Predicting User Behavior in Display Advertising via Dynamic Collective Matrix Factorization

Sheng Li*
Northeastern University
Boston, MA, USA
shengli@ece.neu.edu

Jaya Kawale
Adobe Research
San Jose, CA, USA
kawale@adobe.com

Yun Fu
Northeastern University
Boston, MA, USA
yunfu@ece.neu.edu

ABSTRACT

Conversion prediction and click prediction are two important and intertwined problems in display advertising, but existing approaches usually look at them in isolation. In this paper, we aim to predict the conversion response of users by jointly examining the past purchase behavior and the click response behavior. Additionally, we model the temporal dynamics between the click response and purchase activity into a unified framework. In particular, a novel matrix factorization approach named the dynamic collective matrix factorization (DCMF) is proposed to address this problem. Our model considers temporal dynamics of post-click conversions and also takes advantages of the side information of users, advertisements, and items. Experimental results on a public dataset and a real-world marketing dataset show that our model achieves significant improvements over several baselines.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering

Keywords

Conversion prediction, matrix factorization, temporal dynamic

1. INTRODUCTION

With the proliferation of the Internet and online shopping websites, digital marketing has become an effective way to reach out to consumers. Typically consumers can be influenced via targeted advertisements (ads) on the web to make purchases. Display advertising is a popularly used medium for targeting users and allows advertisers to place graphical ads on the publishers web pages. Some ads (e.g., brand ads) aim to create an awareness of the brand that could possibly lead to potential increase in sales in the future, but most of the ads (e.g., direct response ads) aim to create an impulse leading to a sale right now.

Traditionally the *click-through rate* (CTR) has been used as a central measure for evaluating the performance of a direct response ad campaign and the focus of the publisher so far has been to maximize the number of clicks on an ad. An alternate strategy that has gained attention in the recent past is to maximize the *conversion rate* (CVR) instead of just ad-clicks as many advertisers would prefer not to pay for an ad impression unless it leads to a conversion. The conversion could either imply revenue generated via buying a product or could mean account creation. In both strategies it is critical to understand the user behavior and predict the response so as to have better targeting of the ads and as a result higher conversions.

So far, the problem of click prediction and conversion prediction has mainly been studied in isolation. Researchers have successfully applied novel strategies for click prediction and there has been significant work in the area [1, 2]. Recently the problem of conversion prediction has been studied by [3, 4] to analyze the ad campaigns. However, the objectives of the two problems are often intertwined together and there is a need to study the two problems in conjunction with each other. With the notable exception of [5], there is not much work analyzing the two objectives together to understand the user purchase behavior better. Jointly studying the two problems can help us understand the pathway leading to conversion and can provide answers to several questions. For example, *What ads should be shown to a particular user so that he generates revenue ?* and *Will a given user generate revenue ?*

In this paper, we present a novel approach called as dynamic collective matrix factorization (DCMF) to jointly examine the user click behavior and the purchase activity. The DCMF model is a substantial extension of the collective matrix factorization model used to jointly factorize multiple matrices [6]. Apart from considering the two matrices of click and purchase behavior together as in a CMF model, our model also takes into account the temporal influence of an ad impression/click on conversion. We model the time dependency into the matrix factorization framework to account for the decay in the influence of an ad. An efficient optimization algorithm is devised to solve the problem. Our approach is well suited for an interactive setting where the user preferences and behavior change over time.

Our key contributions can be summarized as follows:

*Part of the work is completed during Sheng Li's internship at Adobe Research.

SIGIR 2015

DOI: http://dx.doi.org/10.1145/2766462.2767781.

- We propose a dynamic collective matrix factorization (DCMF) approach for conversion prediction, which jointly models the temporal relationships between click events and purchase events in display advertising. It also incorporates the side information of users, ads and items via regression.
- Our framework is based upon examining the data in time slices to account for the decayed influence of an ad and we use stochastic gradient descent for optimization. This makes the framework well suited for interactive settings as well as large datasets.
- We evaluate the conversion prediction performance of our approach on real-world datasets and show that our model performs better as compared to the several baseline methods.

2. RELATED WORK

In this section, we briefly review the related works on response prediction and temporal prediction models.

