



Retweet Prediction Using Social-Aware Probabilistic Matrix Factorization

Bo Jiang, Zhigang Lu^(✉), Ning Li, Jianjun Wu, and Zhengwei Jiang

Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
{jiangbo,luzhigang,lining6,wujianjun,jiangzhengwei}@iie.ac.cn

Abstract. Retweet prediction is a fundamental and crucial task in social networking websites as it may influence the process of information diffusion. Existing prediction approaches simply ignore social contextual information or don't take full advantage of these potential factors, damaging the performance of prediction. Besides, the sparsity of retweet data also severely disturb the performance of these models. In this paper, we propose a novel retweet prediction model based on probabilistic matrix factorization method by integrating the observed retweet data, social influence and message semantic to improve the accuracy of prediction. Finally, we incorporate these social contextual regularization terms into the objective function. Comprehensive experiments on the real-world dataset clearly validate both the effectiveness and efficiency of our model compared with several state-of-the-art baselines.

Keywords: Social network · Retweet prediction
Matrix factorization · Social influence · Message semantic

1 Introduction

Online social networks such as Twitter and Facebook have become tremendously popular in recent years. These services are a network structure system formed by interaction among users. The dissemination of information in social networks has brought unprecedented improvement under the structure and has accelerated interpersonal communication and information flow. The retweet mechanism provides a way to allow social users to hold the latest news and help enterprises to carry out marketing on social networking platform. Thus, it is of great practical significance to analyze and explore the retweet behaviors for improving the information propagation and user experience in social networks.

Many approaches has been proposed to model the retweet behaviors based on different social information, such as textual feature [13], social feature [7], social influence [18], visual feature [2], emotion feature [3], or a combination of these various features [13]. Although these methods have made some progress to some extent, the results are unsatisfactory, and can still be improved in a certain space. To improve the performance of prediction, recent works incorporate the observed explicit social information (e.g., social ties) into matrix factorization frameworks

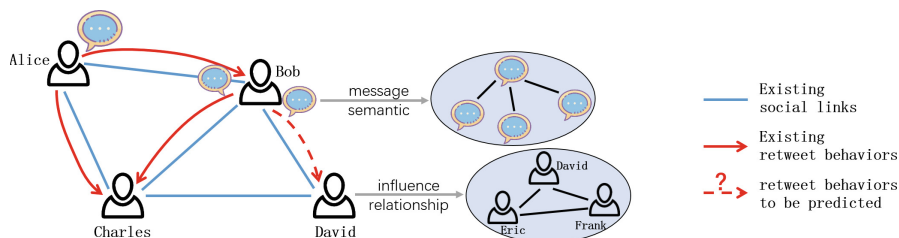


Fig. 1. Illustration of retweet behaviors on online social network.

to design novel models [15, 16]. In fact, it is naturally that the retweet prediction can be viewed as the problem of matrix completion by incorporating additional sources of information about social influence between users and message semantic between short texts. As shown in the example of Fig. 1, when users decide to retweet message, they are interested in the content of message and more likely to retweet messages posted by his close friends due to social relationships. We call this phenomenon for social context. These knowledge can be learnt from social influence and message semantic information. Both of these aspects are important for retweet prediction. However, most existing methods simply ignore contextual information, or don't take full advantage of these potential features.

In this paper, we propose a novel retweet prediction model based on probabilistic matrix factorization by integrating the observed retweet data, social influence and message semantic to improve the accuracy of prediction. Specifically, we first introduce social influence matrix based on network structure and interaction history and message similarity matrix based on document semantic. We then utilize user and message latent feature spaces to learn social influence and message semantic respectively. We incorporate these regularization terms into the objective function. Finally, we conduct several experiments to validate the effectiveness of our model with the state-of-the-art approaches. Experimental results show our model performs better than the baseline models.

The main contributions of this paper are the followings:

- We propose a novel retweet prediction model based on probabilistic matrix factorization by incorporating social influence and message semantic information to improve the performance of prediction.
- We utilize low-rank user latent feature space and message latent feature space to learn social influence and message semantic. The predicted social influence and message semantic can assist the applications such as influencer ranking and information recommendation.
- Various experiments are conducted on real-world social network dataset, and the results demonstrate that our proposed model can achieve better prediction performance than the state-of-the-art methods.

Our paper is organized as follows. Section 2 reviews the related work. Section 3 presents the preliminaries. Our model is proposed in Sect. 4. The experiments are presented in Sect. 5, followed by the conclusion in Sect. 6.

