Retweeting Prediction Using Matrix Factorization with Binomial Distribution and Contextual Information

Bo Jiang[1], Zhigang Lu[1,2], Ning Li[1(✉)], Jianjun Wu[1,2], Feng Yi[3], and Dongxu Han[1,2]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
{jiangbo,luzhigang,lining6,wujianjun,handongxu}@iie.ac.cn
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3] School of Computer Engineering, University of Electronic Science and Technology of China, Zhongshan, China
bjstarbow@gmail.com

Abstract. Retweeting provides an efficient way to expand information diffusion in social networks, and many methods have been proposed to model user's retweeting behaviors. However, most of existing works focus on devising an effective prediction method based on social network data, and few research studies explore the data characteristic of retweeting behaviors which is typical binary discrete distribution and sparse data. To this end, we propose two novel retweeting prediction models, named Binomial Retweet Matrix Factorization (BRMF) and Context-aware Binomial Retweet Matrix Factorization (CBRMF). The two proposed models assume that retweetings are from binomial distributions instead of normal distributions given the factor vectors of users and messages, and then predicts the unobserved retweetings under matrix factorization. To alleviate data sparsity and reduce noisy information, CBRMF first learns user community by using community detection method and message clustering by using short texts clustering algorithm from social contextual information on the basis of homophily assumption, respectively. Then CBRMF incorporates the impacts of homophily characteristics on users and messages as two regularization terms into BRMF to improve the prediction performance. We evaluate the proposed methods on two real-world social network datasets. The experimental results show BRMF achieves better the prediction accuracy than normal distributions based matrix factorization model, and CBRMF outperforms existing state-of-the-art comparison methods.

Keywords: Retweeting prediction · Social network · Binomial distribution · Contextual information · Matrix factorization

G. Li et al. (Eds.): DASFAA 2019, LNCS 11447, pp. 121–138, 2019.
https://doi.org/10.1007/978-3-030-18579-4_8

1 Introduction

Recent years have witnessed an increasing amount of social network services such as Twitter, Facebook and Weibo. One of the distinguishing features of these services is the retweeting mechanism which forwards messages published by other users and shares them with one's own followers. Most of existing studies have shown that retweeting is considered as a key mechanism of information diffusion in social networks [1,18]. Taking full advantage of the function, one can achieve better insights into the process of information diffusion, making a right strategic decision for many tasks of social network applications such as event discovery [10], community detection [3,28], and recommender system [5,16]. Thus, understanding the influencing factors of retweetings from the observed data and predicting the hidden mechanism underlying diffusion is a critical and fundamental task in these applications.

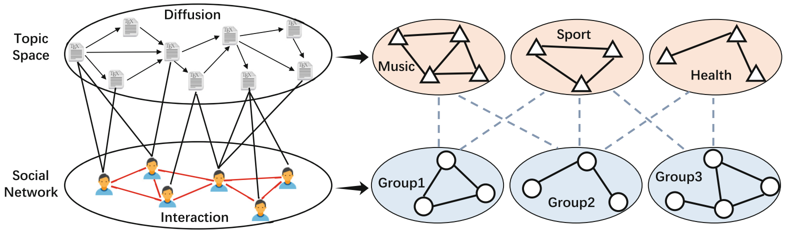


Fig. 1. An example of social groups on social network. Users have diverse topical interests, each group corresponding to a user community with common interests. Messages also have the diversity topics, each topic corresponding to a topic space with similar semantic content.

Considerable work has been carried out on investigating the influencing factors of retweeting decision through user survey [1,18] and statistical analysis [23,30]. These studies find that the interests of user's topics and the strength of social influence are two most important influencing factors when user decides to retweet a message. Compared with the above efforts, more and more studies have been getting to focus on devising an effective retweeting prediction model from different perspectives. For example, a simple but powerful strategy is to consider the retweeting prediction as a classification problem by extracting the different features like user's profile, network structure and message's content, historical interactions [2,9,12]. Although these studies can solve the problem of retweeting prediction to some extent, there is lack of a more principled way to extract the set of related features. The social influence-based models are also proposed to quantify the strength of user influence from the views of network structure and historical interactions [11,32]. However, these models depend on the different types of information, which may not be always available in some platforms. The factor graph-based methods are used to represent factorization of retweeting probability distribution function [19,33]. However, such models are too complex and difficult to be applicable to large-scale applications. Recently, the attention-based deep neural network method is proposed by incorporating social contextual information for this task [34]. However, the training process of neural networks requires a large amount of labeled data, it is not always available in most social networks. Another matrix factorization-based methods convert the problem into matrix completion based on the observed entities [25,26]. These methods leverage the power of matrix factorization model to achieve a relatively high accuracy of prediction by incorporating explicit information and implicit feedback. However, none of methods focus on the data characteristic of retweeting behaviors with binary distributions on social networks.

The present findings demonstrate that social users and information flows naturally form social communities and topic clusters underlying network structure and interactions, as shown in Fig. 1. Moreover, many social networks such as Facebook and Google+ provide "social group" function, which allows the users to be incorporated into a new densely connected subgraph with the same or similar personal interest preference. In this case, user's decision is strongly influenced by the other active neighbors from his friends [32]. On the other hand, information flows can also be mapped into the topic space where clusters of messages form topics. Therefore, we argue that the problem of retweeting prediction should be modeled by considering discrete distribution of social data and the impacts of homophily characteristics on users and messages.