Response Prediction. Response prediction in digital marketing has been widely studied in recent years. Most of the prior work focuses on predicting the click-through-rate (CTR) of online ads or other contents [1, 2]. The ultimate goal of display advertising is conversion. However, there are only a few works that investigate the conversion prediction problem [3, 4]. Lee *et al.* estimated the conversion rate by using the past performance observations along with user, publisher and advertiser data hierarchies [3]. Rosales *et al.* focus on estimating the post-click conversion rate, considering the context of users and pages [4]. Unlike these conversion prediction methods, our approach take advantage of the click response and side information via the collective matrix factorization technique.

The most relevant method in the literature is the hierarchical multi-task learning algorithm presented in [5]. It jointly models the conversion, click and unattributed-conversion problems. There are significant differences between [5] and our work. First, we make use of the explicit temporal relationship between click and purchase events, which is ignored in [5]. Second, unlike the multi-task learning model used in [5], we develop the conversion prediction model based on matrix factorization.

Temporal Prediction Models. The prediction model considering temporal dynamics was first developed for collaborative filtering in [7]. In [8], a dynamic matrix factorization method was presented to model the temporal adoption effects in collaborative filtering.

The idea of using temporal dynamics has been explored in online advertising. Barajas *et al.* proposed a time series approach for evaluating the effectiveness of display advertising [9]. Most recently, Oentaryo *et al.* designed a hierarchical importance-aware factorization machine (HIFM) for click response prediction in mobile advertising [10]. It is able to handle the temporal ad response data. Unlike HIFM, our approach aims to tackle the conversion prediction problem, and model the temporal relationships between click and purchase events.

3. APPROACH

3.1 Modeling Temporal Dynamics

We deal with the conversion prediction problem using the collaborative filtering (CF) technique. The fundamental intuition behind CF is that the "similar" users will have "similar" preferences. There has been work on applying the CF technique to conversion prediction [11], but it models the interactions between pages and ads. In our paper, we directly model the relationships between users and ads/items, which enables us to predict the user behaviors. The intuition is that "similar" users are very likely to click "similar" ads and purchase "similar" items. Figure 1 illustrates our framework.

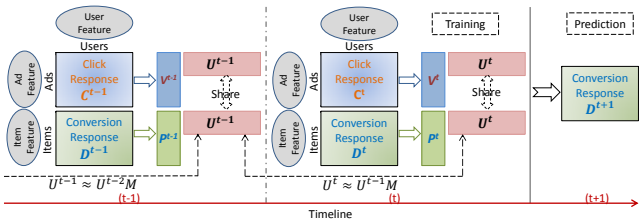

Figure 1: Framework of DCMF approach.

First, we want to jointly analyze the relational data, i.e., the click response and purchase activities of users. Inspired by the collective matrix factorization (CMF) [6], we factorize the click response matrix (User∼Ads) and the purchase activity matrix (User∼Item) simultaneously.

Secondly, we aim to model the temporal information, which is critical in attributing the conversions to the ad clicks. A key observation is that, *the behavior of users may change over time*. For example, if a user has already purchased an item in the previous week, it is unlikely that he/she will purchase the same item again in the next week. Also, the *influence* of an ad impression or a click may only last for a short span of time. Therefore, we incorporate temporal information into CMF.

Thirdly, we need to ensure that the latent features of users do not dramatically change in a short period of time, as in reality the user preferences would evolve smoothly. To address this concern, we leverage the latent features of the users learned in time $t-1$.

Given T pre-defined time slices $t \in \{1, 2, \cdots, T\}$, we use $C^t \in \mathbb{R}^{N_u \times N_a}$ and $D^t \in \mathbb{R}^{N_u \times N_p}$ to denote the click responses and purchase activities from N_u users to N_a ads (or N_p items) in the time slice t, respectively. By exploiting the temporal relationships between click response and purchase events, we notice that the purchase events in time $t+1$ are mainly related to the click events in time t and hence our model needs to account for that. The objective function is:

$$\begin{aligned} \arg\min_{\substack{U^t, V^t, \\ P^t, M}} \quad & f(U^t, V^t, P^t, M) = \alpha \|W^{Ct} \odot (C^t - U^t V^{t\mathrm{T}})\|_{\mathrm{F}}^2 \\ & + (1-\alpha)\|W^{Dt} \odot (D^t - U^t P^{t\mathrm{T}})\|_{\mathrm{F}}^2 \\ & + \lambda_1 \|U^t - U^{t-1}M\|_{\mathrm{F}}^2 \\ & + \lambda_2 (\|U^t\|_{\mathrm{F}}^2 + \|V^t\|_{\mathrm{F}}^2 + \|P^t\|_{\mathrm{F}}^2 + \|M\|_{\mathrm{F}}^2), \end{aligned} \tag{1}$$

where W^{Ct} and W^{Dt} are boolean matrices that indicate the training samples in C^t and D^t, respectively. U^t, V^t and P^t are latent factors; M is a linear transition matrix; $\odot$ denotes the entry-wise product; α, λ_1 and λ_2 are trade-off parameters.