2 Related Work

2.1 Social Recommendation Modeling

Matrix factorization (MF) as well as the variants are widely used in ratings prediction in recommendation system [12]. To enhance the prediction performance of recommender systems with explicit social information, considerable social recommendation models are proposed based on matrix factorization [4, 8, 17, 20]. For example, Ma et al. [9] extend the probabilistic matrix factorization model by additionally incorporating user's social network information to eliminate the data sparsity and improve poor prediction accuracy problems. Hereafter, Ma et al. [8] also propose a social trust ensemble analysis framework by combining users' personal preference and their trusted friends' favors together. Guo et al. [4] design a trust-based MF technique by considering both the explicit and implicit influences of the neighborhood structure of trust information when predicting unknown ratings. Yang et al. [17] design a TrustMF model by integrating sparse rating data given by users and sparse social trust network among these same users. Zhao et al. [20] extend BPR by introducing social positive feedbacks and proposed the SBPR algorithm which achieves better performance in items ranking than BPR. Tang et al. [14] give a recommendation framework SoDim-Rec which incorporates heterogeneity of social relations and weak dependency connections based on social dimensions.

In a word, social context-aware model can take various types of contextual information (e.g., time, location) into account when making recommendations. It also provides a way for the prediction tasks based on context dependent.

2.2 Retweet Behavior Modeling

Many studies have been conducted to identify the influence factors of retweet behavior from different perspectives, including user survey [1, 10], data statistics [13]. In summary, these studies have identified that user's topic interests and social influence are two important aspects for retweet prediction. Meanwhile, research on user's retweet behavior prediction is more attractive to lots of researchers. Representative works include topic-level probabilistic graph model [6], conditional random field [11], social influence factor graph model [18], non-parametric Bayesian model [19], matrix factorization [15, 16]. The above approaches mainly use content and/or structure features to predict retweet behavior. Besides, some works associate with multiple features to predict retweet behavior [7]. However, these studies focus on exploring user-based and message-based features to predict retweet behavior based on the assumption that users and messages are independent and identically distributed. They ignore implicit side information such as social influence among users and semantic structure information among messages.

In summary, research on retweet prediction is still room for improvement. Inspired by social contextual information, we introduce social influence among users and message semantic among messages to devise our retweet prediction method.

3 Preliminaries

Given a message m and a user u , the task of the retweet prediction is to discover whether u retweet m or not. In this work, we use an $M \times N$ user-message retweet matrix $R = \{0, 1\} \in \mathbb{R}^{M \times N}$ to represent the behaviors of users retweet messages, in which user u_i retweet message m_j , R_{ij} is 1, otherwise R_{ij} is 0. Notice that 0s might either be “true” 0s or missing values.

We utilize Probabilistic Matrix Factorization (PMF) [12] to factorize R into user latent feature matrix $U \in \mathbb{R}^{K \times M}$ and message latent feature matrix $V \in \mathbb{R}^{K \times N}$. K is the number of latent features. Also, the retweet matrix R can be approximated by $R' \approx U^T V$. The likelihood function of the observed retweetings is factorised across M users and N messages with each factor as

$$P(R|U, V, \sigma_R^2) = \prod_{i=1}^M \prod_{j=1}^N [\mathcal{N}(R_{ij}|U_i^T V_j, \sigma_R^2)]^{I_{ij}^{(R)}} \quad (1)$$

where $\mathcal{N}(\cdot|\mu, \sigma^2)$ is the probability density function of the normal distribution with mean μ and variance σ^2 . The indicator function $I_{ij}^{(R)}$ is equal to 1 when user u_i retweet message v_j and 0 otherwise. The prior distributions over U and V are defined as

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

We then have the posterior probability by the Bayesian inference as

$$\begin{aligned} P(U, V|R, \sigma_R^2, \sigma_U^2, \sigma_V^2) &\propto P(R|U, V, \sigma_R^2) P(U|\sigma_U^2) P(V|\sigma_V^2) \\ &= \prod_{i=1}^M \prod_{j=1}^N [\mathcal{N}(R_{ij}|g(U_i^T V_j), \sigma_R^2)]^{I_{ij}^{(R)}} \times \prod_{i=1}^M \mathcal{N}(U_i|0, \sigma_U^2 \mathbf{I}) \times \prod_{j=1}^N \mathcal{N}(V_j|0, \sigma_V^2 \mathbf{I}) \end{aligned} \quad (3)$$

