Item-Level Social Influence Prediction with Probabilistic Hybrid Factor Matrix Factorization

Peng Cui
Tsinghua National Lab. for Info. Science and Tech.
Department of Computer Science and Technology
Tsinghua University, Beijing 100084, China

Fei Wang
Healthcare Transformation Group
IBM T J Watson Research Center
Hawthorne, USA

Shiqiang Yang and **Lifeng Sun**
Department of Computer Science and Technology
Tsinghua University, Beijing 100084, China

Abstract

Social influence has become the essential factor which drives the dynamic evolution process of social network structure and user behaviors. Previous research often focus on social influence analysis in network-level or topic-level. In this paper, we concentrate on predicting *item-level* social influence to reveal the users' influences in a more fine-grained level. We formulate the social influence prediction problem as the estimation of a user-post matrix, where each entry in the matrix represents the social influence strength the corresponding user has given the corresponding web post. To deal with the sparsity and complex factor challenges in the research, we model the problem by extending the probabilistic matrix factorization method to incorporate rich prior knowledge on both user dimension and web post dimension, and propose the *Probabilistic Hybrid Factor Matrix Factorization* (PHFMF) approach. Intensive experiments are conducted on a real world online social network to demonstrate the advantages and characteristics of the proposed method.

Introduction

In social computing, people and information are two core dimensions and people sharing information (such as blog, news, album, etc.) is the basic behavior. Actually, the spreading out of information is because of the user sharing in social network. The owner of the information, e.g. the advertisers, hope to maximize the diffusion range of the information (Bao and Chang 2010). This goal makes them desire to target the influencers, who are able to let many friends to click the information they share or even share further to extend the sharing cascades.

Psychologically, people share information with their friends mainly because they want to build their reputations and help others, in which *to influence others* is the important motivation for sharing (Wasko and Faraj 2005). According to the definition of social influence on WIKI, social influence occurs when "*an individual's thoughts, feelings or actions are affected by other people*". This phenomenon is often observed in online social networks like FACEBOOK and TWITTER, where users are often *influenced* to visit, comment, or even forward a web post after friends share it. The aim of this paper is to predict the number of clicks by friends if a user shares a web post, which is defined as the *item-level social influence*.

It should be noted that the item-level social influence is user-post specific. Different from most of the existing research works focusing on users' overall social influence analysis (Newman 2003)(Strogatz 2003) and topical social influence mining (Tang et al. 2009), item-level social influence is not a general measure on users, but on the interactions of users and posts. That is, we need to discriminate a user's social influences with respect to different web posts.

The item-level social influence prediction problem can be formulated as the estimation of a user-post matrix as in our previous paper (Cui et al. 2011), in which the (i, j)-th element represents the number of clicks by friends of user i on her j-th shared web post. Recently, a variety of probabilistic factor based models has been proposed to solve the problem (Marlin and Zemel 2004)(Marlin 2004). However, it is intractable for these models to make exact inference. Either very slow or inaccurate approximations are required to compute the posterior distribution over hidden factors in such models (Salakhutdinov and Mnih 2008).

In (Salakhutdinov and Mnih 2008), a probabilistic matrix factorization (PMF) method is presented, which scales linearly with the number of observations and performs well on large and imbalanced datasets. However, there are two challenges in introducing PMF for item-level social influence prediction:

- Sparsity. The interactions between users and web posts are extremely sparse compared with the total number of user-post pairs. The sparsity is even much more severe than Netflix dataset. According to our statistics of 34K users in the website `www.renren.com`, which is a Facebook style social network site in China, each user only shares 6 web posts in average during a month, compared with a total of 43K web posts; and each post is only shared by 4 users, compared with a total of 34K users.
- Complex factors. There are a volume of factors that affect how many friends will click a shared post, and provide potential clues for user and post grouping.