The first two terms in (1) denote the approximation errors, and the last four terms are regularizations used to prevent overfitting. In (1), the smooth transition of user latent factors are modeled based on the assumption:

$$U^t \approx U^{t-1}M, \tag{2}$$

where U^{t-1} is the latent features of users learned from the previous time slice $t-1$. We assume that the latent features

in time t are closely related to the feature in time $t-1$, which is reasonable in real applications. M is a transition matrix of users' behavior, which tries to capture the mappings between users' behavior in two successive time slices. The intuition is that users' intention on purchasing items should be smoothly transited over time.

3.2 Modeling Side Information

So far, we have seen that the model in (1) is not aware of side information, i.e., the features of users, ads, and items. We can further exploit the additional information to improve the prediction performance. The side information is also particularly useful as the data in conversion and click prediction problems are generally sparse. For example, we do not have any click responses or conversion responses of some new users, which lead to the *cold-start* problem. In this case, the latent features of new users estimated by (1) are not reliable anymore. However, side information provide useful cues from another perspective, and make it possible to learn robust latent features in the cold-start scenario. In this section, we incorporate the side information into (1), and present the DCMF method.

Let X, Y and Z denote the feature matrices for users, ads and items, respectively. We assume that the click response and purchase activity are generated by the inner product of latent factors, and the side information via linear regression. Thus, the matrix approximation can be written as:

$$\begin{aligned} C^t &\approx U^t V^{t\mathrm{T}} + \hat{U}^t Y^{\mathrm{T}} + X\hat{V}^{t\mathrm{T}}, \\ D^t &\approx U^t P^{t\mathrm{T}} + \hat{U}^t Z^{\mathrm{T}} + X\hat{P}^{t\mathrm{T}}, \end{aligned} \tag{3}$$

where $\hat{U}^t$, $\hat{V}^t$ and $\hat{P}^t$ are regression coefficients on user features, ad features and item features, respectively. We treat the three terms used to approximate C^t (or D^t) equally for simplicity. The performance can be enhanced by assigning different weights for them.

By replacing the matrix approximations in (1) with (3), we can then rewrite the objective function as:

$$\begin{aligned} \arg \min_{\substack{U^t,V^t,P^t,M,\\ \hat{U}^t,\hat{V}^t\hat{P}^t}} & \alpha\|W^C \odot (C^t - U^t V^{t\mathrm{T}} - \hat{U}^t Y^{\mathrm{T}} - X\hat{V}^{t\mathrm{T}})\|_{\mathrm{F}}^2 \\ & +(1-\alpha)\|W^D \odot (D^t - U^t P^{t\mathrm{T}} - \hat{U}^t Z^{\mathrm{T}} - X\hat{P}^{t\mathrm{T}})\|_{\mathrm{F}}^2 \\ & +\lambda_1\|U^t - U^{t-1}M\|_{\mathrm{F}}^2 + \lambda_2(\|U^t\|_{\mathrm{F}}^2 + \|V^t\|_{\mathrm{F}}^2 + \|P^t\|_{\mathrm{F}}^2 \\ & +\|\hat{U}^t\|_{\mathrm{F}}^2 + \|\hat{V}^t\|_{\mathrm{F}}^2 + \|\hat{P}^t\|_{\mathrm{F}}^2 + \|M\|_{\mathrm{F}}^2). \end{aligned} \tag{4}$$

With the learned latent factors, we can predict the conversion score of user m for item n at time $t+1$ as:

$$Score(m, n, t+1) = u_m^t p_n^{t\mathrm{T}} + \hat{u}_m^t z_n^{\mathrm{T}} + x_m \hat{p}_n^{t\mathrm{T}}, \tag{5}$$

where u_m^t and $\hat{u}_m^t$ are the m-th row of U^t and $\hat{U}^t$, p_n^t and $\hat{p}_n^t$ are the n-th row of P^t and $\hat{P}^t$, respectively.

