# Unifying User and Message Clustering Information for Retweeting Behavior Prediction

Bo Jiang, Jiguang Liang, Ying Sha(✉), Lihong Wang, Zhixin Kuang, Rui Li, and Peng Li

Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
{jiangbo,liangjiguang,shaying,kuangzhixin,lirui,lipeng}@iie.ac.cn

**Abstract.** Online social networks have been recently increasingly become the dominant platform of information diffusion by user's retweeting behavior. Thus, understanding and predicting who will be retweeted in a given network is a challenging but important task. Existing studies only investigate individual user and message for retweeting prediction. However, social influence and selection lead to formation of groups. The intrinsic and important factor has been neglected for this problem. In the paper, we propose a unified user and message clustering based approach for retweeting behavior prediction. We first cluster users and messages into different groups based on explicit and implicit factors together. Then we model social clustering information as regularization terms to introduce the retweeting prediction framework in order to reduce sparsity of data and improve accuracy of prediction. Finally, we employ matrix factorization method to predict user's retweeting behavior. The experimental results on a real-world dataset demonstrate that our proposed method effectively increases accuracy of retweeting behavior prediction compared to state-of-the-art methods.

**Keywords:** Retweeting behavior · Social networks · Matrix factorization · User clustering · Message clustering

## 1 Introduction

With the advent of social network platforms such as Twitter, Facebook and Weibo, thousands of millions of users have used these sites to share opinions and ideas with each other, and to engage in interesting activities about all kinds of topics and hot events. Social networks encourage connections, interactions and relationships between people. Thus, social network services allow a user to follow other users forming social link. On this basis, as message is forwarded from user to user, large cascades of reshares can be formed. As a result, information dissemination power has a unprecedented improvement via user's retweeting behavior. Retweeting has been considered as a key mechanism of information diffusion in Twitter [15]. Hence, understanding the retweeting effect factors from


B. Cui et al. (Eds.): WAIM 2016, Part II, LNCS 9659, pp. 291–303, 2016.
DOI: 10.1007/978-3-319-39958-4_23

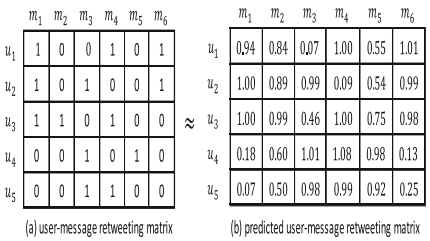



**Fig. 1.** Predicting unobserved retweetings based on observed interaction entities.

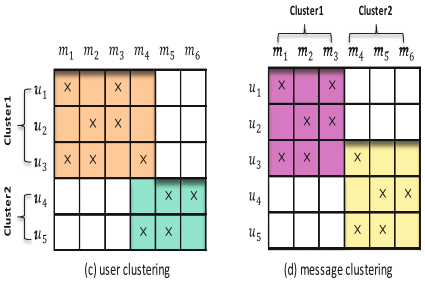



**Fig. 2.** Clustering for users and messages from different dimensions.

user's social footprints and predicting the hidden mechanism underlying diffusion are a critical but challenging task.

A number of research efforts have been performed towards investigating the factors that might affect a user to retweet messages of other users based on user survey [1,2,11,14], statistical analysis [15,17]. Meanwhile, various methods have also been proposed for predicting user's retweeting behavior from different perspectives, such as classifier-based method [7,10], influence-based method [9,19], graph-based method [17]. However, a common property of all the above mentioned models is that users and messages are assumed to be independent each other, respectively. In the real-world scenarios, as social activities grow, social influence and selection lead to formation of different groups, namely users belong to different groups due to the difference of individual preference, and messages belong to different groups due to the difference of referred topic. As a result, we can reach the conclusion that users are more likely to similar each other within the same group than those users who belong to other groups. Messages have the same property. Meanwhile, we argue that the users who belong to the same group are more likely to influence retweeting behavior each other due to their similar interests than these user who belong to other groups. For example in Twitter, a user can create groups of friends, relatives, coworkers and acquaintances that he post and forward on a regular basis. We also investigate that models which take homophily or similarity into account predicts social behavior much better than other more general models which do not take this into account.