In this paper, we propose two novel matrix factorization methods for the retweeting prediction, named Binomial Retweet Matrix Factorization (BRMF) and Context-aware Binomial Retweet Matrix Factorization (CBRMF). The two models assume retweetings are from binomial distributions instead of normal distributions. CBRMF is an extended BRMF model by considering social circles and message clustering to alleviate the data sparsity and reduce the noisy information based on the impacts of homophily.

Contributions. In this paper, we make the following three contributions:

- We propose two novel matrix factorization methods for retweeting prediction problem. The two methods assume retweetings are from binomial distributions instead of normal distributions. To the best of our knowledge, this is the first work for retweeting prediction to exploit binomial distributions with the matrix factorization model.
- We utilize user community and document clustering as contextual regularization terms into binomial retweet matrix factorization based on homophily assumption to alleviate the data sparsity and reduce the noisy information, as well as improve the performance of prediction.
- We conduct several analysis experiments with two real-world social network datasets, the experimental results demonstrate our proposed models outperform state-of-the art comparison methods.

Outline. The rest of this paper is organized as follows. Section 2 introduces the related work. Section 3 provides the formal definition of the retweeting prediction problem and introduces notations used in this paper. In Sect. 4, we present two retweeting prediction model based on binomial distributions and its learning and inference procedures. Section 5 evaluates our proposed models with the real large social network datasets in terms of accuracy. Section 6 gives a conclusion of this work.

2 Related Work

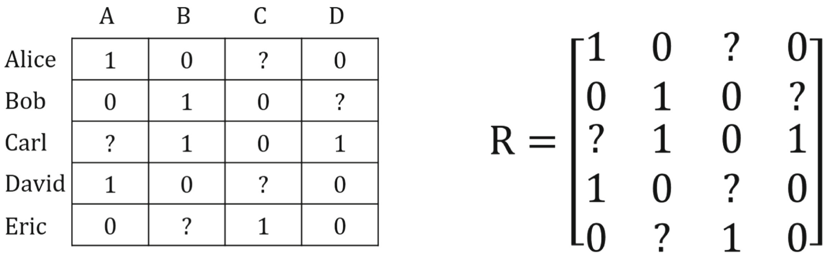
Social Recommendation with Matrix Factorization. Recommender systems are used as an efficient tool for dealing with the information overload problem. Many social recommendation approaches have been developed in recent years. For example, SoRec [13] employs both users' social network information and rating records to solve the data sparsity and poor prediction accuracy problems based on probabilistic matrix factorization. Similarly, Context MF [8] incorporates individual preference and interpersonal influence based on probabilistic matrix factorization for improving the accuracy of social recommendation. Ensemble methods predict a missing rating for a given user by a linear combination of ratings from the user and the social network. STE [14] represents the formulation of the social trust restrictions by fusing the users' tastes and their trusted friends' favors together on the recommender systems. mTrust [24] studies multi-faceted trust relationships between users for rating prediction. Regularization methods focus on a user's preference and assume that a user's preference should be similar to that of her social network. SocialMF [6] incorporates the mechanism of trust propagation into the matrix factorization approach for recommendation in social networks. Social Regularization [15] imposes social regularization terms to constrain matrix factorization objective functions based on users' social friend information. TBPR [27] studies the effects of distinguishing strong and weak ties by using neighbourhood overlap to approximate tie strength in social recommendation. In summary, social recommendation methods have been successfully applied in the missing values prediction tasks.

Retweet Behavior Modeling. Existing studies focus on exploring the influencing factors of retweeting behaviors by conducting user survey [1,18] and performing empirical analysis [17,22,29]. These results indicate that the intention of user retweeting message is positively influenced by sharing the informative content of message, and enhancing social influence from social relationships. On the other hand, more efforts have been guided on modeling how a message be retweeted in social network. For instance, feature-based methods take the set of related features to predict retweeting behavior such as social features [7,12], visual features [4]. Social influence-based methods also have been proposed to model user retweeting behavior [32]. Most of the above methods apply heuristic methods to extract the set of features for retweeting prediction. However, some of these features may be computationally expensive or not always available in some social networks. Zhang et al. [33] propose a novel nonparametric Bayesian model adapted from the hierarchical Dirichlet process to combine textual, structural, and temporal information for the task. Subsequently, Zhang et al. [34]

also propose a attention-based deep neural network retweet model by incorporating the user, author, user interests, and similarity information between the tweets and user interests. Besides, other studies employ matrix factorization by using social contextual information from user and content dimensions to solve the problem [26]. Nevertheless, no research considers discrete distributions of retweeting values on social networks. The retweeting prediction has still some unsolved problems such as exploring the data characteristic and studying the role of group structure on users and messages.

3 Problem Preliminaries

We give some necessary notations used in this paper and present a formal representation of the retweeting prediction problem under the probabilistic matrix factorization model.



	A	B	C	D
Alice	1	0	?	0
Bob	0	1	0	?
Carl	?	1	0	1
David	1	0	?	0
Eric	0	?	1	0

(a) An example of a binary data with unknowns. Users are presented in rows, while messages are in columns.

$$R = \begin{bmatrix} 1 & 0 & ? & 0 \\ 0 & 1 & 0 & ? \\ ? & 1 & 0 & 1 \\ 1 & 0 & ? & 0 \\ 0 & ? & 1 & 0 \end{bmatrix}$$

(b) Retweeting data can be represented in the matrix R, which is usually sparse with a high percentage of missing values.

Fig. 2. Matrix representation with retweeting data.