3.3 Optimization

As the stochastic gradient descent (SGD) algorithm is efficient in practice, we utilize SGD to solve (4).

First, we fix M, and update other variables, $U^t = \{u_1^t, \cdots, u_r^t\}$, $V^t = \{v_1^t, \cdots, v_r^t\}$, $P^t = \{p_1^t, \cdots, p_r^t\}$, $\hat{U}^t = \{\hat{u}_1^t, \cdots, \hat{u}_r^t\}$, $\hat{V}^t = \{\hat{v}_1^t, \cdots, \hat{v}_r^t\}$, and $\hat{P}^t = \{\hat{p}_1^t, \cdots, \hat{p}_r^t\}$. For simplicity, we use f to denote the objective in (4). After selecting a pair of random training points C_{ij}^t and D_{ik}^t, we only need to update u_i^t, v_j^t, p_k^t, $\hat{u}_i^t$, $\hat{v}_j^t$ and $\hat{p}_k^t$ using:

$$\begin{aligned} u_i^t &= u_i^t - \gamma\tfrac{\partial}{\partial u_i^t} f, v_j^t = v_j^t - \gamma\tfrac{\partial}{\partial v_j^t} f, p_k^t = p_k^t - \gamma\tfrac{\partial}{\partial p_k^t} f, \\ \hat{u}_i^t &= \hat{u}_i^t - \gamma\tfrac{\partial}{\partial \hat{u}_i^t} f, \hat{v}_j^t = \hat{v}_j^t - \gamma\tfrac{\partial}{\partial \hat{v}_j^t} f, \hat{p}_k^t = \hat{p}_k^t - \gamma\tfrac{\partial}{\partial \hat{p}_k^t} f, \end{aligned} \tag{6}$$

Table 1: RMSE and MAE (with standard deviation) on Movielens-1M dataset.

Method	RMSE	MAE
PMF [12]	0.9168 (0.0019)	0.7197 (0.0013)
LIBMF [13]	0.9161 (0.0017)	0.7185 (0.0014)
SVDFeature [14]	0.9152 (0.0021)	0.7163 (0.0017)
HBMFSI [15]	0.9089 (0.0025)	0.7146 (0.0016)
CMF [6]	0.9145 (0.0017)	0.7156 (0.0012)
DCMF (Ours)	**0.9018 (0.0027)**	**0.7082 (0.0019)**

where γ is the learning rate. The detailed gradients for each variable are omitted here due to the space limit.

Next, we fix all the other variables, and update M. By ignoring all the irrelevant terms with respect to M, the objective (4) reduces to:

$$\arg\min_M f(M) = \lambda_1\|U^t - U^{t-1}M\|_{\mathrm{F}}^2 + \lambda_2\|M\|_{\mathrm{F}}^2. \tag{7}$$

We can then update M using:

$$M = M - \gamma\tfrac{\partial}{\partial M} f(M), \tag{8}$$

where gradient is $\frac{\partial f}{\partial M} = -\lambda_1 U^{(t-1)\mathrm{T}}(U^t - U^{t-1}M) + \lambda_2 M$.

The above process is repeated until convergence.

4. EXPERIMENTS

In this section, we first compare our approach and baselines on the public dataset. Then we evaluate the conversion prediction performance on a proprietary dataset.

4.1 Experiments on Public Data

Data and Settings. We choose a public benchmark dataset Movielens-1M [1], which is widely used to evaluate matrix factorization methods in collaborative filtering. It contains 1 million ratings with 6040 users and 3952 movies, and provides side information for both users and items. In the experiments, the side information are encoded into binary valued vectors. The size of rating matrix R is 6040×3952. To testify the performance on jointly factorizing multiple matrices, we divide the matrix R into two parts, $R_a = R_{1\sim6040,1\sim1900}$ and $R_b = R_{1\sim6040,1901\sim3952}$. In this way, two rating matrices have the same set of users, but different items. We compare our approach with the following baselines: PMF [12], LIBMF [13], SVDFeature [14], HBMFSI [15], and CMF [6]. Two popular evaluation metrics are utilized, which are root mean squared error (RMSE) and mean absolute value (MAE).

Results. We aim to predict the ratings in R_b. The single matrix factorization methods (i.e., PMF, LIBMF, SVDFeature and HBMFSI) are trained only using the rating matrix R_b. CMF and our methods are trained using both R_a and R_b. The ratings in R_b are split into training set and test set according to the temporal information. Ten training/test sets are generated in total. Table 1 shows the average RMSE and MAE with standard deviations. Two collective factorization methods, CMF and ours, outperforms PMF, LIBMF and SVDFeature. By taking advantages of the side information, HBMFSI and our DCMF approach achieve much lower RMSE and MAE than all the other methods, implying that side information are very important especially when the data are sparse.