Inspired by this, we propose a unified social clustering framework based on matrix factorization method through incorporating user and message clustering information to improve the accuracy of user's retweeting behavior prediction. Specifically, we factorize the user-message retweeting matrix into two intermediated latent matrices: latent user feature matrix and latent message feature matrix. The predicted user-message retweeting matrix is approximated as the product of user feature matrix and message feature matrix under some constraints, as illustrated in Fig. 1. Moreover, we employ clustering information in retweeting prediction to reduce sparsity of data and by doing so to improve accuracy of prediction, as illustrated in Fig. 2. We have conducted experiments

on real social network dataset from Weibo. The results show incorporating cluster information from users and messages can reduce the data sparsity, and our method greatly outperforms the baseline methods by a large margin.

The main contributions of this work can be summarized as follows.

- We formulate the retweeting prediction problem as a predicting missing value task based matrix factorization, namely given the sets of users and messages, our goal is to find who will be retweeted based on partially observed entities.
- We exploit user and message clustering information as regularization terms to constrain objective function to reduce the sparsity of data and improve the performance of prediction.
- With extensive experiments on a real world dataset collected from Weibo, we empirically show the effectiveness and efficiency of our approach. Our approach outperforms state-of-the art methods with a significant margin.

The rest of the paper is organized as follows. Related work is introduced in Sect. 2. Our retweeting prediction model is proposed in Sect. 3 and experimental results are reported in Sect. 4. Conclusion comes in Sect. 5.

## 2 Related Work

There have been significant interests in algorithms for predicting retweeting behavior in social networks [4,5,7,9,12,13,18,20]. Here, we only summarize some representative investigations. For example, Yang et al. [17] proposed a factor graph model to predict user's retweeting behavior by analyzing influence that user, information, and time had on retweeting behavior. Luo et al. [10] employed a learning to rank based framework to discover the users who are most likely to retweet a specific post. Zhang et al. [19] demonstrated the existence of influence locality in social network and predicted user's retweeting behavior based on social influence locality via a logistic regression classifier. Jiang et al. [7] explored a wide range of features, such as user-based, content-based, relationship-based, and time-based, and then used classifier model as the solution to predict retweeting behavior. Most of the above methods are typically based on the effectiveness of leveraging the extracted of features for retweeting prediction. Choosing an appropriate feature set is the most critical part of these algorithms. However, some of these features may be computationally expensive for large social networks.

Recently, some works using matrix factorization for retweeting behavior prediction have been proposed. As far as we know, Wang et al. [16] utilized nonnegative matrix factorization to predict retweeting behavior from user and content dimensions by employing strength of social relationship to constrain objective function. However, this approach does not consider clustering information of user preferences and message referred topics. Hence social relationship undergo the data sparsity and limit the contribution of social regularization. Jiang et al. [6] proposed centroid-based and similarity-based message clustering retweeting prediction models which improve the prediction accuracy. It does not take into account influence from user clustering information.

Therefore, in our work, we give consideration to user and message clustering information from explicit and implicit dimensions, and integrate these clustering factors into matrix factorization model to reduce the data sparsity and improve the performance of retweeting prediction.

## 3 Social Clustering Prediction Model

In this section, we first present a formulation of the problem, and then introduce social clustering information from users and messages to reduce the data sparsity and improve the prediction performance. Finally, we give a unified framework for retweeting prediction, named SCRP (Social Clustering Retweeting Prediction).

### 3.1 Problem Formulation

We first formally define the problem of retweeting behavior prediction from the perspective of matrix factorization. Suppose that we are given $M$ users and $N$ messages, where the $i^{th}$ user denotes as $u_i$ and the $j^{th}$ message denotes as $m_j$. The behaviors of users retweeting messages are represented in an $M \times N$ user-message retweeting matrix $\mathbf{R} = [\mathbf{r}_1, \cdots, \mathbf{r}_N]$, in which each row corresponds to a user and each column corresponds to a message. Meanwhile, whether user decide to retweet a message or not is a binary value task, hence the $(i, j)^{th}$ entry with $\mathbf{R} \in \mathbb{R}^{M \times N}$ can be represented as

$$R_{ij} = \begin{cases} 1 & \text{if } u_i \text{ retweeted } m_j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Then we can model the problem of retweeting behavior prediction as a matrix completion task, where the unobserved entries in matrix $\mathbf{R}$ can be predicted based on the observed retweeting behaviors and other social factors.