This model is learned by maximizing posterior probability, which is equivalent to minimizing sum-of-squares of factorization error with regularization terms

$$\mathcal{L} = \frac{1}{2} \sum_{i=1}^M \sum_{j=1}^N I_{ij}^{(R)} (R_{ij} - g(U_i^T V_j))^2 + \frac{\eta}{2} \|U\|_F^2 + \frac{\lambda}{2} \|V\|_F^2 \quad (4)$$

where the logistic function $g(x) = 1/(1 + \exp(-x))$ maps the value of $U_i^T V_j$ to the range $(0, 1)$, $\eta = \frac{\sigma_R^2}{\sigma_U^2}$, $\lambda = \frac{\sigma_R^2}{\sigma_V^2}$ and $\|\cdot\|_F$ denotes the Frobenius norm.

As we have described above, the retweet prediction can be considered as a matrix completion task, where the unobserved retweetings in matrix R can be predicted based on the observed retweet behaviors. However, R is highly sparse, it is extremely difficult to directly learn the optimal latent spaces for users and messages only by the observed retweeting entries. We argue that social contextual information can assist in prediction. For example, people with social relations are more likely to share same preferences, and users pay close attention to their interested topics. By this idea, we incorporate these social contextual information into our prediction method.

4 Social-Aware Prediction Model

4.1 Modeling Social Influence

User's action can be affected with others in the process of information spread. For example, whether user like message or not will be affected by the publisher to some extent. Here, we argue that the user's retweet behaviors are affected by his direct neighbors due to social influence in social networks. Thus, we employ social influence to improve the prediction performance.

We denote the social influence matrix $F \in \mathbb{R}^{M \times M}$, in which each entry F_{ij} represent the strength of social influence user u_i has on user u_j based on network structure and interaction behaviors. Similarly, we factorize F into user latent feature matrix $U \in \mathbb{R}^{K \times M}$ and factor latent feature matrix $Z \in \mathbb{R}^{K \times M}$. We define the conditional distribution over the observed social influence as

$$P(F|U, Z, \sigma_F^2) = \prod_{i=1}^M \prod_{f=1}^M [\mathcal{N}(F_{if}|g(U_i^T Z_f), \sigma_F^2)] \quad (5)$$

We also place prior distributions on U and F as

$$P(U|\sigma_U^2) = \prod_{i=1}^M \mathcal{N}(U_i|0, \sigma_U^2 \mathbf{I}), \quad P(Z|\sigma_Z^2) = \prod_{f=1}^M \mathcal{N}(Z_f|0, \sigma_Z^2 \mathbf{I}) \quad (6)$$

We quantify the strength of influence based on network structure and interaction behaviors. Specifically, we first explore the utilization of network structure to quantify influence. For example, in social network, a user is a high influencer if he is followed by many users. Based on the idea, we denote the network structure influence matrix F_{ij}^S with its (i, j) -th entry as

$$F_{ij}^S = \frac{n_{u_i}^{in}}{n_{u_i}^{in} + n_{u_i}^{out}} \times I_{ij}^{(S)} \quad (7)$$

where $n_{u_i}^{in}$ is the follower number of u_i and $n_{u_i}^{out}$ is the following number of u_i . The indicator function $I_{ij}^{(S)}$ is equal to 1 if u_j is a follower of u_i and 0 otherwise.

We also measure interaction influence from the user interaction history in social networks. Similarly, we compute the (i, j) -th entry for the interaction behavior influence matrix F_{ij}^B as

$$F_{ij}^B = \frac{\sum_{y \in Y(i,j)} (A_{iy} - \bar{A}_i) \cdot (A_{jy} - \bar{A}_j)}{\sqrt{\sum_{y \in Y(i,j)} (A_{iy} - \bar{A}_i)^2} \cdot \sqrt{\sum_{y \in Y(i,j)} (A_{jy} - \bar{A}_j)^2}} \quad (8)$$

where $Y(i, j)$ represents the set of messages accepted by both users u_i and u_j , \bar{A}_i represents the average acceptance of user u_i . To guarantee non negativity, we use the sigmod function to map F_{ij} into $(0, 1)$.