Thus, it is clear that we need subtle and effective prior knowledge and predictive factor selection for user and post grouping to alleviate the sparsity and complex factor problem. In this paper, We proposed a *Probabilistic Hybrid Fac-*

tor Matrix Factorization (PHF-MF) algorithm for item-level social influence modeling. In this model, we try to find out the common hidden vector space for both the users and the posts, where their multiplication can well approximate the observed training user-post matrix. Meanwhile, in order to alleviate the sparsity problem, we construct the priors on users and posts by incorporating the user-specific factors and post-specific factors, and apply gradient descent to solve the PHF-MF problem.

It is worthwhile to highlight the key contributions of this paper.

- We extend the Probabilistic Matrix Factorization model by introducing constraints on two dimensions of the interaction matrix in PHF-MF, which make it feasible to incorporate prior knowledge in a probabilistic matrix factorization framework.
- The proposed PHF-MF gives a probabilistic interpretation of the previously proposed Hybrid Factor Non-Negative Matrix Factorization (HF-NMF) in (Cui et al. 2011), which makes the model more theoretically solid.
- We conducted intensive experiments on real social network datasets, and the results show that the PHF-MF can achieve a better performance compared with other competitors.

Problem Formulation

First, to make the paper self-contained, we formally define the problem of item-level social influence prediction as (Cui et al. 2011). Suppose we have M users with the i-th user denoted as u_i and N postings with the j-th post denoted as p_j. We use $\mathcal{N}(u_i)$ to denote the collection of u_i's first-order friends (i.e. the nodes that directly link to u_i). As mentioned in the introduction, two key factors that need to be formulated in our model are:

- **Item-level social influence**: The strength of u_i's influence on $\mathcal{N}(u_i)$ given the web post p_j, denoted as f_{ij}, is defined the number of u_i's friends who clicked post p_j. We assume that the influence should be *specific* on each user-post pairs: (1) different users have different influence power to their friends; (2) different posts have different influence power (more intuitively, attraction) to users who are interested in; and (3) users' influences manifest differently for different posts. Therefore, only item-level social influence can reveal the users' real influence on friends, and the strength of influence should definitely be user-post specific.
- **Social influence prediction**: The social influence prediction is to predict the unobserved social influences $\hat{f_{ij}}$ based on the observed f_{ij}'s and those predictive factors. One issue that is worthy of emphasizing here is that the user factors and post factors are essential for the predictive modeling. On one hand, the user-post interactions are very sparse. We need to find effective factors to "group" those users and posts to alleviate the sparsity problem. On the other hand, the user and post-specific factors also provide some effective prior knowledge to complement the inference from pure user-post interactions.

With the above terminologies, we can formally define the task of item-level social influence prediction. We denote the user-post influence matrix as $\tilde{\mathbf{X}} \in \mathbb{R}^{M \times N}$, with its (i,j)-th entry

$$\tilde{X}_{ij} = \begin{cases} f_{ij} & \text{if } u_i \text{ shared } p_j \\ 0 & otherwise \end{cases} \quad (1)$$

If we use g_i to denote the number of u_i's friends (i.e. $g_i = |\mathcal{N}(u_i)|$, where $|.|$ is the cardinality of a collection), then $f_{ij} \leq g_i$.

To measure the influence strength of different users in the same scale, we propose the following *percentile* influence matrix

$$X_{ij} = \begin{cases} \frac{f_{ij}}{g_i} & \text{if } u_i \text{ shared } p_j \\ 0 & otherwise \end{cases} \quad (2)$$

so that X_{ij}'s are normalized into the range of $[0, 1]$.

The user-post influence matrix $\tilde{X}$ can be reconstructed by

$$\tilde{\mathbf{X}} = Diag(\mathbf{g}) \cdot \mathbf{X} \quad (3)$$

where $\mathbf{g} = [g_1, g_2, \cdots, g_N]^\top \in \mathbb{R}^N$, and $Diag(\mathbf{g})$ is the diagonal matrix with $\mathbf{g}$ on the diagonal line.

