partial gradient for user's site-specific preferences P_i^d and the movie's latent factor ϕ_k^d are as follows: As the gradient calculation is rather straight forward and is similar with the original MF, thus we omit the details here.

4.3.2 User Preference Learning

In this module, we need to update users' universal preferences $\{U_i\}$ as well as the preference transferring parameters T^d, T^w. We still follow Gibbs sampling and update them individually according to their posterior distribution conditioned on the others.

The posterior distribution of user's universal preferences U_i according to the general model is:

$$P(U_i|T, P, U_{-i}, \cdot) \propto \mathcal{N}(U_i|0, \alpha) \cdot \prod_q P(P_i^q|U_i, T^q, \sigma^q) \quad (9)$$

where $P(P_i^q|U_i, T^q, \sigma^q)$ indicates probability of user having P_i^q as preference in site q under prior transferred from U_i and T^q. When applying to text-based and item-based sites as stated in previous section, we have:

$$\begin{aligned} P(U_i|\cdot) \propto & \mathcal{N}(U_i|0, \alpha) \cdot \\ & \mathcal{N}(P_i^d|T^d U_i, \sigma^d) \cdot Dir(P_i^w|\sigma^w T^w U_i) \end{aligned} \quad (10)$$

The posterior distribution for preference transfer model's parameter T^q for general scenario is:

$$P(T^q|U, P, \cdot) \propto P(T^q|\gamma^q) \cdot \prod_i P(P_i^q|U_i, T^q, \sigma^q) \quad (11)$$

where $P(T^q|\gamma^q)$ is the non-informative prior for T^q and $P(P_i^q|U_i, T^q, \sigma^q)$ is the same as in Eq. (9). Plugging the setting for text-based and item-based sites (normal distribution prior for T^w, T^d and the corresponding $P(P_i^q|U_i, T^q, \sigma^q)$), we have:

$$\begin{aligned} P(T^d|U, P, \cdot) &\propto \mathcal{N}(T^d|0, \gamma^d) \cdot \prod_i \mathcal{N}(P_i^d|T^d U_i, \sigma^d) \\ P(T^w|U, P, \cdot) &\propto \mathcal{N}(T^w|0, \gamma^w) \cdot \prod_i Dir(P_i^q|\sigma^w T^w U_i) \end{aligned} \quad (12)$$

Based on these posterior distribution, Gibbs sampling with Metropolis-Hasting [4] or gradient descend technique can be applied for the learning. For simplicity of inference in text-based sites cases, when updating user preference modules we may assume the site-specific preference P_i^q follows multivariate normal distribution with same mean and variance of the original distribution.

5. EXPERIMENTS

5.1 Experiment Settings

We conduct the experiments using real data over 148,470 aligned users from Weibo and Douban, with microblog retweeting histories and movie rating histories as user actions in the two sites respectively. Details are discussed in Section 3.1.

5.1.1 Evaluation Methodology

As there is no ground truth for user preferences, we can only evaluate the quality of user modeling by the performance of recommender systems.

We design experiments for both text-based site and item-based site to show that joint user modeling can transfer the user preferences and improve the performances in both directions. Specifically, we implement microblog and movie recommender systems for Weibo and Douban respectively using both JUMA and the existing state-of-the-art user modeling techniques as the underlying user preferences modeling.

For ground truth, we consider the user is interested in the movie if he/she rates it, and similar for microblogs. We use a portion (40% to 90%) of the user actions for training the model and the rest for testing. The recommender system ranks the potential movies/microblogs (the ones that user has no action with in the training data) according to the estimated user preferences. Then we employ Area Under the Curve (AUC) as the metric to evaluate the resulted rank.

5.1.2 Comparing Algorithms

For the text-based site Weibo, we compare the followings:

- **LDA:** Latent Dirichlet Allocation, a generative topic model [2]. We apply it to capture topic distributions of users and microblogs, and then employ KL-divergence to calculate the pairwise similarities.
- **CTR:** The collaborative tweet ranking model proposed in [3]. It integrates topic features as well as additional side information such as retweet count, author profile, post time and etc.
- **JUMA:** Our approach in this paper.
- **JUMA$^+$:** JUMA with additional side information.

Note that LDA and JUMA use only the textual information while CTR and JUMA$^+$ also employ side information. For fair comparison, we should compare them accordingly.