Finally, social influence from user u_i to user u_j is calculated as

$$F_{ij} = g(\rho F_{ij}^S + (1 - \rho) F_{ij}^B) \quad (9)$$

where $\rho \in (0, 1)$ controls the effects of network topology structure and history of interaction.

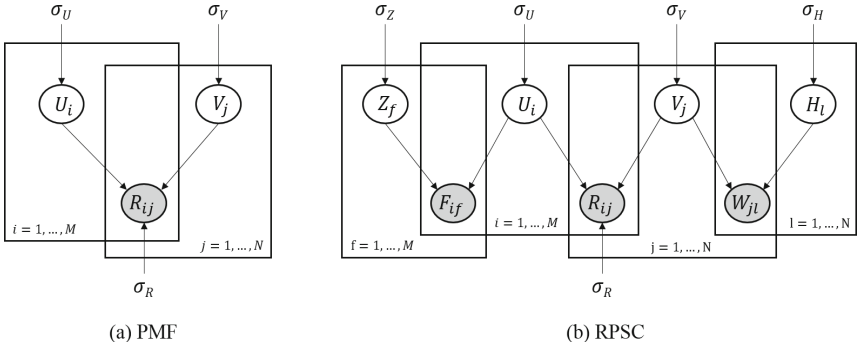


Fig. 2. The graphical representation of (a) PMF and (b) our proposed method.

4.2 Modeling Message Semantic

The findings have been indicated that the content of message is an important factor when users retweet message [1, 10]. We argue that the topic distribution of messages can reflect user’s personal topic interests. Therefore, we also explore the utilization of message semantic to improve the retweet prediction.

We introduce the content similarity matrix $W \in \mathbb{R}^{N \times N}$ to represent message similarity information. Each entry W_{ij} denotes the similarity score between messages m_i and m_j . Similarly, we factorize W into message latent feature matrix $V \in \mathbb{R}^{K \times N}$ and factor latent feature matrix $H \in \mathbb{R}^{K \times N}$. We then define the conditional distribution over the observed message semantic as

$$P(W|V, H, \sigma_W^2) = \prod_{j=1}^N \prod_{l=1}^N [\mathcal{N}(W_{jl}|g(V_j^T H_l), \sigma_W^2)] \tag{10}$$

We also place zero-mean Gaussian priors on W and H as

$$P(V|\sigma_V^2) = \prod_{j=1}^N \mathcal{N}(V_j|0, \sigma_V^2 \mathbf{I}), \quad P(H|\sigma_H^2) = \prod_{l=1}^N \mathcal{N}(H_l|0, \sigma_H^2 \mathbf{I}) \tag{11}$$

In this paper, we employ GPU-DMM [5] method to infer latent topic structure of short texts. After performing GPU-DMM, we can represent each document with its topic distribution $p(z|d)$. Hence, we can compute content similarity matrix W based on cosine similarity method between messages m_i and m_j as

$$W_{ij} = p(z|d_i)p(z|d_j) \tag{12}$$

where $p(z|d_i)$ denotes the topic distribution of message m_i .

4.3 Learning and Prediction

Next, we incorporate social influence and message semantic similarity information into the framework of probabilistic matrix factorization and solve the optimization. The corresponding graphical model is presented in Fig. 2.

Based on Bayesian inference, we model the conditional distribution of U , V , Z and H over social influence and message semantic similarity as

$$\begin{aligned}
& P(Z, H, U, V | R, F, W, \sigma_R^2, \sigma_Z^2, \sigma_H^2, \sigma_U^2, \sigma_V^2) \propto P(R|U, V, \sigma_R^2) \\
& P(F|U, Z, \sigma_F^2) P(W|V, H, \sigma_W^2) P(U|\sigma_U^2) P(V|\sigma_V^2) P(Z|\sigma_Z^2) P(H|\sigma_H^2) \\
& = \prod_{i=1}^M \prod_{j=1}^N [\mathcal{N}(R_{ij} | g(U_i^T V_j), \sigma_R^2)]^{I_{ij}^{(R)}} \times \prod_{i=1}^M \prod_{f=1}^M [\mathcal{N}(F_{if} | g(U_i^T Z_f), \sigma_F^2)] \\
& \times \prod_{j=1}^N \prod_{l=1}^N [\mathcal{N}(W_{jl} | g(V_j^T H_l), \sigma_W^2)] \times \prod_{i=1}^M \mathcal{N}(U_i | 0, \sigma_U^2 \mathbf{I}) \times \prod_{j=1}^N \mathcal{N}(V_j | 0, \sigma_V^2 \mathbf{I}) \\
& \times \prod_{f=1}^M \mathcal{N}(Z_f | 0, \sigma_Z^2 \mathbf{I}) \times \prod_{l=1}^N \mathcal{N}(H_l | 0, \sigma_H^2 \mathbf{I})
\end{aligned} \tag{13}$$