Our formulation of the item-level social influence prediction problem is quite different from existing works on social network analysis. First, we measure the social influence in item-level, compared with the structure-level analysis (Newman 2003)(Strogatz 2003) and topic-level analysis works (Tang et al. 2009). Second, the goal of the problem is to predict the users' social influence for unobserved data, which is in contrast with the majority of existing works to analyze the influence patterns from observed data (Anagnostopoulos, Kumar, and Mahdian 2008)(Bakshy, Karrer, and Adamic 2009).

Probabilistic Matrix Factorization

We suppose that there exists a joint latent space for both users and posts with common dimensionality k, such that the user-post specific social influences are modeled as the inner product between user-post vector pairs in that space. Accordingly, the user u_i is associated with an user vector $\mathbf{U}_i \in \mathbb{R}^k$, and the post p_j is associated with a post vector $\mathbf{V}_j \in \mathbb{R}^k$. We also define two matrices $\mathbf{U} = [\mathbf{U}_1, \mathbf{U}_2, \cdots, \mathbf{U}_M] \in \mathbb{R}^{k \times M}$, $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_N] \in \mathbb{R}^{k \times N}$. As the model performance is evaluated by root mean square error (RMSE) on the test set, we adopt a probabilistic linear model with Gaussian observation noise as in (Salakhutdinov and Mnih 2008). Here we define the the conditional distribution over the observed entries in $\mathbf{X}$ as

$$P(\mathbf{X}|\mathbf{U}, \mathbf{V}, \sigma_X^2) = \prod_{i=1}^{M} \prod_{j=1}^{N} \left[\mathcal{N}(X_{ij}|\mathbf{U}_i^\top \mathbf{V}_j, \sigma_X^2)\right]^{Y_{ij}} \quad (4)$$

where Y_{ij} is the indicator of user u_i sharing web post p_j.

We assume zero-mean spherical Gaussian priors on user and post feature vectors:

$$P(\mathbf{U}|\sigma_U^2) = \prod_{i=1}^{M} \mathcal{N}(U_i|0, \sigma_U^2 \mathbf{I}) P(\mathbf{V}|\sigma_V^2) = \prod_{j=1}^{N} \mathcal{N}(V_i|0, \sigma_V^2 \mathbf{I})$$

Then the log posterior distribution over the user and post feature vectors is calculated by

$$\ln P(\mathbf{U},\mathbf{V}|\mathbf{X},\sigma_X^2,\sigma_U^2,\sigma_V^2) \quad (5)$$
$$\propto -\frac{1}{2\sigma_X^2}\sum_{i,j} Y_{ij}(X_{ij}-\mathbf{U}_i^\top\mathbf{V}_j)^2-\frac{1}{2\sigma_U^2}\sum_i \mathbf{U}_i^\top\mathbf{U}_i-\frac{1}{2\sigma_V^2}\sum_i \mathbf{V}_i^\top\mathbf{V}_i$$

Therefore to obtain the *Maximum A Posteriori* (MAP) estimation of $\mathbf{U}$ and $\mathbf{V}$ is equivalent to minimize

$$\frac{1}{2\sigma_X^2}\left\|\mathbf{Y}\odot(\mathbf{X}-\mathbf{U}\mathbf{V}^\top)\right\|_F^2+\frac{1}{2\sigma_U^2}\left\|\mathbf{U}\right\|_F^2+\frac{1}{2\sigma_V^2}\left\|\mathbf{V}\right\|_F^2 \quad (6)$$

where $||.||_F^2$ is the Frobenius Norm, and the $\odot$ is the Hardamard Product.

Hybrid Factors

As mentioned above, the severe sparsity of $\mathbf{X}$ makes it very challenging to directly learn the latent spaces for users and posts from only observed user-post interaction entries. That's the reason why we need to make full use of the user-specific and post-specific factors to compress the degrees of freedom, so that the correlation within users and web posts can be exploited to alleviate the sparsity problem.

Incorporating User-Specific Factors

We selected two effective user-oriented predictive factors: the *percentage of active friends*, and the *average friend tie strength*, which are defined as follows.