For the item-based site Douban, we compare:

- **PMF:** Probabilistic Matrix Factorization, a widely used approach for item-based recommendation [22].
- **SVD++:** the state-of-the-art matrix factorization approach for movie recommendation [14].
- **TMF:** Topic-Based Matrix Factorization, explicitly adding user's topic distribution in Weibo as features into feature-based matrix factorization.
- **mmTM:** Multi-Modal Topic Model, a naive approach for joint user modeling proposed in Section 3 for preliminary analysis.
- **JUMA:** our approach in this paper.

For all approaches, latent factor dimension for user preferences is set to 10 and number of topic is 20. All hyper-parameters are set to 1. The results are conducted over 5-fold cross validation.

There is no suitable cross-domain recommendation techniques to compare with under this experiment setting. As we discussed in Section 2, most existing cross-domain works focus on transferring user preferences between homogeneous user actions, thus can not apply for our setting between microblogging site and movie rating site. There are works modeling between text-based and item-based actions [21], however they require aligned actions for the model learning. For ComSoc in [34], although it also leverages multiple platform for joint user modeling, it focuses on borrowing the social relations instead of user generated contents thus also not suitable for comparing.

5.2 Experiment Results

5.2.1 General Scenario

We first evaluate in the general scenario. We vary the training ratio from 40% to 90% and show the corresponding results in Table 3.

For the text-based site Weibo, our approach JUMA achieve relative improvements of 3.8% and 4.7% comparing to LDA when training ratio is 90% and 40% respectively. The results indicate JUMA out-performs comparing method and is more robust to training size. For advanced and realistic

Table 3: Experimental Results, Varying Training Information Ratio

TARGET	ALGS	AUC SCORE, VARYING TRAINING INFORMATION RATIO					
		0.4	0.5	0.6	0.7	0.8	0.9
Text-Based (Weibo)	LDA	0.6514 ± 0.0017	0.6605 ± 0.0015	0.6694 ± 0.0016	0.6769 ± 0.0018	0.6839 ± 0.0015	0.6928 ± 0.0014
	JUMA	$\mathbf{0.6824 \pm 0.0014}$	$\mathbf{0.6892 \pm 0.0016}$	$\mathbf{0.6976 \pm 0.0014}$	$\mathbf{0.7058 \pm 0.0017}$	$\mathbf{0.7120 \pm 0.0012}$	$\mathbf{0.7194 \pm 0.0013}$
	CTR	0.7021 ± 0.0021	0.7133 ± 0.0017	0.7262 ± 0.0018	0.7352 ± 0.0017	0.7432 ± 0.0016	0.7532 ± 0.0015
	JUMA$^+$	$\mathbf{0.7338 \pm 0.0017}$	$\mathbf{0.7420 \pm 0.0015}$	$\mathbf{0.7502 \pm 0.0018}$	$\mathbf{0.7592 \pm 0.0015}$	$\mathbf{0.7670 \pm 0.0015}$	$\mathbf{0.7743 \pm 0.0014}$
Item-Based (Douban)	PMF	0.7275 ± 0.0016	0.7323 ± 0.0013	0.7384 ± 0.0015	0.7428 ± 0.0016	0.7485 ± 0.0014	0.7521 ± 0.0013
	SVD++	0.7856 ± 0.0012	0.7929 ± 0.0010	0.7986 ± 0.0016	0.8055 ± 0.0012	0.8089 ± 0.0013	0.8112 ± 0.0011
	TMF	0.7872 ± 0.0015	0.7946 ± 0.0012	0.8001 ± 0.0013	0.8071 ± 0.0019	0.8102 ± 0.0014	0.8132 ± 0.0012
	mmTM	0.6929 ± 0.0019	0.6940 ± 0.0011	0.6943 ± 0.0012	0.6963 ± 0.0015	0.7034 ± 0.0010	0.7064 ± 0.0018
	JUMA	$\mathbf{0.8127 \pm 0.0017}$	$\mathbf{0.8172 \pm 0.0016}$	$\mathbf{0.8219 \pm 0.0013}$	$\mathbf{0.8235 \pm 0.0011}$	$\mathbf{0.8243 \pm 0.0015}$	$\mathbf{0.8259 \pm 0.0013}$

scenario where side information is available, JUMA$^+$ also out competes TM$^+$.

For the item-based site Douban, JUMA achieves better performance comparing to both within-site approachs (PMF and SVD++) and joint modeling extended from existing approaches (TMF and mmTM). Also note that topic-based matrix factorization (TMF) has only slight improvement over traditional MF. By detailed analysis we find that topic-related feature's impact does not emerge in tMF because the signals coming from topic distributions are not as strong as the action histories within the same site.