Given M users and N messages, the behaviors of users retweeting messages are represented in an $M \times N$ retweeting matrix $R = [\mathbf{r}_1, \cdots, \mathbf{r}_N]$, in which each row corresponds to a user and each column corresponds to a message. Retweeting has only two states, where R_{ij} takes the value of 1 if u_i retweets m_j and 0 otherwise. Let $U \in \mathbb{R}^{K \times M}$ and $V \in \mathbb{R}^{K \times N}$ be the latent user and message feature matrices respectively, where U_i represents a user and V_j represents a message in latent feature space. K is the number of the latent features. The likelihood function of the observed retweetings is factorized across M users and N messages with each factor based on Probabilistic Matrix Factorization (PMF) [20] as

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

where $\mathcal{N}(\cdot|\mu, \sigma^2)$ is the probability density function of the Gaussian distribution with mean μ and variance σ^2.

PMF is learned by maximizing posterior probability of matrices U and V, which is equivalent to minimizing sum-of-squares of factorization error as

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

where the prior distributions over U and V are assumed to zero-mean Gaussian, $(\|U\|_F^2 + \|V\|_F^2)$ can prevent overfitting, $|| \cdot ||_F$ denotes the Frobenius norm of the matrix.

It's known that PMF assumes that all ratings are from normal distributions given the corresponding user and item factor vectors. However, the normal assumption in the retweeting prediction problem is not suitable since the values of retweetings only have 0s and 1s, as shown in Fig. 2. For this reason, we will explore how to use binomial distributions to solve the problem in the following section.

4 The Proposed Model

With the assumption that retweetings are from binomial distributions, we propose Binomial Retweet Matrix Factorization (BRMF) and Context-aware Binomial Retweet Matrix Factorization (CBRMF) for the retweeting prediction task. CBRMF is an extended BRMF model by incorporating user community and document clustering as social contextual regularization terms to alleviate the data sparsity and reduce the noisy information and improve the performance of prediction. The detailed descriptions of our proposed models are as follows.

4.1 Binomial Retweet Matrix Factorization

Since retweeting has only two states, i.e., {retweet, not retweet} in our datasets, the assumption of normal distributions sampled from retweet data doesn't draw the retweeting behaviors. Instead, we assume that all retweetings are from binomial distributions with different preference parameters.

The binomial distributions for retweetings satisfy the following conditions: (1) retweetings are independent and identically distributed, (2) retweeting has only two choices, and (3) the retweeting probability of each user is approximately equal to the ratio of the occurring time of retweetings and the total number of retweetings in our datasets. Here, we replace the Gaussian distribution with the Bernoulli distribution in Eq. (1) as

$$P(R|U,V) = \prod_{i=1}^{M} \prod_{j=1}^{N} \mathcal{B}(R_{ij}|\mathcal{S}, U_i^T V_j) \tag{3}$$

where $\mathcal{B}(k|n,p)$ is the binomial probability mass function with parameters n and p, and $\mathcal{S}$ is the number of retweetings in our datasets. For given a user u_i and a message m_j, our goal is to maximize the probability of u_i retweets m_j as

$$P(R|U,V) = \prod_{i=1}^{M} \prod_{j=1}^{N} p(R_{ij})^{I_{ij}}(1 - p(R_{ij}))^{1-I_{ij}} \tag{4}$$

where $p(\cdot)$ is the probability that u_i retweets m_j. $I \in \{0,1\}$ is an indicator matrix where I_{ij} is equal to 1 if u_i retweets m_j and 0 otherwise. The sigmoid function $g(x) = 1/(1 + exp(-x))$ bound the range of $U_i^T V_j$ denoting the user-message association.

To learn the model, we maximize the following likelihood objective function as

$$P(R|U,V) = \prod_{i=1}^{M}\prod_{j=1}^{N}(g(R_{ij})^{I_{ij}}(1 - g(R_{ij}))^{1-I_{ij}})) \tag{5}$$

The log of posterior distribution with Eq. (5) is given by

$$\begin{aligned}\mathcal{L}(U,V|R) = &\sum_{i=1}^{M}\sum_{j=1}^{N}[I_{ij} ln\frac{g(R_{ij})}{1 - g(R_{ij})} + ln(1 - g(R_{ij}))] \\ &- \frac{1}{2\sigma_U^2}\sum_{i=1}^{M} U_i^T U_i - \frac{1}{2\sigma_V^2}\sum_{j=1}^{N} V_j^T V_j + Const\end{aligned} \tag{6}$$

where users and messages draw zero-mean Gaussian distributions.

Maximizing Eq. (6) is equivalent to minimizing the following objective function as

$$\sum_{i=1}^{M}\sum_{j=1}^{N}[ln(e^{U_i^T V_j} + 1) - U_i^T V_j I_{ij}] + \frac{\lambda}{2}(\sum_{i=1}^{M}\|U_i\|_F^2 + \sum_{j=1}^{N}\|V_j\|_F^2) \tag{7}$$

where $(||U_i||_F^2 + ||V_j||_F^2)$ are also to avoid overfitting.