4.2 Conversion Prediction on Marketing Data

Data and Settings. To evaluate the performance on conversion prediction, we examine a subset of the market-

[1] `http://grouplens.org/datasets/movielens/`

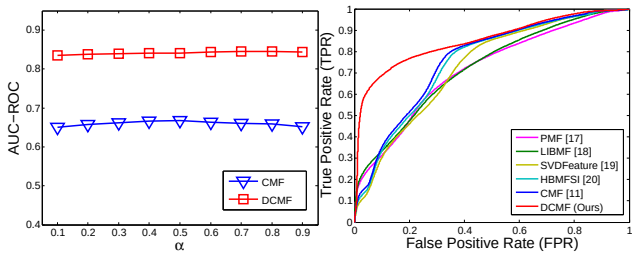

Figure 2: Left: AUC-ROC of CMF and DCMF in Case-3 with different α. Right: ROC curves in Case-1.

Table 2: AUC-ROC of each methods on marketing data.

Method	Case 1	Case 2	Case 3	Case 4	Case 5
PMF [12]	0.7286	0.7180	0.6451	0.7119	0.7143
LIBMF [13]	0.7305	0.7197	0.6532	0.7255	0.7229
SVDFeature [14]	0.7425	0.7286	0.6605	0.7308	0.7314
HBMFSI [15]	0.7498	0.7371	0.6689	0.7397	0.7412
CMF [6]	0.7511	0.7464	0.6678	0.7356	0.7369
DCMF (**Ours**)	**0.8504**	**0.8514**	**0.8411**	**0.8312**	**0.8372**

ing data from Oct. 1, 2013 $\sim$ Nov. 30, 2013. The dataset constitutes of behavioral characteristics of 448,158 number of users and 737 ads. Along with the impression records, we also have click and purchase activity information for all the users. We empirically choose one week as the time window t in our model. To construct the binary response tables, we denote the click events and purchase events as positive responses, and the impressions (without any following events) as negative responses. All the other entries are treated as missing values. As the click and purchase are rare events in reality, our data set is extremely sparse. To collect the side information, we select some features of users, ads and items, respectively. For each user, we encode the demographic information (e.g., country, state, domain) into a binary valued vector. The attributes of ads (e.g., advertiser, ad size) and items (e.g., type, price) are also encoded into binary vectors, respectively. To conduct fair comparisons, we set up 5 different training/test cases along the timeline. Each training set consists of a click events table and a purchase events table. Each test set only contains a table of purchase events, as our goal is to predict the conversions.

For PMF, LIBMF, SVDFeature and HBMFSI, we only use the purchase data for training, due to the intrinsic limitation of these methods. For CMF and our approach, we use both the click data and purchase data for training. In particular, we use the click events and purchase events at time t to predict the purchase events in time $t+1$ (i.e., D^{t+1}). Following [5], we use the ROC curve and the area under ROC (AUC-ROC) as our evaluation metrics.

For SGD based matrix factorization methods (e.g., PMF, LIBMF, SVDFeature and CMF), the major parameters are the learning rate γ and the trade-off parameter λ for regularization terms. For Bayesian method HBMFSI, we follow the settings in [15]. We sample a validation set from the training data, and tune these parameters empirically. The parameters γ, λ_1 and λ_2 are set to 0.003, 0.001 and 0.02, respectively. In CMF and our approach, another important parameter is α. To understand its sensitivity, we observe the AUC-ROC of CMF and DCMF approach with various choices of α. Figure 2 (left) shows the AUC-ROC values in Case-3. We can observe that our approach is not sensitive to the values of α. We achieve better performance when α falls into the range $[0.5, 0.8]$. In the following experiments, α is set to 0.6. In addition, the dimension of latent features is set to 20 for each method. To initialize our approach in Case-1, we borrow the latent factors of CMF learned from Oct. 1$\sim$Oct. 7 as the input U^{t-1}.