- *The percentage of active friends*. The activeness of a friend u_r, denoted by $act(u_r)$, is the number of posts she visited during a given time period. Then the percentage of active friends is calculated by

$$uf_1(u_i)=\frac{1}{|\mathcal{N}(u_i)|}\sum_{u_r\in\mathcal{N}(u_i)}\delta(act(u_r)\geq\tau) \quad (7)$$

  where τ is the time threshold for active user, and $\delta(.)$ is the Delta function.

- *The Average Friend Tie Strength*. We define the *tie strength* between a user u_i and one of her friends u_r as the number of shared posts (by user u_i) friend u_r visits, which is denoted as $tie(u_i,u_r)$. Then, the average friend tie strength is calculated by

$$uf_2(u_i)=\frac{1}{|\mathcal{N}(u_i)|}\sum_{u_r\in\mathcal{N}(u_i)}\frac{tie(u_i,u_r)}{\sum_j Y_{ij}}. \quad (8)$$

We make use of these two factors to measure the similarity between u_i and u_j as

$$W_{ij}=\rho_1|uf_1(u_i)-uf_1(u_j)|+\rho_2|uf_2(u_i)-uf_2(u_j)| \quad (9)$$

In our experiments, we set $\rho_1=\rho_2=0.5$.

In this way, we can construct the user-user similarity matrix $\mathbf{W}\in\mathbb{R}^{M\times M}$. We further assume that $\mathbf{W}$ can be approximated by the inner product of the latent user matrix. Thus, the conditional distribution over the user-user similarity matrix is defined in a similar way as (4):

$$P(\mathbf{W}|\mathbf{U},\sigma_U^2)=\prod_{p=1}^M\prod_{q=1}^M\mathcal{N}(W_{pq}|\mathbf{U}_p^\top\mathbf{U}_q,\sigma_W^2) \quad (10)$$

Incorporating Post-Specific Factors

From our investigations on `www.renren.com`, the social influence is strongly correlated with the content of the web posts. For example, the posts on popular topics attract more clicks in average. We denote the post content matrix as $\mathbf{C}\in\mathbb{R}^{N\times d}$, where d is the dimensionality of the posts, which is constructed by implementing *Latent Dirichlet Allocation* (LDA) (Blei, Ng, and Jordan 2003) on the post corpus to discover 100 topics. Then the content of each post is represented as the topic distributions over the 100 topics. Similar to *latent semantic analysis* (Hofmann 1999), we assume there exists matrix $\mathbf{G}\in\mathbb{R}^{d\times k}$ which indicates topic group identity, and the conditional distribution over $\mathbf{C}$ is defined as

$$P(\mathbf{C}|\mathbf{V},\mathbf{G},\sigma_C^2)=\prod_{m=1}^N\prod_{n=1}^d\mathcal{N}(C_{mn}|\mathbf{V}_m^\top(\mathbf{G}^\top)_n,\sigma_C^2) \quad (11)$$

where $(\mathbf{G}^\top)_n$ is the n-th column of $\mathbf{G}^\top$.

Probabilistic Hybrid Factor MF

In this section, we will introduce a method, *Probabilistic Hybrid Factor Matrix Factorization*, to integrate all the above factors and get an estimation of the latent user and post matrices.

The Model

Given the observed matrices $\mathbf{X},\mathbf{W},\mathbf{C}$, the posterior distribution over the user and web post latent features is given by

$$P(\mathbf{U},\mathbf{V}|\mathbf{X},\mathbf{W},\mathbf{C},\Omega)=\frac{P(\mathbf{X},\mathbf{W},\mathbf{C}|\mathbf{U},\mathbf{V},\Omega)P(\mathbf{U},\mathbf{V}|\Omega)}{P(\mathbf{X},\mathbf{W},\mathbf{C})} \quad (12)$$
$$\propto P(\mathbf{X}|\mathbf{U},\mathbf{V},\Omega)P(\mathbf{W}|\mathbf{U},\Omega)P(\mathbf{C}|\mathbf{V},\mathbf{G},\Omega)P(\mathbf{U}|0,\Omega)P(\mathbf{V}|0,\Omega)$$

where $\Omega=\{\sigma_X^2,\sigma_W^2,\sigma_C^2,\sigma_U^2,\sigma_V^2\}$ is the hyperparameter set including the observed noise variance and prior variance on user and feature vectors.