The local minimum of the objective function given by Eq. (7) can be found by performing stochastic gradient descent (SGD) approach on feature vectors U_i and V_j as

$$\frac{\partial \mathcal{L}}{\partial U_i} = \sum_{j=1}^{N}\frac{e^{U_i^T V_j}}{e^{U_i^T V_j} + 1} V_j - I_{ij} V_j + \lambda U_i \tag{8}$$

$$\frac{\partial \mathcal{L}}{\partial V_j} = \sum_{i=1}^{M}\frac{e^{U_i^T V_j}}{e^{U_i^T V_j} + 1} U_i - I_{ij} U_i + \lambda V_j \tag{9}$$

4.2 Context-Aware Binomial Retweet Matrix Factorization

As mentioned above, contextual information is an indispensable factor for retweeting prediction due to its effect on users' decisions. Thus, we propose Context-aware Binomial Retweet Matrix Factorization (CBRMF) by considering user community and document clustering learned from contextual information for retweeting behaviors.

User Community Modeling. According to sociology and psychology, users under the effect of network structure and information diffusion together gradually form communities corresponding to close social circles or interest groups with the similar personal preference. In this case, behaviors are more likely to be effected each other in the same social community, e.g., some users like to follow others's message and are easily influenced by the active neighbors' decisions.

Based on the above facts, we introduce user community with the hypothesis that the users that are similar in hidden user space have similar personal preferences. We here apply a new community detection algorithm building upon the distance dynamics [21], which automatically spots communities in a network by examining the changes of distances among nodes. Specifically, we first construct an undirected interaction graph $G = (V, E, W)$, where V is the set of users, E is the set of edges and W is the corresponding set of weights. $e = \{u, v\} \in E$ indicates a social relationship between the users u and v. $w(u, v)$ denotes the weight of edge e. There are various explicit and implicit relations in social networks. For instance, following represents that a user pays close attention to another user, and retweeting reflects that two users appear in the same message with action relevancy. These behaviors can be modeled as the interaction relation graph G, in which w_{ij} is the association weight measuring the co-occurrences of users u_i and u_j, i.e. $w_{ij} = \sum(\#\text{follow}, \#\text{retweet}, \#\text{mention})$ to represent the sum of occurrence times with these actions, where # denotes the number of occurrences for users u_i and u_j given an action. In particular, we set $w_{ij} = 1$ in case of the users u_i and u_j only connected by following relationship.

After constructing the interaction graph, the next crucial step is to obtain the user communities by performing Attractor method [21]. For the Attractor algorithm, the cohesion parameter λ is used to determine the positive or negative interaction influence on the distances from exclusive neighbors. We use the implementation provided by the authors and the recommended settings as in their paper[1]. Once the clustering is done, we denote user community matrix as $W \in \mathbb{R}^{M \times M}$, where W_{ij} takes the value of 1 if C_{u_i} and C_{u_j} belong to the same community and 0 otherwise.

The user communities make different users with the same group become similar in the latent hidden space. Then we can arrive at the following user social community regularizer as

$$\mathcal{L}_1(U) = \|W - g(U^T U)\|_F^2 \tag{10}$$

where the same community for users indicates the two users should be very close and could be large otherwise.

Document Clustering Modeling. An empirical observation is the documents with similar content in observed space have similar semantic distance in hidden space, and the similar messages have a high retweeting probability when they have retweeted in the past. However, short text on social networks is very sparse and exists the noisy data, making it hard to find sufficient statistical factors to discover syntactic and semantic dependencies. Here, to alleviate the problem

[1] https://github.com/YcheCourseProject/CommunityDetection.

of data sparseness and noisy, we use a collapsed Gibbs sampling method by Dirichlet multinomial mixture (DMM) model for short text clustering, called GSDMM [31]. The model has some good property for the problem of short text clustering, such as fast to converge and cope with the sparse and high dimensional problem of short texts. The model code used are publicly available[2].

More specifically, we cluster all short texts into different groups by using GSDMM model that documents are similar to one another within the same cluster and are dissimilar to documents in other clusters. After clustering, in our proposed model, we represent document clustering matrix as $H \in \mathbb{R}^{N \times N}$, where H_{ij} takes the value of 1 if C_{d_i} and C_{d_j} belong to the same clustering and 0 otherwise.

Similarly, once document clusters are finished, we may arrive at the following document similarity cluster regularizer

$$\mathcal{L}_2(V) = \|H - g(V^T V)\|_F^2 \tag{11}$$

where the same group for documents indicates the latent distance should be very close and could be large otherwise.

Prediction Approach. We demonstrate how to construct user community and document clustering regularization terms in the above section. Next, we factorize user and message latent factors with matrices W and H collaboratively as

$$\begin{aligned}\mathcal{L}(U, V|R) = &\sum_{i=1}^{M}\sum_{j=1}^{N} ln[g(R_{ij})^{I_{ij}}(1 - g(R_{ij})^{1-I_{ij}}))] \\ &- \frac{1}{2\sigma_W^2}\sum_{i=1}^{M}\sum_{k=1}^{M} I_{ik}^{(W)}(W_{ik} - g(U_i^T U_k))^2 \\ &- \frac{1}{2\sigma_H^2}\sum_{j=1}^{N}\sum_{k=1}^{N} I_{jk}^{(H)}(H_{jk} - g(V_j^T V_k))^2 \\ &- \frac{1}{2\sigma_U^2}\sum_{i=1}^{M} U_i^T U_i - \frac{1}{2\sigma_V^2}\sum_{j=1}^{N} V_j^T V_j + Const\end{aligned} \tag{12}$$

where I^W is a user indicator matrix where I_{ij}^W is equal to 1 if users u_i and u_i belong to the same social group and 0 otherwise, and I^H is a message indicator matrix where I_{ij}^H is equal to 1 if message m_i and m_j belong to the same topic group and 0 otherwise.

[2] https://github.com/rwalk/gsdmm.