We evaluate the performance of each compared method in 5 training/test splits. Figure 2(right) shows the ROC in Case-1, and Table 2 lists the AUC-ROC of all cases. Form Figure 2 and Table 2, we make the following observations: (1) Incorporating side information improves the prediction results significantly. SVDFeature and HBMFSI employ the features of users and items, they achieve much better performance than PMF and LIBMF, which do not utilize any side information. Similarly, our DCMF approach performs very well by virtue of the side information. (2) Modeling click response is helpful in predicting purchase events, since CMF and our approach outperforms PMF and LIBMF in each case. (3) Temporal dynamics is critical in conversion prediction. Our approach outperform all the other competitors, as they leverage the temporal relationships between click and purchase events. In particular, our DCMF approach obtains better results than CMF, indicating that the latent features learned from previous time slice are useful. (4) DCMF achieves the best results in each case, which demonstrates that side information and temporal dynamic could be complementary to each other.

5. CONCLUSIONS

In this paper, we presented a novel matrix factorization approach DCMF for conversion prediction in display advertising. DCMF jointly examines the click events and purchase events in an online learning fashion by leveraging the temporal information and side information. Extensive experimental results on a public dataset and a real-world marketing dataset demonstrate the superiority of our approach over existing methods.

Acknowledgements

This research is supported in part by the NSF CNS award 1314484, ONR award N00014-12-1-1028, ONR Young Investigator Award N00014-14-1-0484, and U.S. Army Research Office Young Investigator Award W911NF-14-1-0218.

6. REFERENCES

[1] Chenyan Xiong, Taifeng Wang, Wenkui Ding, Yidong Shen, and Tie-Yan Liu. Relational click prediction for sponsored search. In *WSDM*, pages 493–502, 2012.

[2] Deepak Agarwal, Bo Long, Jonathan Traupman, Doris Xin, and Liang Zhang. Laser: a scalable response prediction platform for online advertising. In *WSDM*, pages 173–182, 2014.

[3] Kuang chih Lee, Burkay Orten, Ali Dasdan, and Wentong Li. Estimating conversion rate in display advertising from past performance data. In *KDD*, pages 768–776, 2012.

[4] Rómer Rosales, Haibin Cheng, and Eren Manavoglu. Post-click conversion modeling and analysis for non-guaranteed delivery display advertising. In *WSDM*, pages 293–302, 2012.

[5] Amr Ahmed, Abhimanyu Das, and Alexander J. Smola. Scalable hierarchical multitask learning algorithms for conversion optimization in display advertising. In *WSDM*, pages 153–162, 2014.

[6] Ajit Paul Singh and Geoffrey J. Gordon. Relational learning via collective matrix factorization. In *KDD*, pages 650–658, 2008.

[7] Yehuda Koren. Collaborative filtering with temporal dynamics. In *KDD*, pages 447–456, 2009.

[8] Freddy Chong Tat Chua, Richard Jayadi Oentaryo, and Ee-Peng Lim. Modeling temporal adoptions using dynamic matrix factorization. In *ICDM*, pages 91–100, 2013.

[9] Joel Barajas, Ram Akella, Marius Holtan, Jaimie Kwon, and Brad Null. Measuring the effectiveness of display advertising: a time series approach. In *WWW (Companion Volume)*, pages 7–8, 2011.

[10] Richard Jayadi Oentaryo, Ee-Peng Lim, Jia-Wei Low, David Lo, and Michael Finegold. Predicting response in mobile advertising with hierarchical importance-aware factorization machine. In *WSDM*, pages 123–132, 2014.

[11] Aditya Krishna Menon, Krishna Prasad Chitrapura, Sachin Garg, Deepak Agarwal, and Nagaraj Kota. Response prediction using collaborative filtering with hierarchies and side-information. In *KDD*, pages 141–149, 2011.

[12] Ruslan Salakhutdinov and Andriy Mnih. Probabilistic matrix factorization. In *NIPS*, 2007.

[13] Yong Zhuang, Wei-Sheng Chin, Yu-Chin Juan, and Chih-Jen Lin. A fast parallel sgd for matrix factorization in shared memory systems. In *RecSys*, pages 249–256, 2013.

[14] Tianqi Chen, Weinan Zhang, Qiuxia Lu, Kailong Chen, Zhao Zheng, and Yong Yu. Svdfeature: a toolkit for feature-based collaborative filtering. *Journal of Machine Learning Research*, 13:3619–3622, 2012.

[15] Sunho Park, Yong-Deok Kim, and Seungjin Choi. Hierarchical bayesian matrix factorization with side information. In *IJCAI*, 2013.