Let $\mathbf{U} \in \mathbb{R}^{M \times K}$ be the latent user feature matrix, and $\mathbf{V} \in \mathbb{R}^{K \times N}$ be the latent message feature matrix, where $K$ ($K \ll M, N$) is the number of the latent features. We also assume that each row $U_i = [U_{i1}, U_{i2}, \cdots, U_{iK}]^T$ in $\mathbf{U}$ corresponds to a user and each column $V_j = [V_{1j}, V_{2j}, \cdots, V_{Kj}]^T$ in $\mathbf{V}$ corresponds to a message in latent feature space, respectively.

Now, the retweeting matrix $\mathbf{R}$ can be approximated by the product of two matrices: the latent user feature matrix $\mathbf{U}$ and the latent message feature matrix $\mathbf{V}$, i.e. $\mathbf{R}_{ij} \approx \mathbf{U}_i \mathbf{V}_j$. To learn the optimal latent feature matrices $\mathbf{U}$ and $\mathbf{V}$, we minimize the following objective function based on unobserved entries and observed entries.

$$\min_{U,V} \mathcal{J}(R, U, V) = \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{N} I_{ij} (R_{ij} - U_i V_j)^2 + \frac{\gamma}{2} \|U\|_F^2 + \frac{\lambda}{2} \|V\|_F^2 \tag{2}$$

where $\|\cdot\|_F$ denotes the Frobenius norm fitting constraint, $\gamma$ and $\lambda$ are regularization parameters. In order to focus more on the observed entries, we introduce

an indicator function $I_{ij}$ that is equal to 1 if $u_i$ retweeted $m_j$ and equal to 0 otherwise. The last two regularization terms are added to avoid overfitting.

Due to the severe sparsity of the retweeting matrix $\mathbf{R}$, it is impossible for directly learning the optimal latent spaces for users and messages by relying solely on observed retweeting entries. To alleviate the sparsity problem and improve the accuracy of prediction, we employ user and message clustering information to constraint the objective function.

### 3.2 User Clustering Factor

Users from social network are more likely to form a cohesive group due to social influence and selection. From a social and anthropological standpoint, people who have a common interest preference or a similar lifestyle are probably held together. We also argue that the users from the same group are more likely to similar each other than these users from the other groups. Hence user's interests and behavior pattern can be better represented by other users from the same group in the context of data sparsity.

Based on the above observation, we have the following assumptions that (1) the similar taste preference among users in observed spaces are consistent with the latent spaces; (2) users belonging to the same group should lie close to each other in the latent space; (3) each user can be represented by a linear combination of other users from the same group in the latent space.

In order to reduce the data sparsity and improve the accuracy of prediction, we perform K-means algorithm on the set of users $\mathcal{U}$ before predicting. More precisely, the set of users $\mathcal{U}$ can be divided into $\mathcal{U}_1, \mathcal{U}_2, \cdots, \mathcal{U}_p$, where $\mathcal{U}_i \cap \mathcal{U}_j = \varnothing$ and $p$ is the number of users clustering. To formulate this, we introduce a user clustering sharing matrix $\mathbf{G} \in \mathbb{R}^{M \times M}$ with its $(i,j)^{th}$ entry defined as

$$g_{ij} = \begin{cases} 1 & \text{if } C_{u_i} = C_{u_j} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $C_{u_i}$ and $C_{u_j}$ are the clustering labels of users $u_i$ and $u_j$, respectively. Then, to minimize the latent difference between users $u_i$ and $u_j$ who belong to the same group, we impose a social regularization term

$$\mathcal{J}_1 = \sum_{i=1}^{M} \sum_{j=1}^{M} g_{ij} S_u(i,j) \|U_i - U_j\|_F^2 \tag{4}$$

where $S_u(i,j)$ represents the similarity between $u_i$ and $u_j$.