Maximizing log-posterior distribution on U , V , Z and H is equivalent to minimizing sum-of-of-squared errors function with quadratic regularization terms as

$$\begin{aligned}
\min_{U, V} \mathcal{L} &= \frac{1}{2} \sum_{i=1}^M \sum_{j=1}^N I_{ij}^{(R)} (R_{ij} - g(U_i^T V_j))^2 \\
&+ \frac{\alpha}{2} \sum_{i=1}^M \sum_{f=1}^M \|F_{if} - g(U_i^T Z_f)\|_F^2 + \frac{\beta}{2} \sum_{j=1}^N \sum_{l=1}^N \|W_{jl} - g(V_j^T H_l)\|_F^2 \\
&+ \frac{\gamma}{2} \|U\|_F^2 + \frac{\eta}{2} \|V\|_F^2 + \frac{\varphi}{2} \|Z\|_F^2 + \frac{\rho}{2} \|H\|_F^2
\end{aligned} \tag{14}$$

where $\alpha = \frac{\sigma_R^2}{\sigma_F^2}$, $\beta = \frac{\sigma_R^2}{\sigma_W^2}$, $\gamma = \frac{\sigma_R^2}{\sigma_U^2}$, $\eta = \frac{\sigma_R^2}{\sigma_V^2}$, $\varphi = \frac{\sigma_R^2}{\sigma_Z^2}$, $\rho = \frac{\sigma_R^2}{\sigma_H^2}$. In order to reduce the model complexity, we set $\gamma = \eta = \varphi = \rho$ in all of the experiments.

The local minimum of the objective function given by Eq.(14) can be found by using stochastic gradient descent on feature vectors U_i , V_j , Z_f and H_l as

$$\frac{\partial \mathcal{L}}{\partial Z_f} = \sum_{i=1}^M g'(U_i^T Z_f) (g(U_i^T Z_f) - F_{if}) U_i + \varphi Z_f \tag{15}$$

$$\frac{\partial \mathcal{L}}{\partial H_l} = \sum_{j=1}^N g'(V_j^T H_l) (g(V_j^T H_l) - W_{jl}) V_j + \rho H_l \tag{16}$$

$$\frac{\partial \mathcal{L}}{\partial U_i} = \sum_{j=1}^N I_{ij}^{(R)} g'(U_i^T V_j) (g(U_i^T V_j) - R_{ij}) V_j + \alpha \sum_{f=1}^M g'(U_i^T Z_f) (g(U_i^T Z_f) - F_{if}) Z_f + \gamma U_i \tag{17}$$

$$\frac{\partial \mathcal{L}}{\partial V_j} = \sum_{i=1}^M I_{ij}^{(R)} g'(U_i^T V_j) (g(U_i^T V_j) - R_{ij}) U_i + \beta \sum_{l=1}^N g'(V_j^T H_l) (g(V_j^T H_l) - W_{jl}) H_l + \eta V_j \tag{18}$$

where $g'(x)$ is the derivative of logistic function $g'(x) = \exp(x)/(1 + \exp(x))^2$. After learning U and V , an unknown retweet entry can be estimated as $U_i^T V_j$.

5 Experimental Analysis

5.1 Dataset Description

We use a real-world dataset collected from Weibo which is a social network in China like Twitter. Weibo allows user to generate following and follower relationships, and retweet the interested message posted by other people. In this paper, we use publicly available Weibo dataset to evaluate the validity of our proposed method [18]. The dataset contains the content of message, the relationships of user’s following and follower, and the information of retweet behaviors. The data statistics are illustrated in Table 1. It can be seen from the statistical results that user behaviors data are very sparse on social networks.

Table 1. Retweet data statistics

Dataset	#Users	#Tweets	#Retweets	#Relations	Density
Weibo	1,787,443	300,000	23,755,810	308,489,739	0.005%

5.2 Comparative Algorithms

We compare our method (RPSC) with the following baseline algorithms.