Similarly, maximizing the posterior distribution is equivalent to minimizing the sum-of-squared errors function with hybrid quadratic regularization terms:

$$
\begin{aligned}
\min_{U,V} \mathcal{L}(U,V|R) = &\sum_{i=1}^{M}\sum_{j=1}^{N}[ln(e^{U_i^T V_j}+1) - U_i^T V_j I_{ij}] \\
&+ \frac{\alpha}{2}\sum_{i=1}^{M}\sum_{k=1}^{M} I_{ik}^{(W)}(W_{ik} - g(U_i^T U_k))^2 \\
&+ \frac{\beta}{2}\sum_{j=1}^{N}\sum_{k=1}^{N} I_{jk}^{(H)}(H_{jk} - g(V_j^T V_k))^2 \\
&+ \frac{\lambda}{2}(\sum_{i=1}^{M}\|U_i\|_F^2 + \sum_{j=1}^{N}\|V_j\|_F^2)
\end{aligned}
\qquad (13)
$$

Training Model. Since the objective function is convex with regard to each parameter, a local minimum can be achieved by updating each parameter iteratively. We also directly use SGD method to update the feature vectors U_i and V_j given by Eq. (13) as

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial V_j} = &\sum_{i=1}^{M}\frac{e^{U_i^T V_j}}{e^{U_i^T V_j}+1}U_i - I_{ij}U_i + \lambda V_j \\
&+ \beta\sum_{k=1}^{N} I_{jk}^{(H)} g'(V_j^T V_k)(g(V_j^T V_k) - H_{jk})V_k
\end{aligned}
\qquad (14)
$$

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial V_j} = &\sum_{i=1}^{M}\frac{e^{U_i^T V_j}}{e^{U_i^T V_j}+1}U_i - I_{ij}U_i + \lambda V_j \\
&+ \beta\sum_{k=1}^{N} I_{jk}^{(H)} g'(V_j^T V_k)(g(V_j^T V_k) - H_{jk})V_k
\end{aligned}
\qquad (15)
$$

where $g'(x) = exp(-x)/(1+exp(-x))^2$ is the derivative of $g(x)$. In each iteration, U and V are updated based on the latent variables from the previous iteration.

5 Experiments

5.1 Experimental Settings

We use two real-world social network datasets to evaluate the validity of our proposed model. Weibo is one of the most popular social network platforms in China. In this paper, we use publicly available Weibo dataset [32]. The dataset contains 1,787,443 users, and 300,000 popular messages and 23,755,810 retweetings. In term of messages, we randomly choose 100,000 popular messages and extract the corresponding relationship and retweetings as our experimental dataset. Besides, we also collect Twitter data using the RESTful APIs with the crawling process from August 10, 2015 to December 10, 2015. The crawl strategy is designed as followings: randomly select 100 users and then collect their tweet lists, the content of each tweet and following relationships among them. Finally, the dataset contains 4,913 users, 275,820 messages and 570,314 retweetings. The basic statistical information is shown in Table 1.

Table 1. Statistics of experimental dataset.

Dataset	#Users	#Messages	#Relations	#Retweetings	Sparseness
Weibo	71,649	100,000	1,125,365	7,198,730	0.1%
Twitter	4,913	275,820	1,075,820	570,314	0.04%

We evaluate the quality of the approximate values for retweeting matrix R using the Root Mean Square Error (RMSE). We can see that a smaller RMSE value means a better performance. Due to the nature of the problem, both the observation and prediction are binary. Hence, we also use the Precision, Recall, F_1 and Accuracy to evaluate the performance of the proposed algorithms. We randomly split each dataset into two disjoint sets, 80% for training and 20% for testing, and perform 5-fold cross validation.

5.2 Baseline Methods

We compare our models against the following traditional and state-of-the-art models:

- **PMF.** This method assume that retweetings are from the normal distributions given the factor vectors of users and messages. The user's retweeting behaviors can be predicted by the inner products of user and message factor vectors based on matrix completion in missing data [20].
- **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 [32].
- **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 [26].
- **SUA-ACNN.** The model proposes a novel attention-based deep neural network to incorporate user's attention interests and social information for this task by embedding to represent the user, the user's attention interests, the author and message respectively [34].
- **HCFMF.** This method learns message embedding by jointly taking the message co-occurrence, semantics, social patterns into consideration, then decomposes the user-message matrix and message-message similarity matrix based on a co-factorization model for learning user's retweeting behavior [25].

We also implement the different configurations of our proposed model to verify the effectiveness of our algorithm.

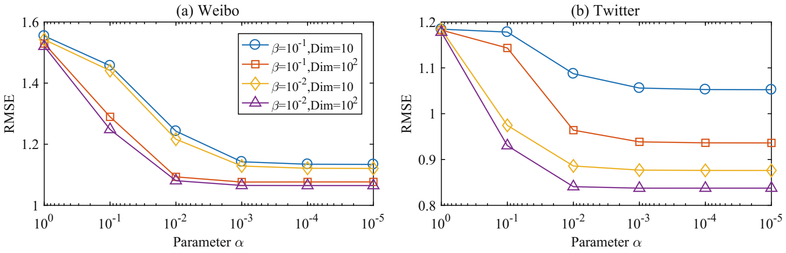


Fig. 3. Impact of α with CBRMF model on two datasets.