The log of the posterior distribution over the user and post latent features is calculated by

$$\begin{aligned}&\ln(P(\mathbf{U},\mathbf{V}|\mathbf{X},\mathbf{W},\mathbf{C},\Omega))\\ &\propto -\frac{1}{2\sigma_X^2}\sum_{i,j}Y_{ij}(X_{ij}-\mathbf{U}_i^\top\mathbf{V}_j)^2-\frac{1}{2\sigma_W^2}\sum_{p,q}(W_{pq}-\mathbf{U}_p^\top\mathbf{U}_q)^2\\ &\quad-\frac{1}{2\sigma_C^2}\sum_{m,n}(\mathbf{C}_{mn}-\mathbf{V}_m^\top(\mathbf{G}^\top)_n)^2-\frac{1}{2\sigma_U^2}\sum_i\mathbf{U}_i^\top\mathbf{U}_i\\ &\quad-\frac{1}{2\sigma_V^2}\sum_i\mathbf{V}_i^\top\mathbf{V}_i\end{aligned}$$

Maximize the posterior distribution is equivalent to minimize the sum-of-squared errors function with hybrid quadratic regularization terms:

$$\begin{aligned}\mathcal{J}=&||\mathbf{Y}\odot(\mathbf{X}-\mathbf{U}^\top\mathbf{V})||_F^2+\alpha||\mathbf{W}-\mathbf{U}^\top\mathbf{U}||_F^2\\ &+\beta||\mathbf{C}-\mathbf{V}\mathbf{G}^\top||_F^2+\gamma||\mathbf{U}||_F^2+\delta||\mathbf{V}||_F^2.\end{aligned} \quad (13)$$

where $\alpha=\frac{\sigma_X^2}{\sigma_W^2}$, $\beta=\frac{\sigma_X^2}{\sigma_C^2}$, $\gamma=\frac{\sigma_X^2}{\sigma_U^2}$, and $\delta=\frac{\sigma_X^2}{\sigma_V^2}$.

user information	post information
user id (UID)	post id
friend links	post content
shared post id list	visiting friend UID list

Table 1: Data set information.

Solution

Although the objective function is not jointly convex with respect to $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{G}$, it is convex with each of them with the other two fixed. Therefore we can adopt a *block coordinate descent* scheme to solve the problem (Bertsekas 1999). That is, starting from some random initialization on $\mathbf{U}, \mathbf{V}, \mathbf{G}$, we solve each of them alternatively with the other two fixed, and proceed step by step until convergence[1]. In this paper, we use the gradient search method to solve the problem (Bertsekas 1999). Specifically, the gradients of the objective with respect to the variables are

$$\begin{aligned}
\frac{\partial \mathcal{J}}{\partial \mathbf{U}} &= 2\Big(-(\mathbf{Y}\odot\mathbf{X})\mathbf{V} + (\mathbf{Y}\odot\mathbf{U}\mathbf{V}^\top)\mathbf{V} \\
&\quad -2\alpha\mathbf{W}\mathbf{U} + 2\alpha\mathbf{U}\mathbf{U}^\top\mathbf{U} + \gamma\mathbf{U}\Big) \\
\frac{\partial \mathcal{J}}{\partial \mathbf{V}} &= 2\Big(-(\mathbf{Y}^\top\odot\mathbf{X}^\top)\mathbf{U} + (\mathbf{Y}^\top\odot\mathbf{V}\mathbf{U}^\top)\mathbf{U} \\
&\quad -\beta\mathbf{C}\mathbf{G} + \beta\mathbf{V}\mathbf{G}^\top\mathbf{G} + \delta\mathbf{V}\Big) \\
\frac{\partial \mathcal{J}}{\partial \mathbf{G}} &= 2\beta\Big(-\mathbf{C}^\top\mathbf{V} + \mathbf{G}\mathbf{V}^\top\mathbf{V}\Big)
\end{aligned}$$

Experiments

Dataset Information

We perform our experiments on a real online social network dataset, which is crawled from `http://renren.com`, a Facebook style social network web site in China. We have 34k users, and 43k web posts in the dataset, and the basic information we used for each user and post are listed in Table .