The similarity can refer to different dimensions. Here, we not only consider the similarity of user's taste preferences, but also take into account the similarity of user's interaction behaviors. The former can be profiled in the content of messages posted by user, and the latter can be reflected in user's social actions (e.g., posting, forwarding, commenting, etc.) which adopt the same message among users may have similar interests. Therefore, the similarity among users can be calculated based on the combine of taste preferences and social behaviors.

In order to calculate explicit taste preferences, we exploit LDA [8], which learns fixed-length feature representations from texts, to learn the vector representations of messages on the collection of user's messages. Then we calculate the taste preferences similarity between users $u_i$ and $u_j$ as following:

$$S_{taste}(i,j) = \frac{I(i)I(j)}{\|I(i)\| \, \|I(j)\|} \tag{5}$$

where $I(i) = \frac{1}{|D(i)|} \sum_{a \in D(i)} T_a$, $D(i)$ is the set of messages posted by user $u_i$, $T_a$ is the learned vector representations for message $a$.

We also have the behavior footprint information of social users, and the behavior similarity between two users can be calculated by measuring the adopted interaction of these two users. To quantitatively measure the behavior similarity, we opt to choose Pearson Correlation Coefficient (PCC) [3], which is proposed to solve this problem that different users have different social action styles.

$$S_{behavior}(i,j) = \frac{\sum_{f \in I(i,j)} (R_{if} - \overline{R}_i) \cdot (R_{jf} - \overline{R}_j)}{\sqrt{\sum_{f \in I(i,j)} (R_{if} - \overline{R}_i)^2} \cdot \sqrt{\sum_{f \in I(i,j)} (R_{jf} - \overline{R}_j)^2}} \tag{6}$$

where $I(i,j)$ denotes the set of messages adopted by both $u_i$ and $u_j$, $\overline{R}_i$ represents the average adopt of user $u_i$. Due to $S_{behavior}(i,j) \in [-1,1]$, we also employ a sigmod function to map behavior similarities into $[0,1]$.

Finally, the similarity between users $u_i$ and $u_j$ is calculated as following:

$$S_u(i,j) = \rho S_{taste}(i,j) + (1-\rho) S_{behavior}(i,j) \tag{7}$$

where $\rho$ is employed to control the contribution of each factor.

### 3.3 Message Clustering Factor

As is mentioned above, messages posted by various users have different structure styles and referred different topics. Therefore, messages can be divided into different groups based on structural and semantic information of texts. We also have the following assumptions that (1) the similar among messages in observed spaces are consistent with the latent spaces; (2) messages are more likely to similar within the same group compare to different latent groups; (3) each message can be represented by a linear combination of other messages from the same group in the latent space. Similarly, we also consider two dimensions of similarity for messages: structural information and semantic information.

Similar to user clustering, we also use K-means algorithm to perform messages clustering. More precisely, the set of messages $\mathcal{M}$ can be grouped into $\mathcal{M}_1, \mathcal{M}_2, \cdots, \mathcal{M}_q$, where $\mathcal{M}_i \cap \mathcal{M}_j = \varnothing$ and $q$ is the number of messages clustering. To formulate this, we also construct a message clustering sharing matrix $\mathbf{H} \in \mathbb{R}^{N \times N}$ with its $(i,j)^{th}$ entry defined as

$$h_{ij} = \begin{cases} 1 & \text{if } C_{m_i} = C_{m_j} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $C_{m_i}$ and $C_{m_j}$ are the clustering labels of messages $m_i$ and $m_j$, respectively. Then, to minimize the latent difference between messages $m_i$ and $m_j$ which belong to the same group, we impose a social regularization term

$$\mathcal{J}_2 = \sum_{j=1}^{N}\sum_{j=1}^{N} h_{ij} S_m(i,j) \|V_i - V_j\|_F^2 \tag{9}$$

where $S_m(i,j)$ represents the similarity between $m_i$ and $m_j$ which can be calculated by the combine of structure and semantic vector of two messages.