- **PMF**: This method doesn’t take into account social contextual information and only uses user-message matrix for the retweet prediction [12].
- **LRC-BQ**: The method designs social influence locality based on pairwise influence and structural diversity, and adds the basic features and influence locality features into the logistic regression to predict retweet behavior [18].
- **MNMF_{RP}**: This method measures the social relationship strength based on network structure and interaction history, and introduces social relationship regularizer under non-negative matrix factorization to predict retweets [16].
- **HCFMF**: The model provides a co-factor matrix factorization by modeling message co-occurrence similarity based on microblog content, word semantic similarity based on word embeddings, and user social similarity based on author information into collaborative filtering when predicting retweets [15].

We also consider the variants of the proposed model to verify the effectiveness. Let $\mathcal{L}_o = \frac{1}{2} \sum_{i=1}^M \sum_{j=1}^N I_{ij}^{(R)} (R_{ij} - g(U_i^T V_j))^2 + \frac{\gamma}{2} \|U\|_F^2 + \frac{\eta}{2} \|V\|_F^2$, then we have

- **RPSC-U**: This method only considers user’s social influence information in our proposed model. The adjusted function is

$$\mathcal{L}(R, U, V, Z) = \mathcal{L}_o + \frac{\alpha}{2} \sum_{i=1}^M \sum_{f=1}^M \|F_{if} - g(U_i^T Z_f)\|_F^2 + \frac{\varphi}{2} \|Z\|_F^2 \quad (19)$$

- **RPSC-M**: This method only utilizes message semantic information for the proposed model. The degenerated function is

$$\mathcal{L}(R, U, V, H) = \mathcal{L}_o + \frac{\beta}{2} \sum_{j=1}^N \sum_{l=1}^N \|W_{jl} - g(V_j^T H_l)\|_F^2 + \frac{\rho}{2} \|H\|_F^2 \quad (20)$$

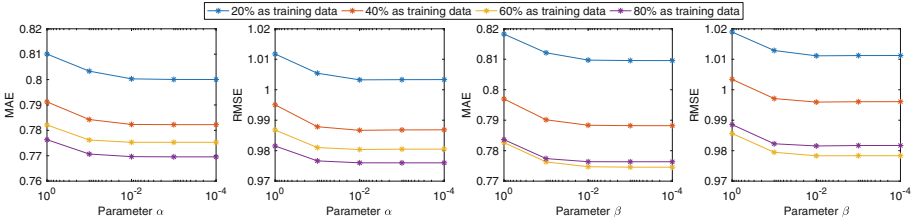


Fig. 3. MAE and RMSE vs. α , β on the different training data settings.

5.3 Evaluation Measures

For the evaluation metrics, we use the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) to measure the model accuracy, defined as

$$MAE = \frac{\sum_{R_{ij} \in \mathcal{R}} |R_{ij} - U_i^T V_j|}{|\mathcal{R}|}, \quad RMSE = \sqrt{\frac{\sum_{R_{ij} \in \mathcal{R}} (R_{ij} - U_i^T V_j)^2}{|\mathcal{R}|}} \quad (21)$$

where R_{ij} denotes the retweet value given message m_j by user u_i . $|\mathcal{R}|$ denotes the number of tested entries. A smaller MAE or RMSE means a better performance.

We also employ Precision, Recall, and F_1 -score to evaluate whether user retweets or not when receiving message. A simple strategy is that hide some observed entries as unobserved entries for evaluation, and perform classification after training. We perform 5-fold cross validation and report their average values.

5.4 Parameter Settings

Tradeoff Parameters. In our proposed model, the parameters α and β are used to control the strength of social influence and the weight of message semantic respectively, and the rest parameters γ , η , φ and ρ is used to prevent overfitting. In this paper, we use different amounts of training data (20%, 40%, 60%, 80%) to find the optimal values of parameters. From the results shown in Fig. 3, we can see that MAE and RMSE gradually decrease and achieve the best performance when α is around 10^{-2} . Hence, we set $\alpha = 10^{-2}$ for the following experiments. Meanwhile, the impact of β shares the same trends as α . We thus set $\beta = 10^{-2}$. We also conduct the extensive experiments and observe similar results with γ , η , φ and ρ . We directly give $\gamma = \eta = \varphi = \rho = 10^{-2}$ due to the space limitation.