In our experiments, we randomly sample different number of users and select the web posts shared by these sampled users to form datasets with different sizes, including 500users dataset, 2000users dataset, 5000users dataset, and the 10000users dataset. They are used to evaluate the detail performance of the proposed method.

Comparative Methods

Besides the proposed PHF-MF method, we also implement the following methods for comparison.

- **Logistic Regression** (LR): If we regard the user and post factors as variables, and the strength of social influence as the response, then the prediction of social influence can be formulated as a regression problem. Thus, we firstly use the LR model to linearly combine the user factors and post factors, and learn the regression coefficients of the factors from the observed training data.
- **Cox Proportional Hazards Regression** (CoxPH): Different from LR, the user factors and post factors are combined in an exponential way, as is used in (Yang and Counts 2010), which aims to predict the speed of diffusion of tweets in Twitter.
- **User Averaging Influence** (AvgU): As users have different overall influences regardless of web posts, we can predict unobserved social influence by the average over observed ones, i.e.,

$$f_{i,\cdot} = \frac{\sum_j f_{ij}}{\sum_j Y_{ij}}. \tag{14}$$

- **Post Averaging Influence** (AvgP): As in AvgU, we can also predict the social influence by the web posts' averaging influence regardless of users:

$$f_{\cdot,j} = \frac{\sum_i f_{ij}}{\sum_i Y_{ij}}. \tag{15}$$

- **Basic Matrix Factorization** (bMF): In this method, we only consider the user-post interaction matrix, and find the joint latent space for users and posts by solving the objective function:

$$\min_{\mathbf{U},\mathbf{V}} \left\|\mathbf{X} - \mathbf{U}\mathbf{V}^\top\right\|_F^2 + \gamma\left\|\mathbf{U}\right\|_F^2 + \delta\left\|\mathbf{V}\right\|_F^2 \tag{16}$$

- **User Factors Constrained MF** (bMF+UF): By incorporating the user factors in to the bMF, we find the joint latent space for users and posts by solving:

$$\min_{\mathbf{U},\mathbf{V}} \left\|\mathbf{X} - \mathbf{U}\mathbf{V}^\top\right\|_F^2 + \alpha\left\|\mathbf{W} - \mathbf{U}\mathbf{U}^\top\right\|_F^2 + \gamma\left\|\mathbf{U}\right\|_F^2 + \delta\left\|\mathbf{V}\right\|_F^2$$

- **Post Factors Constrained MF** (bMF+PF): By incorporating the post factors in to the bMF, we find the joint latent space for users and posts by solving:

$$\min_{\mathbf{U},\mathbf{V}} \left\|\mathbf{X} - \mathbf{U}\mathbf{V}^\top\right\|_F^2 + \beta\left\|\mathbf{C} - \mathbf{V}\mathbf{G}^\top\right\|_F^2 + \gamma\left\|\mathbf{U}\right\|_F^2 + \delta\left\|\mathbf{V}\right\|_F^2$$

The quality of the prediction will be evaluated using the *Root Mean Square Error* (RMSE).

$$RMSE = \sqrt{\frac{\sum_{X_{ij}\in\mathcal{X}} (X_{ij} - \mathbf{U}_i\mathbf{V}_j^\top)^2}{|\mathcal{X}|}} \tag{17}$$

Parameter Settings

In this section, we will investigate the effect of different parameter settings when implementing PHF-MF, include tradeoff parameters, dimension of hidden space, and number of projected gradient iterations, on the performance.