There is a wealth of evidence to suggest that structure features from messages are significantly associated with user's retweetability [15]. Here, we extract hashtag, URL, mention as a feature set in our proposed model. Specifically, we use a feature vector $\mathcal{V}_{structure}(j)$=(#hashtag, #URL, #mention) to represent the set of these features, where #hashtag/#URL/#mention denote the number of hashtag/URL/mention occurred for message $m_j$, respectively. Moreover, user's retweeting behavior is strongly correlated with the content of messages. Hence, we also use LDA method to measure the semantic information for messages. For a message $m_j$, we use $\mathcal{V}_{semantic}(j)$ to denote $m_j$'s semantic feature vector. Now, we can combine structural vector $\mathcal{V}_{structure}(j)$ and semantic vector $\mathcal{V}_{semantic}(j)$ for message $m_j$ into a compound vector $\mathcal{V}(j)$. In addition, we also use the two vectors mentioned above to calculate the similarities among messages. More specifically, the similarity between messages $m_i$ and $m_j$ is calculated as

$$S_m(i,j) = \lambda S_{structure}(i,j) + (1-\lambda) S_{semantic}(i,j) \tag{10}$$

where $\lambda$ is the parameter controlling the contribution of each factor. $S_{structure}$ $(i,j)$ and $S_{semantic}(i,j)$ are cosine similarities based on structural and semantic vectors, respectively.

### 3.4 Unified Prediction Model

Based on the above discussed, we demonstrate how to construct user clustering regularization and message clustering regularization, respectively. Now, we solve the optimization problem by combining $\mathcal{J}_1, \mathcal{J}_2$ with $\mathcal{J}$:

$$\begin{aligned}\min_{U,V} \mathcal{J}(R,U,V) = & \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N} I_{ij}(R_{ij} - U_i V_j)^2 \\ & + \frac{\alpha}{2}\sum_{i=1}^{M}\sum_{k=1}^{M} g_{ik} S_u(i,k)\|U_i - U_k\|_F^2 \\ & + \frac{\beta}{2}\sum_{j=1}^{N}\sum_{l=1}^{N} h_{jl} S_m(j,l)\|V_j - V_l\|_F^2 \\ & + \frac{\gamma}{2}\|U\|_F^2 + \frac{\eta}{2}\|V\|_F^2\end{aligned} \tag{11}$$

where $\alpha \geq 0$ and $\beta \geq 0$ are the parameters controlling user clustering regularization and message clustering regularization on $U_i$ and $V_j$, respectively.

A local minimum of the objective function given by Eq. (11) can be found by employing gradient descent method in feature vectors $U_i$ and $V_j$, respectively.

$$\begin{aligned} \frac{\partial \mathcal{J}}{\partial U_i} &= \sum_{j=1}^{N} I_{ij}(U_i V_j - R_{ij})V_j + \gamma U_i + \alpha \sum_{k=1}^{M} g_{ik} S_u(i,k)(U_i - U_k) \\ \frac{\partial \mathcal{J}}{\partial V_j} &= \sum_{i=1}^{M} I_{ij}(U_i V_j - R_{ij})U_i + \eta V_j + \beta \sum_{l=1}^{N} h_{jl} S_m(j,l)(V_j - V_l) \end{aligned} \tag{12}$$

## 4 Experimental Analysis

### 4.1 Dataset Description

We use a publicly available dataset released by [19] to evaluate the performance of our model. The dataset was collected from Weibo, which allows users to follow other users and receive messages from followed users. Like Twitter, it also provides retweeting function to encourage users to spread information. Specifically, in this paper, we randomly sample 10,000 messages retweeted by 690,787 users from the above dataset. Since the dataset doesn't contain the messages published by retweeters, we also collect messages posted by retweeters in order to calculate similarities among retweeters. Table 1 lists statistics of the dataset used in this paper.