By these experiments, we show that topic model and matrix factorization both can achieve better performance when integrating with JUMA. For other unevaluated techniques, we believe that similar improvement can also be achieved.

5.2.2 Cold-Start Scenario

Recall that cold-start problem is the potential failure of recommender system when dealing with new users with few or even no actions. We simulate cold-start scenarios by limiting the number of historic actions used for training. We depict the relative AUC improvements over users with different number of training actions (different cold-start level) in Figure 4. Results in both sites show that the performance improvement is much higher when dealing with cold users comparing to non-cold users. We achieve a relative improvement of 12.8% for the users with no historic actions in Weibo and 9.3% in Douban, indicating that JUMA successfully leverages cross-site action histories to alleviate the cold-start problem without special treatment.

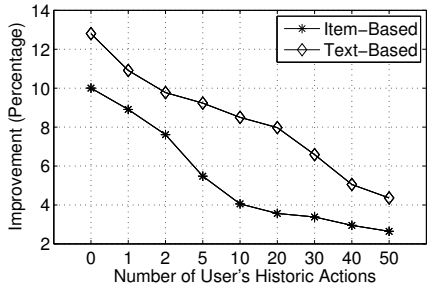


Figure 4: Experiment Result for Cold Start Scenario

5.2.3 Comprehensiveness of User Modeling

Now we focus on opposite direction of cold-start where users have more actions than average. We show the results in Table 4, indicating JUMA still outperforms comparing algorithms. For those users, the traditional approaches should be able to precisely model user's preference in the target site. The advantage of JUMA is comprehensiveness. For example, it is hard to model one's political preferences based on his movie histories, but is rather simple based on his microblogs. Such information might be helpful when recommending political related movies/documentaries.

Table 4: Non-Cold Users Scenarios

TARGET	ALGS	USER HISTORY COUNT			
		100	150	200	300
Text-Based (Weibo)	LDA	0.6952	0.7041	0.7105	0.7192
	JUMA	0.7205	0.7259	0.7342	0.7428
	CTR	0.7625	0.7662	0.7715	0.7796
	JUMA$^+$	0.7792	0.7846	0.7895	0.7930
Item-Based (Douban)	PMF	0.7558	0.7597	0.7632	0.7684
	SVD++	0.8142	0.8193	0.8235	0.8327
	TMF	0.8165	0.8204	0.8255	0.8342
	mmTM	0.7130	0.7185	0.7243	0.7304
	JUMA	0.8293	0.8337	0.8379	0.8436

6. CONCLUSION & FUTURE WORK

In this paper, we aim at improving comprehensiveness and accuracy of user modeling by joint user modeling, i.e. simultaneously modeling user actions in multiple aligned heterogeneous sites. We make the assumption that user's underlying preferences are consistent across-sites and further conduct analysis on real data to support it. We propose a modularized probabilistic graphical model framework JUMA for joint user modeling over aligned sites, which can integrate state-of-the-art site-specific approaches for a united user modeling. We also collect and publish a large scale data set from Weibo and Douban. Based on the data set, we conduct extensive experiments to evaluate JUMA's performance in various scenarios. Results show that JUMA out performs existing works and can alleviate cold-start problem without special treatment. For future works, we may consider integrating social relations into JUMA, and developing interpretation or visualization for user's underlying universal preferences.

7. REFERENCES

[1] S. Berkovsky, T. Kuflik, and F. Ricci. Cross-domain mediation in collaborative filtering. In *User Modeling 2007*, pages 355–359. Springer, 2007.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JML*, 2003.

[3] K. Chen, T. Chen, G. Zheng, O. Jin, E. Yao, and Y. Yu. Collaborative personalized tweet recommendation. In *SIGIR*, 2012.

[4] S. Chib and E. Greenberg. Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4):327–335, 1995.

[5] I. Fernández-Tobías, I. Cantador, M. Kaminskas, and F. Ricci. Cross-domain recommender systems: A survey of the state of the art. In *SCIR*, 2012.

[6] H. Gao, J. Tang, and H. Liu. Addressing the cold-start problem in location recommendation using geo-social correlations. *Data Mining and Knowledge Discovery*, pages 1–25, 2014.

[7] F. Godin, V. Slavkovikj, W. De Neve, B. Schrauwen, and R. Van de Walle. Using topic models for twitter hashtag recommendation. In *WWW*, 2013.