[1]Here the objective is obviously lower bounded by 0, and the alternating gradient search procedure will decrease it monotonically. Thus the algorithm is guaranteed to be convergent.

α	β	γ	δ	$RMSE$
0.00001	0.001	0.001	0.001	0.15564
0.0002	**0.005**	**0.01**	**0.01**	**0.15135**
0.0001	0.01	0.01	0.01	0.15234
0.001	0.1	0.1	0.1	0.17742

Table 2: PHF-MF tradeoff parameters setting and evaluation.

Tradeoff Parameters The tradeoff parameters $\alpha, \beta, \gamma, \delta$ in PHF-MF play the role of adjusting the strength of different terms in the objective function. As the value range of the latter 4 components in equation (13) are different, the parameter setting should be consistent with the corresponding component's value range. Considering the roles of different components, we test the three sets of tradeoff parameters as shown in Table 2, and use the 2000users dataset for validation.

Although the parameters setting is similar with HF-NMF in (Cui et al. 2011), the physical significance of the parameters (i.e. the variance ratio of matrices) indicates a way for better understanding the parameter tuning process. The results in Table 2 show that the parameter set $\alpha = 0.0002, \beta = 0.005, \gamma = \delta = 0.01$ produce the best performance. In our following experiments, we just use this parameter setting.

Dimensionality of the Hidden Space The goal of PHF-MF is to find a k-dimensional joint latent space for users and web posts. How to set k is important for prediction performance. If k is too small, the users and web posts cannot be well represented and discriminated in the latent space. If k is too large, the computational complexity will be greatly increased. Thus, we conduct 5 experiments with k ranging from 5 to 40 on the 2000users dataset. The results are shown in Figure. 1, from which we can see that with the increase on the dimension k, $RMSE$ will reduce gradually. When $k > 30$, the RMSE reduces rather slow. For the concern of the tradeoff between efficiency and prediction precision, we choose $k = 20$ as the latent space dimension in our experiments.

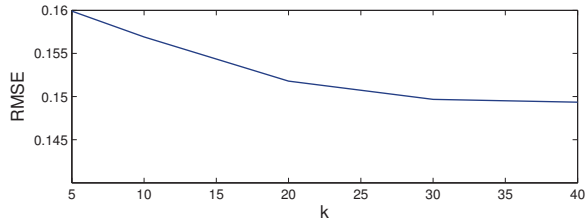

Figure 1: Prediction performance v.s. hidden space dimension.

Prediction Performance

In this section, we will demonstrate the prediction performance of the proposed method, and compare it with other methods.

We randomly select 50%, 70% and 90% of the observed entries in matrix $\mathbf{X}$ of difference sizes of datasets (including 500users dataset, 2000users dataset, 5000users dataset and 10000users dataset) as the training data, and the rest as the testing data. The random selection was carried out 10 times, and the average RMSE is reported. The same experiments are conducted on the proposed method and the 7 comparative methods listed in subsection . The results are shown in the Table 3.

From the Table 3, we can observe that:

- The probabilistic interpretation of the HF-NMF model gives a new way for indicating the parameter tuning process, and gains more insights on the physical significance of the parameters.
- The proposed PHF-MF algorithm, which incorporates the user, post and the user-post interaction factors together in a probabilistic way, achieves the best performance compared with other comparative methods, including the HF-NMF in (Cui et al. 2011).
- The more entries used for training, the lower RMSE the methods can achieve. This is consistent with the intuitive assumption that the prediction performance depend heavily on the percentage of training data, especially in sparse dataset where the model can be hardly sufficiently trained.

Conclusions

In this paper, we propose a *Probabilistic Hybrid Factor Matrix Factorization* (PHF-MF) method for item-level social influence prediction. By extending the baseline probabilistic matrix factorization method, and interpret the HF-NMF method in a probabilistic way, PHF-MF is endowed with a good ability to incorporate prior knowledge on tow dimensions of the interaction matrix, and a more solid theory foundation. Experimental results on a real social network dataset demonstrate that the proposed method can achieve better performance in social influence prediction compared with baseline methods.