Number of Latent Features. The dimensionality of latent factor indicates the power of feature representation, and the proper dimensionality can be more effective to predict the retweet behaviors. Thus, we train latent feature matrices U , V , Z and H to find the optimal latent space to represent users and messages. In this paper, we use different training datasets to discover the appropriate K . From the results shown in Fig. 4, we can observe that MAE and RMSE decreases gradually when the latent feature K increase. Finally, we choose $K = 100$ as the feature dimension in our experiments due to computational cost.

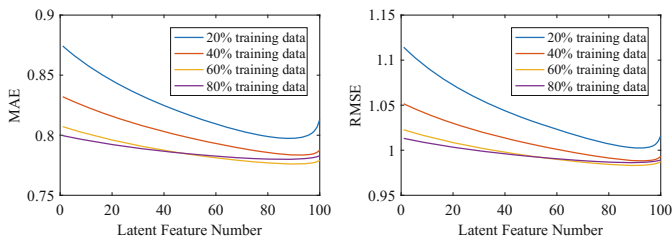


Fig. 4. MAE and RMSE vs. Latent Feature on the different training data settings.

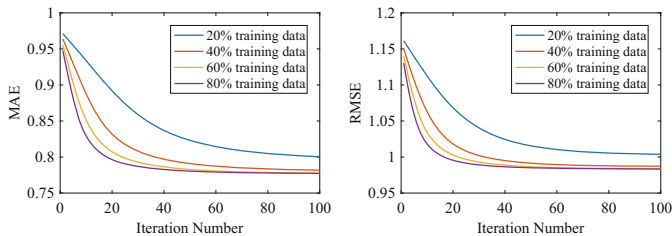


Fig. 5. MAE and RMSE vs. Iteration on the different training data settings.

Number of Iterations. Minimizing the objective function of our proposed method need to seek a proper number of iterations so that the algorithm has a better convergence performance while avoid overfitting. In this paper we try to the different number of iterations with various proportions of training data. The MAE and RMSE values are recorded in each iteration, shown in Fig. 5. From the results, we can conclude that MAE and RMSE values decrease gradually when increasing the number of iterations. Finally, we choose a limited number of iterations (i.e., 100) as the stop condition of our method.

5.5 Performance and Analysis

The underlying intuition that whether user retweet message or not is a binary value. Thus, in this section, we are consider the problem of retweet prediction as the task of classification. Specifically, we use the learned approximate entries as feature for Naïve Bayes classifier, and then evaluate the performance of our model and other baselines on the different training data settings. The experiment results on social network data are shown in Table 2. From the results, we can draw the following observations: (1) our proposed method outperforms all the other baseline methods on the different training datasets and improves the user’s retweet prediction to some extent; (2) HCFMF, MNMFRP and RPSC outperforms PMF, which demonstrates that utilizing social contextual information is more effective than simply ignore such social context; (3) the relative improvement of HCFMF, MNMFRP and RPSC over LRC-BQ shows that the matrix factorization method is more suited to the retweet prediction task; (4)

among the RPSC variants, RPSC-U performs better in improving prediction performance in terms of F_1 -score, indicating that social influence contributes more than message semantic. The possible explanation is that social influence comprehensively reflects user behavior patterns. In summary, these results suggest that our proposed model can achieve the better performance by casting the prediction problem into the solution of probabilistic matrix factorization with combining social contextual information.

Table 2. Performance of retweet prediction with different training data settings.

Method	60% as training data			100% as training data		
	Precision	Recall	F_1 -score	Precision	Recall	F_1 -score
PMF	0.484	0.434	0.458	0.584	0.534	0.558
LRC-BQ	0.518	0.677	0.587	0.698	0.770	0.733
MNMFPR	0.674	0.715	0.694	0.796	0.791	0.793
HCFMF	0.787	0.805	0.796	0.802	0.834	0.818
RPSC-U	0.791	0.811	0.801	0.804	0.847	0.825
RPSC-M	0.785	0.815	0.799	0.799	0.829	0.814
RPSC	0.801	0.827	0.814	0.806	0.853	0.829

6 Conclusion

In this paper, we propose a novel model to predict user’s retweet behaviors based on probabilistic matrix factorization method, in which incorporates social influence learned from network structure and user’s interaction behavior and message semantic obtained by modeling the topic distribution of the message. Then we combine user-user social influence and message-message semantic similarity regularization terms to constrain objective function under probabilistic matrix factorization. To validate the effectiveness and efficient of our model, we construct extensive experiments. The experimental results reveal that the proposed method can effectively improve the accuracy of retweet prediction compare with the state-of-the-art baseline methods. As future work, we plan to extend this work incorporating time delay factor between the posted message and the received user and explore how the deep learning model can be employed so that the feature vectors of users and messages can be further learned efficiently.