**Table 1.** Retweeting data statistics

| Dataset | #Users | #Retweeter's tweets | #Tweets | #Retweets | Sparseness |
|---|---|---|---|---|---|
| Weibo | 690,787 | 131,129,186 | 10,000 | 1,435,720 | 0.02 % |

### 4.2 Experimental Settings

For the above dataset, we randomly sample 80 % of the retweetings from user-message retweeting matrix as the training data to predict the remaining 20 % of retweetings. The corresponding entries in $\mathbf{R}$ of positive instances for testing data are set to 0. We determine the number of clusters in the proposed model using rule of thumb: $k \approx \sqrt{n/2}$ with $n$ as the number of users/messages. Meanwhile, we empirically set the number of topics to 100 and parameters $\rho = \lambda = 0.5$.

### 4.3 Comparative Algorithms

We implement the following baselines for comparison with our social clustering based retweeting prediction model (SCRP).

- **Naive Bayes**: The retweeting predication can be considered as a binary classification task, where each message is labelled either positive or negative instance to represent whether it will be retweeted or not.
- **LRC-BQ**: The method proposes a notion of social influence locality based on pairwise influence and structural diversity, and then uses a logistic regression classifier to predict user's retweeting behavior [19].
- **MNMFRP**: This method utilizes nonnegative matrix factorization to predict retweeting behavior from user and content dimensions, respectively, by using strength of social relationship to constrain objective function [16].
- **CRPM & IRPM**: The two methods use the clustering relationships of messages to predict retweeting behavior based on matrix factorization [6]. These models don't take into account clustering information from users.
- **SCRP-U**: This method only considers user clustering information in our proposed retweeting prediction model.
- **SCRP-M**: This method only utilizes message clustering information for user retweeting prediction model.

### 4.4 Evaluation Measures

To quantitatively evaluate the performance of the proposed model, we divide the constructed data set into training and test data, and perform 10-fold cross validation to alleviate the effects of random selection. We evaluate the performance of retweeting prediction in terms of Precision, Recall, $F_1$-score, and Accuracy.

### 4.5 Parameter Settings

In this section, we will investigate the effect of different parameter settings for our proposed model, including tradeoff parameters, dimension of latent features, and number of projected gradient iterations, on the performance.

**Tradeoff Parameters:** In our proposed method in this paper, the tradeoff parameters $\alpha$, $\beta$, $\gamma$ and $\eta$ play the role of adjusting the strengths of different terms in the objective function. They control how much our method should incorporate the clustering information for retweeting prediction model. Taking the scales of $U$ and $V$ into account, we scan orders of magnitude and try different combinations of parameters as shown in Table 2. The results in Table 2 show that the parameter set $\alpha = \beta = \gamma = \eta = 10^{-4}$ produce the best performance. In our following experiments, we just use this parameter setting.

**Number of Latent Features:** To find a K-dimensional joint latent space for users and messages, we train $U$ and $V$ using gradient descent method. More specifically, we conduct extensive experiments with $K$ from 2 to 80 on the constructed dataset. The results are shown in Fig. 3, from which we can see conclude that with the latent feature number $K$ increasing, $F_1$-score increases gradually. We can also observe that $F_1$-score grow more slowly when $K > 50$. Considering the computation efficiency and storage cost, we choose $K = 50$ as the latent

**Table 2.** Tradeoff parameters on Weibo dataset (50 Hidden Features and 50 Iterations)

| $\alpha$ | $\beta$ | $\gamma$ | $\eta$ | $F_1$-score | Accuracy |
|---|---|---|---|---|---|
| $10^{-6}$ | $10^{-6}$ | $10^{-6}$ | $10^{-6}$ | 0.808 | 0.768 |
| $10^{-5}$ | $10^{-5}$ | $10^{-5}$ | $10^{-5}$ | 0.835 | 0.808 |
| $10^{-4}$ | $10^{-4}$ | $10^{-5}$ | $10^{-5}$ | 0.839 | 0.825 |
| $10^{-4}$ | $10^{-4}$ | $10^{-4}$ | $10^{-4}$ | **0.847** | **0.831** |
| $10^{-4}$ | $10^{-4}$ | $10^{-3}$ | $10^{-3}$ | 0.823 | 0.788 |
| $10^{-3}$ | $10^{-3}$ | $10^{-3}$ | $10^{-3}$ | 0.816 | 0.780 |