Acknowledgements

This work is supported by National Natural Science Foundation of China, No. 60933013 and No. 61003097; National Basic Research Program of China, No. 2011CB302206; China Postdoctoral Science Foundation, No. 20100470285, and National Key Project Series, No. 2011ZX01042-001-002.

References

Anagnostopoulos, A.; Kumar, R.; and Mahdian, M. 2008. Influence and correlation in social networks. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*.

Bakshy, E.; Karrer, B.; and Adamic, L. A. 2009. Social influence and the diffusion of user-created content. In *Proceedings of the 10th ACM conference on Electronic commerce*.

	LR	CoxPH	AvgP	AvgU	bMF	bMF+PF	bMF+UF	PHF-MF
500 users								
50%	0.2026	0.202	0.199376	0.159798	0.158551	0.154719	0.153801	**0.152467**
70%	0.2093	0.2089	0.185823	0.150239	0.147858	0.149715	0.149009	**0.145877**
90%	0.1827	0.1822	0.178479	0.145262	0.1416632	0.1420855	0.1407717	**0.1398865**
2000 users								
50%	0.266	0.2642	0.226263	0.173654	0.171527	0.17034	0.169171	**0.168576**
70%	0.2304	0.2292	0.204045	0.163817	0.16154	0.161642	0.1597654	**0.158863**
90%	0.1699	0.1742	0.192715	0.1581	0.154478	**0.15044**	0.15273	0.15056
5000 users								
50%	0.2249	0.2241	0.250079	0.185797	0.183837	0.182474	0.179382	**0.176895**
70%	0.2288	0.2206	0.226694	0.1743	0.169922	0.169994	0.170496	**0.169372**
90%	0.2307	0.2324	0.210018	0.170393	0.167686	0.164501	0.164983	**0.163873**
10000 users								
50%	0.2615	0.2591	0.254009	0.189941	0.188926	0.185849	0.182247	**0.180954**
70%	0.2104	0.2073	0.23416	0.175754	0.174362	0.179101	0.172264	**0.170521**
90%	0.2354	0.234	0.210646	0.17097	0.167127	0.167633	0.165806	**0.164857**

Table 3: Prediction performance comparisons.

Bao, H., and Chang, E. Y. 2010. Adheat: an influence-based diffusion model for propagating hints to match ads. In *Proceedings of the 19th international conference on World wide web*.

Bertsekas, D. P. 1999. *Nonlinear Programming*. MIT: MIT Press.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993C1022.

Cui, P.; Wang, F.; Yang, S. Q.; and Sun, L. F. 2011. Who should share what? item-level social influence prediction for users and posts ranking. In *Proceeding of international ACM SIGIR conference on Research and development in information retrieval*.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd International conference on Research and development in information retrieval*.

Marlin, B., and Zemel, R. S. 2004. The multiple multiplicative factor model for collaborative filtering. In *Proceedings of the international conference on Machine learning*.

Marlin, B. 2004. Modeling user rating profiles for collaborative filtering. In *Proceeding of the Advances in neural information processing*.

Newman, M. E. J. 2003. The structure and function of complex networks. *SIAM Reviews* 45:167–256.

Salakhutdinov, R., and Mnih, A. 2008. Probabilistic matrix factorization. In *Proceeding of the Advances in neural information processing*.

Strogatz, S. H. 2003. Exploring complex networks. *Nature* 410:268C276.

Tang, J.; Sun, J.; Wang, C.; and Yang, Z. 2009. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*.

Wasko, M. M., and Faraj, S. 2005. Why should i share? examining social capital and knowledge contribution in electronic networks of practice. *Management Information System Quarterly* 29:35–57.

Yang, J., and Counts, S. 2010. Predicting the speed, scale, and range of information diffusion in twitter. In *Proceeding of the international AAAI conference on Weblogs and social media*.