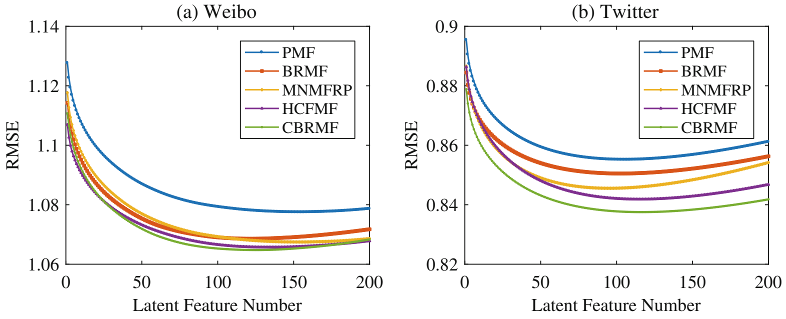


Fig. 4. RMSE vs. latent feature number on two datasets.

- **CBRMF-U.** This method only employs user community factor by eliminating the effect of document clustering regularization term with setting $\beta = 0$ in Eq. (13).
- **CBRMF-M.** This method only uses document clustering factor by eliminating the effect of user community regularization term with setting $\alpha = 0$ in Eq. (13).

5.3 Experimental Results

This section presents some important results settings and also give the results in more details by comparing these baseline methods and discuss the impact of different factors.

Parameter Settings. The parameters α and β provide important contribution strengths for social contextual information in our CBRMF model. The impact of β generally shares the same trend as α. Hence, we here only illustrate the results of α due to the space limitation. The experimental results of α on Weibo and Twitter datasets are shown in Fig. 3. From the figure, we can observe that the RMSE values gradually decrease while parameter β and dimension of latent features increasing on two datasets, and α become stable around 10^{-3}. Hence, we set $\alpha = 10^{-3}$ in our experimental setup. Similarly, we also empirically set the parameters $\beta = 10^{-3}$ and $\eta = 10^{-2}$ on two datasets in our models.

Number of Latent Features. We perform PMF, BRMF, MNMFRP, HCFMF, and CBRMF models to discover the proper number of latent features on Weibo and Twitter datasets, respectively. The conducted experimental results are shown in Fig. 4 with number of latent features K from 2 to 200. From these results, we can find that with the latent feature number K increasing, the RMSE first decreases and then increases, and reach the lowest point around 100. Considering the calculation effect and time efficiency, we choose $K = 100$ as the dimension of latent feature space.

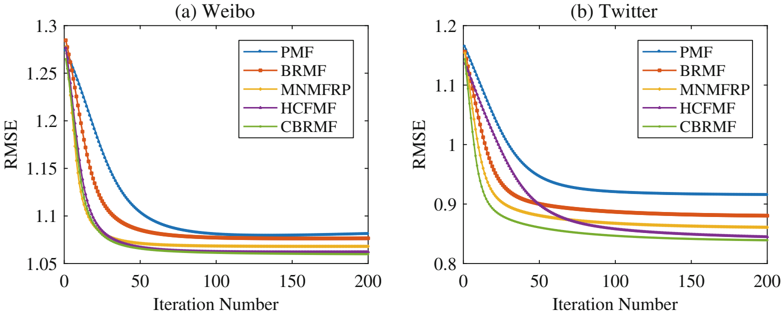


Fig. 5. RMSE vs. Iteration Number on two datasets.

Number of Iterations. Similarly, to find the proper number of updating iterations to get a good performance while avoid overfitting, we record the RMSE values for each iteration. Figure 5 illustrates the impacts of the number of iterations on two datasets. From the results, we can see that RMSE values on two datasets decrease gradually with the number of iterations increasing. To reach a converged result with an acceptable computational cost, we set number of iterations to 100 in our proposed models.

Performance Comparison for Retweeting Prediction. We demonstrate the prediction performance of our proposed methods and all baseline methods to find who will retweet. Specifically, we run all methods for 5 runs, and report the average results of each method in Table 2. From these results, we can observe the following conclusions: (1) our proposed CBRMF, which incorporates user community and document clustering together, significantly outperforms the baseline methods in our experimental results; (2) the proposed BRMF outperforms PMF, which indicates it is reasonable that retweetings are from binomial distributions instead of normal distributions given the factor vectors of users and messages; (3) the comparison between CBRMF-U vs. CBRMF and CBRMF-M vs. CBRMF, reveals that the user factors and message factors are comparable results compared with the social factors for retweeting prediction. (4) most of matrix factorization methods such as MNMFRP, HCFMF and CBRMF for retweeting prediction can achieve the accuracy of prediction relatively well. These results suggest that matrix factorization is suitable for the task. (5) SUA-ACNN performs slightly better than most of baseline methods, and also the improvements are statistically significant compared to BRMF. The results demonstrate that the attention-based deep neural network can benefit the performance. We also notice slightly different performance of the two datasets. For example, the BRMF and CBRMF methods achieve better prediction results on Weibo dataset than on Twitter dataset. A possible reason is the average number of retweetings per user on Weibo dataset is much higher than the number on Twitter dataset. In this case, user-based method can generally generate better results since every user has more information to use. This is the possible cause of why the user-based method has better performance. In summary, we conclude that BRMF and CBRMF following binomial distributions are a reasonable assumption, and improve the prediction accuracy by using social contextual information.

Table 2. Performance of retweeting prediction with different baseline methods on Weibo and Twitter datasets.