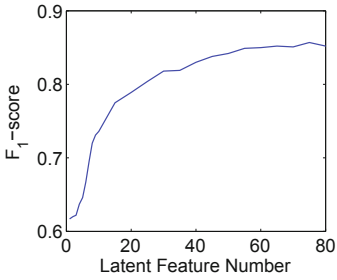



**Fig. 3.** Latent feature number on Weibo dataset (50 Iterations)

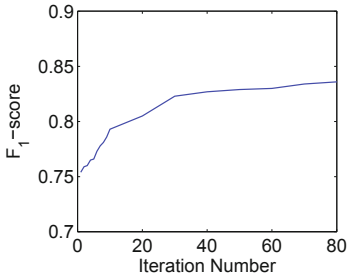



**Fig. 4.** Iteration number on Weibo dataset (50 Hidden Features)

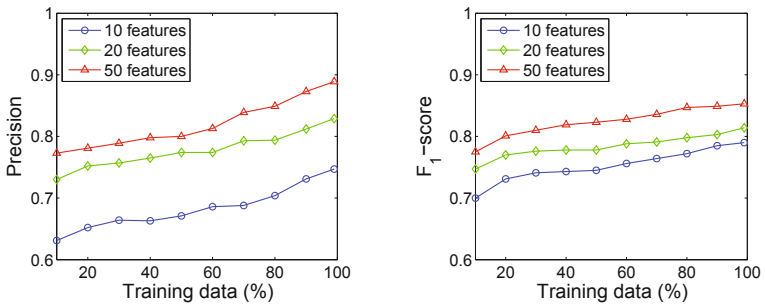
space dimension in our experiments. Although it is not the perfect one, the following experiments demonstrate it is adequate.

**Number of Iterations:** When using gradient descent method to solve the objective function, we need to predefine a proper number of updating iterations to get a good performance while avoid overfitting. Figure 4 illustrates the impacts of the number of iterations on $F_1$-score. Considering the trade-off between the computational efficiency and the accuracy of prediction, we conduct 50 iterations for the solution in our experiments.

### 4.6 Effect of Sparseness

Based on the above parameter settings, we further exploit different training data sets to test the sensibility of the proposed model on constructed dataset. For example, training data 80 % means we randomly select 80 % of the retweeting behavior instances from user-message retweeting matrix as the training data to predict the remaining 20 % of retweeting entities. The overall performance of our proposed approach with different training set is illustrated in Fig. 5. From these figures, we can see conclude that the performance of our proposed SCRP method improves gradually as the number of training positive instances increase.

Moreover, we have also the following observation that the performances of our model change within a narrow range in the dataset with different sparseness which shows our model have good robustness. In general, social clustering based retweeting prediction performs better when observed retweeting instances are relatively more in the training data. This indicates that each user/message can be better represented by a linear combination of other users/messages from the same group in the latent space.




(a) Precision v.s. Training Data (b) $F_1$-score v.s. Training Data

**Fig. 5.** Different training data settings to test our proposed SCRP model.

### 4.7 Prediction Performance

Our goal is to find who will be retweeted based on partially observed retweeting instances. Therefore, in this section, we will demonstrate the prediction performance of the proposed method, and compare it with other methods. Specifically, we set the optimal parameters when running the baselines. Then all experiments are performed 5 runs with the 50 dimensions to represent the latent features. We list the average results of each method in Table 3. Noted that both LRC-BQ model [16] and MNMFRP model [19] use the same original dataset with us. From these results, we can observe the following conclusions: (1) The proposed SCRP model, which incorporates user and message clustering factors together, significantly outperforms the baseline methods in our experimental results; (2) The prediction performance of SCRP is better than (CRPM &IRPM), which reveals that user clustering information is effectiveness of factor for retweeting prediction; (3) The comparison between SCRP-U v.s. MNMFRP reveals that the strategy of incorporating user clustering information to predict missing entities in the objective function is more effective compared with considering the strength of social relationship. In general, incorporating user and message clustering information can reduce the sparsity of data and improve the performance of prediction.