Acknowledgement. This work is supported by Natural Science Foundation of China (No. 61702508), and National Key Research and Development Program of China (No. 2016YFB0801004, 2016QY04w0905). This work is also partially supported by Key Laboratory of Network Assessment Technology, Chinese Academy of Sciences and Beijing Key Laboratory of Network security and Protection Technology.

References

1. Abdullah, N.A., Nishioka, D., Tanaka, Y., Murayama, Y.: User's action and decision making of retweet messages towards reducing misinformation spread during disaster. *J. Inf. Process.* **23**(1), 31–40 (2015)
2. Can, E.F., Oktay, H., Manmatha, R.: Predicting retweet count using visual cues. In: *CIKM*, pp. 1481–1484. ACM (2013)
3. Chen, J., Liu, Y., Zou, M.: User emotion for modeling retweeting behaviors. *Neural Netw.* **96**, 11–21 (2017)
4. Guo, G., Zhang, J., Yorke-Smith, N.: TrustSVD: collaborative filtering with both the explicit and implicit influence of user trust and of item ratings. In: *AAAI*, pp. 123–129 (2015)
5. Li, C., Wang, H., Zhang, Z., Sun, A., Ma, Z.: Topic modeling for short texts with auxiliary word embeddings. In: *SIGIR*, pp. 165–174. ACM (2016)
6. Liu, L., Tang, J., Han, J., Jiang, M., Yang, S.: Mining topic-level influence in heterogeneous networks. In: *CIKM*, pp. 199–208. ACM (2010)
7. Luo, Z., Osborne, M., Tang, J., Wang, T.: Who will retweet me?: finding retweeters in twitter. In: *SIGIR*, pp. 869–872. ACM (2013)
8. Ma, H., King, I., Lyu, M.R.: Learning to recommend with social trust ensemble. In: *SIGIR*, pp. 203–210 (2009)
9. Ma, H., Yang, H., Lyu, M.R., King, I.: SoRec: social recommendation using probabilistic matrix factorization. In: *CIKM*, pp. 931–940 (2008)
10. Metaxas, P.T., Mustafaraj, E., Wong, K., Zeng, L., O'Keefe, M., Finn, S.: What do retweets indicate? Results from user survey and meta-review of research. In: *ICWSM*, pp. 658–661 (2015)
11. Peng, H.-K., Zhu, J., Piao, D., Yan, R., Zhang, Y.: Retweet modeling using conditional random fields. In: *ICDM*, pp. 336–343. IEEE (2011)
12. Salakhutdinov, R., Mnih, A.: Probabilistic matrix factorization. In: *NIPS*, pp. 1257–1264 (2007)
13. Suh, B., Hong, L., Pirolli, P., Chi, E.H.: Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network. In: *SocialCom*, pp. 177–184. IEEE (2010)
14. Tang, J., Wang, S., Hu, X., Yin, D., Bi, Y., Chang, Y., Liu, H.: Recommendation with social dimensions. In: *AAAI*, pp. 251–257 (2016)
15. Wang, C., Li, Q., Wang, L., Zeng, D.D.: Incorporating message embedding into co-factor matrix factorization for retweeting prediction. In: *IJCNN*, pp. 1265–1272. IEEE (2017)
16. Wang, M., Zuo, W., Wang, Y.: A multidimensional nonnegative matrix factorization model for retweeting behavior prediction. *Math. Probl. Eng.* **2015**, 1–10 (2015)
17. Yang, B., Lei, Y., Liu, J., Li, W.: Social collaborative filtering by trust. In: *IJCAI*, pp. 2747–2753 (2013)
18. Zhang, J., Tang, J., Li, J., Liu, Y., Xing, C.: Who influenced you? Predicting retweet via social influence locality. *ACM TKDD* **9**(3), 25 (2015)
19. Zhang, Q., Gong, Y., Guo, Y., Huang, X.: Retweet behavior prediction using hierarchical dirichlet process. In: *AAAI*, pp. 403–409 (2015)
20. Zhao, T., McAuley, J., King, I.: Leveraging social connections to improve personalized ranking for collaborative filtering. In: *CIKM*, pp. 261–270. ACM (2014)