Method	Weibo dataset				Twitter dataset			
	Precision	Recall	F_1	Accuracy	Precision	Recall	F_1	Accuracy
PMF	0.628	0.607	0.612	0.619	0.611	0.654	0.631	0.621
BRMF	**0.669**	**0.645**	**0.657**	**0.668**	**0.657**	**0.612**	**0.635**	**0.643**
LRC-BQ	0.698	0.770	0.733	0.719	0.669	0.638	0.653	0.656
MNMFRP	0.754	0.705	0.729	0.757	0.711	0.688	0.699	0.693
SUA-ACNN	0.746	0.733	0.739	0.753	0.728	0.713	0.720	0.725
HCFMF	0.756	0.742	0.749	0.763	0.739	0.725	0.732	0.735
CBRMF-U	0.723	0.702	0.712	0.723	0.727	0.712	0.719	0.743
CBRMF-M	0.733	0.716	0.724	0.738	0.762	0.751	0.757	0.778
CBRMF	**0.802**	**0.785**	**0.794**	**0.797**	**0.795**	**0.774**	**0.784**	**0.785**

Impact of the Number of Clusters. The number of clusters with user and message has a great effect on the performance of our proposed CBRMF model. Figure 6 shows the accuracy of retweeting prediction with various values of user and document clusters found on two datasets. From the result, we can see that (1) these datasets have a optimal value for number of clusters when we enlarge the number of cluster; (2) The best number of clusters found by CBRMF is near the latent factor number of groups which means CBRMF can infer the number of clusters automatically when the number of cluster is large enough. Therefore, we can conclude that it is a better practice to set user communities and document clusters to 40 and 80 on Weibo dataset, and 40 and 70 on Twitter dataset, respectively.

Performance Comparison with Different Community Detection and Short Text Clustering Methods. We investigate the effects of different

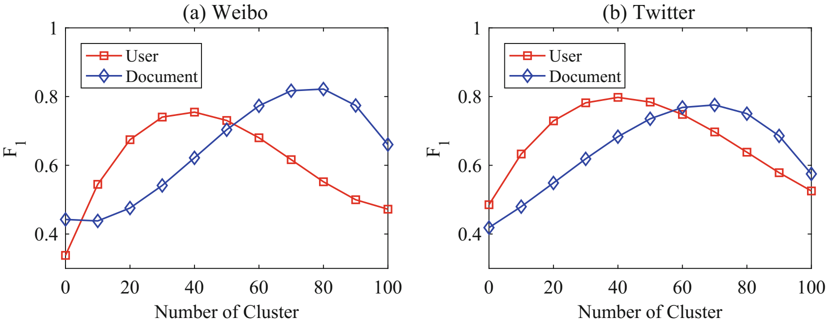


Fig. 6. F_1 vs. different number of cluster on two datasets.

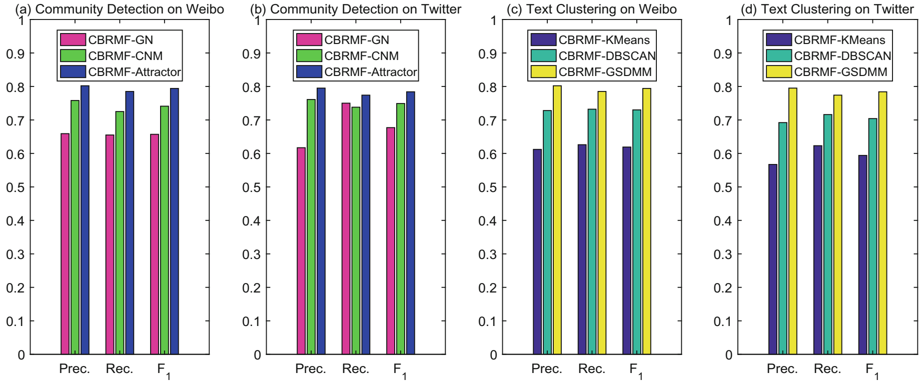


Fig. 7. Performance comparison of CBRMF with different community detection and short text clustering methods on datasets.

methods of community detection and short text clustering in our proposed model. The detailed description is as

- **CBRMF-GN.** The method utilizes the Girvan-Newman algorithm to detect user communities.
- **CBRMF-CNM.** This method uses the Clauset-Newman-Moore community detection method for large networks.
- **CBRMF-KMeans.** This method employs k-means clustering to partition short text documents into clusters.
- **CBRMF-DBSCAN.** This method uses density-based spatial clustering of applications with noise to group short texts documents into clusters.

Figure 7 shows the prediction performance with different community detection and short text clustering methods on Weibo and Twitter datasets. From these results, we can see when choosing Attractor algorithm to perform user's community detection, CBRMF-Attractor outperforms CBRMF-GN and

CBRMF-CNM on two datasets. We have the similar observation CBRMF-GSDMM performs better than CBRMF-KMeans and CBRMF-DBSCAN over short text clustering while the GSDMM is done. In a word, these results suggest the Attractor and GSDMM models are a better choice for detecting user community and clustering short text documents on social network datasets.