**Table 3.** Performance of retweeting behavior prediction.

| Method | Precision | Recall | $F_1$-score | Accuary |
|---|---|---|---|---|
| Naive Bayes | 0.562 | 0.555 | 0.558 | 0.555 |
| LRC-BQ | 0.698 | 0.770 | 0.733 | 0.719 |
| MNMFRP | 0.796 | 0.791 | 0.793 | N/A |
| CRPM | 0.814 | 0.833 | 0.823 | 0.821 |
| IRPM | 0.817 | 0.833 | 0.825 | 0.823 |
| SCRP-U | 0.846 | 0.809 | 0.827 | 0.809 |
| SCRP-M | 0.847 | 0.811 | 0.829 | 0.811 |
| SCRP | **0.863** | **0.831** | **0.847** | **0.831** |

## 5 Conclusion

In this paper, we propose a novel method, which incorporates the users and messages clustering information together, to predict user's retweeting behavior. The proposed model measures the similarities among users and messages using an ensemble from explicit and implicit dimensions, and then utilizes matrix factorization method to predict unobserved retweeting behaviors by employing cluster information of users and messages to constrain objective function. Experimental results demonstrate that the proposed method can achieve better performance than state-of-the-art methods.

**Acknowledgments.** This work was supported by National Key Technology R & D Program(No.2012BAH46B03), and the Strategic Leading Science and Technology Projects of Chinese Academy of Sciences(No.XDA06030200).

## 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. Boyd, D., Golder, S., Lotan, G.: Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In: HICSS, pp. 1–10 (2010)
3. Breese, J.S., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. In: UAI, pp. 43–52 (1998)
4. Can, E.F., Oktay, H., Manmatha, R.: Predicting retweet count using visual cues. In: CIKM, pp. 1481–1484 (2013)
5. Feng, W., Wang, J.: Retweet or not?: personalized tweet re-ranking. In: WSDM, pp. 577–586 (2013)
6. Jiang, B., Liang, J., Sha, Y., Wang, L.: Message clustering based matrix factorization model for retweeting behavior prediction. In: CIKM, pp. 1843–1846 (2015)
7. Jiang, B., Sha, Y., Wang, L.: A multi-view retweeting behaviors prediction in social networks. In: Cheng, R., Cui, B., Zhang, Z., Cai, R., Xu, J. (eds.) Web Technologies and Applications. LNCS, vol. 9313, pp. 756–767. Springer, Heidelberg (2015)

8. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
9. Liu, L., Tang, J., Han, J., Jiang, M., Yang, S.: Mining topic-level influence in heterogeneous networks. In: CIKM, pp. 199–208 (2010)
10. Luo, Z., Osborne, M., Tang, J., Wang, T.: Who will retweet me? Finding retweeters in Twitter. In: SIGIR, pp. 869–872 (2013)
11. Metaxas, P., 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)
12. Naveed, N., Gottron, T., Kunegis, J., Alhadi, A.C.: Bad news travel fast: A content-based analysis of interestingness on twitter. In: WebSci, p. 8 (2011)
13. Petrovic, S., Osborne, M., Lavrenko, V.: RT to win! Predicting message propagation in Twitter. In: ICWSM (2011)
14. Recuero, R., Araujo, R., Zago, G.: How does social capital affect retweets? In: ICWSM (2011)
15. 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 (2010)
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, Z., Guo, J., Cai, K., Tang, J., Li, J., Zhang, L., Su, Z.: Understanding retweeting behaviors in social networks. In: CIKM, pp. 1633–1636 (2010)
18. Zaman, T.R., Herbrich, R., Van Gael, J., Stern, D.: Predicting information spreading in twitter. In: NIPS, pp. 17599–601 (2010)
19. 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)
20. Zhang, Q., Gong, Y., Guo, Y., Huang, X.: Retweet behavior prediction using hierarchical dirichlet process. In: AAAI, pp. 403–409 (2015)