[8] N. Golbandi, Y. Koren, and R. Lempel. Adaptive bootstrapping of recommender systems using decision trees. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 595–604. ACM, 2011.

[9] F. Hu and Y. Yu. Interview process learning for top-n recommendation. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 331–334. ACM, 2013.

[10] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff. Identifying users across social tagging systems. In *ICWSM*, 2011.

[11] M. Jamali and M. Ester. A matrix factorization technique with trust propagation for recommendation in social networks. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 135–142. ACM, 2010.

[12] H. Kautz, B. Selman, and M. Shah. Referral web: combining social networks and collaborative filtering. *Communications of the ACM*, 40(3):63–65, 1997.

[13] N. Koenigstein, G. Dror, and Y. Koren. Yahoo! music recommendations: modeling music ratings with temporal dynamics and item taxonomy. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 165–172. ACM, 2011.

[14] Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 426–434. ACM, 2008.

[15] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, pages 30–37, 2009.

[16] X. N. Lam, T. Vu, T. D. Le, and A. D. Duong. Addressing cold-start problem in recommendation systems. In *Proceedings of the 2nd international conference on Ubiquitous information management and communication*, pages 208–211. ACM, 2008.

[17] G. Linden, B. Smith, and J. York. Amazon. com recommendations: Item-to-item collaborative filtering. *Internet Computing, IEEE*, 7(1):76–80, 2003.

[18] H. Liu and P. Maes. Interestmap: Harvesting social network profiles for recommendations. *Beyond Personalization-IUI*, page 56, 2005.

[19] J. Liu, F. Zhang, X. Song, Y.-I. Song, C.-Y. Lin, and H.-W. Hon. What's in a name?: an unsupervised approach to link users across communities. In *Proceedings of the 6th ACM international conference on Web search and data mining*, pages 495–504. ACM, 2013.

[20] S. Liu, S. Wang, F. Zhu, J. Zhang, and R. Krishnan. Hydra: Large-scale social identity linkage via heterogeneous behavior modeling. In *SIGMOD*, 2014.

[21] J. McAuley and J. Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *RecSys*, pages 165–172. ACM, 2013.

[22] A. Mnih and R. Salakhutdinov. Probabilistic matrix factorization. In *NIPS*, pages 1257–1264, 2007.

[23] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *Security and Privacy, 2009 30th IEEE Symposium on*, pages 173–187. IEEE, 2009.

[24] A. Narayanan and V. Shmatikov. Myths and fallacies of personally identifiable information. *Communications of the ACM*, 53(6):24–26, 2010.

[25] S. Sahebi and P. Brusilovsky. Cross-domain collaborative recommendation in a cold-start context: The impact of user profile size on the quality of recommendation. In *UMAP*. 2013.

[26] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock. Methods and metrics for cold-start recommendations. In *SIGIR*, pages 253–260, 2002.

[27] S. Tan, J. Bu, X. Qin, C. Chen, and D. Cai. Cross domain recommendation based on multi-type media fusion. *Neurocomputing*, 127:124–134, 2014.

[28] S. Tan, Z. Guan, D. Cai, X. Qin, J. Bu, and C. Chen. Mapping users across networks by manifold alignment on hypergraph. In *AAAI*, 2014.

[29] J. Tang, S. Wu, J. Sun, and H. Su. Cross-domain collaboration recommendation. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1285–1293. ACM, 2012.

[30] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

[31] J. Zhang, X. Kong, and P. S. Yu. Predicting social links for new users across aligned heterogeneous social networks. In *2013 IEEE 13th International Conference on Data Mining*, pages 1289–1294. IEEE, 2013.

[32] Y. Zhang and J. Koren. Efficient bayesian hierarchical user modeling for recommendation system. In *SIGIR*, pages 47–54. ACM, 2007.

[33] Z.-K. Zhang, C. Liu, Y.-C. Zhang, and T. Zhou. Solving the cold-start problem in recommender systems with social tags. *EPL*, 92(2):28002, 2010.

[34] E. Zhong, W. Fan, J. Wang, L. Xiao, and Y. Li. Comsoc: adaptive transfer of user behaviors over composite social network. In *Proceedings of the 18th ACM SIGKDD*, pages 696–704. ACM, 2012.

[35] K. Zhou, S.-H. Yang, and H. Zha. Functional matrix factorizations for cold-start recommendation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 315–324. ACM, 2011.