6 Conclusion

We propose two novel retweet prediction models based on binomial distributions and contextual information. The proposed two models assume retweetings are from binomial distributions instead of normal distributions under matrix factorization. CBRMF is an extended BRMF model by incorporating user community and document clustering as social regularization terms to alleviate the data sparsity and reduce the noisy information. By experimental evaluation using two real-world social network datasets, we can conclude it is reasonable to assume that retweetings are from binomial distributions, and our proposed methods outperform existing state-of-the-art comparison methods.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (No. 61702508, No. 61802404), the National Key Research and Development Program of China (No. 2016QY06X1204), and the Strategic Priority Research Program of the CAS (XDC02000000). 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. Bae, Y., Ryu, P.-M., Kim, H.: Predicting the lifespan and retweet times of tweets based on multiple feature analysis. J. Electron. Telecommun. Res. Inst. **36**(3), 418–428 (2014)
3. Bedi, P., Sharma, C.: Community detection in social networks. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. **6**(3), 115–135 (2016)
4. Can, E.F., Oktay, H., Manmatha, R.: Predicting retweet count using visual cues. In: CIKM, pp. 1481–1484. ACM (2013)
5. Gao, H., Tang, J., Hu, X., Liu, H.: Content-aware point of interest recommendation on location-based social networks. In: AAAI, pp. 1721–1727 (2015)
6. Jamali, M., Ester, M.: A matrix factorization technique with trust propagation for recommendation in social networks. In: RecSys, pp. 135–142 (2010)
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.) APWeb 2015. LNCS, vol. 9313, pp. 756–767. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25255-1_62
8. Jiang, M., et al.: Social contextual recommendation. In: CIKM, pp. 45–54 (2012)

9. Lee, K., Mahmud, J., Chen, J., Zhou, M., Nichols, J.: Who will retweet this? Detecting strangers from Twitter to retweet information. ACM Trans. Intell. Syst. Technol. (TIST) **6**(3), 31 (2015)
10. Li, C., Bendersky, M., Garg, V., Ravi, S.: Related event discovery. In: WSDM, pp. 355–364. ACM (2017)
11. Liu, L., Tang, J., Han, J., Jiang, M., Yang, S.: Mining topic-level influence in heterogeneous networks. In: CIKM, pp. 199–208. ACM (2010)
12. Luo, Z., Osborne, M., Tang, J., Wang, T.: Who will retweet me? Finding retweeters in Twitter. In: SIGIR, pp. 869–872. ACM (2013)
13. Ma, H., Yang, H., Lyu, M.R., King, I.: SoRec: social recommendation using probabilistic matrix factorization. Comput. Intell. **28**(3), 289–328 (2008)
14. Ma, H., King, I., Lyu, M.R.: Learning to recommend with social trust ensemble. In: SIGIR, pp. 203–210 (2009)
15. Ma, H., Zhou, D., Liu, C., Lyu, M.R., King, I.: Recommender systems with social regularization. In: WSDM, pp. 287–296. ACM (2011)
16. Macedo, A.Q., Marinho, L.B., Santos, R.L.T.: Context-aware event recommendation in event-based social networks. In: RecSys, pp. 123–130. ACM (2015)
17. Mahdavi, M., Asadpour, M., Ghavami, S.M.: A comprehensive analysis of tweet content and its impact on popularity. In: IST, pp. 559–564. IEEE (2016)
18. 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)
19. Peng, H.-K., Zhu, J., Piao, D., Yan, R., Zhang, Y.: Retweet modeling using conditional random fields. In: 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW), pp. 336–343. IEEE (2011)
20. Salakhutdinov, R., Mnih, A.: Probabilistic matrix factorization. In: NIPS, pp. 1257–1264 (2007)
21. Shao, J., Han, Z., Yang, Q., Zhou, T.: Community detection based on distance dynamics. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2015, New York, NY, USA, pp. 1075–1084. ACM (2015)
22. Shi, J., Chen, G., Lai, K.K.: Factors dominating individuals' retweeting decisions. In: CyberC, pp. 161–168. IEEE (2016)
23. Suh, B., Hong, L., Pirolli, P., Chi, Ed.H.: Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In: SocialCom (2010)
24. Tang, J., Gao, H., Liu, H.: mTrust: discerning multi-faceted trust in a connected world, pp. 93–102 (2012)
25. 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)
26. Wang, M., Zuo, W., Wang, Y.: A multidimensional nonnegative matrix factorization model for retweeting behavior prediction. Math. Probl. Eng. **2015**, 10 (2015)
27. Wang, X., Lu, W., Ester, M., Wang, C., Chen, C.: Social recommendation with strong and weak ties. In: CIKM, pp. 5–14. ACM (2016)
28. Xie, J., Szymanski, B.K.: Towards linear time overlapping community detection in social networks. In: Tan, P.-N., Chawla, S., Ho, C.K., Bailey, J. (eds.) PAKDD 2012. LNCS (LNAI), vol. 7302, pp. 25–36. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-30220-6_3
29. Yang, C., Liu, L., Jiao, Y., Chen, L., Niu, B.: Research on the factors affecting users' reposts in microblog. In: ICSSSM, pp. 1–6. IEEE (2017)

30. Yang, Z., et al.: Understanding retweeting behaviors in social networks. In: CIKM, pp. 1633–1636 (2010)
31. Yin, J., Wang, J.: A Dirichlet multinomial mixture model-based approach for short text clustering. In: KDD, pp. 233–242. ACM (2014)
32. Zhang, J., Tang, J., Li, J., Liu, Y., Xing, C.: Who influenced you? Predicting retweet via social influence locality. ACM Trans. Knowl. Discov. Data (TKDD) **9**(3), 25 (2015)
33. Zhang, Q., Gong, Y., Guo, Y., Huang, X.: Retweet behavior prediction using hierarchical Dirichlet process. In: AAAI, pp. 403–409 (2015)
34. Zhang, Q., Gong, Y., Wu, J., Huang, H., Huang, X.: Retweet prediction with attention-based deep neural network. In: CIKM, pp. 75–84. ACM (2016)