Predicting Individual Retweet Behavior by User Similarity: A Multi-Task Learning Approach

Xing Tang[a], Qiguang Miao[a,*], Yining Quan[a], Jie Tang[b], Kai Deng[a]

[a] *School of Computer Science and Technology, Xidian University, China*
[b] *Department of Computer Science and Technology, Tsinghua University, China*

Abstract

Users read microblogs and retweet the most "interesting" tweets to their friends in online social networks. Predicting retweet behavior is extremely challenging due to various reasons. First, the most of existing approaches primarily discuss a global retweet predicting model, with a goal of finding a uniform model that fits all users, but ignore individual behavior. And while social influence plays an important role in information diffusion, this fact has been largely ignored in conventional research. In this paper, we adopt a "microeconomics" approach to a model, and predict the individual retweet behavior. We study relationships between users by considering social similarity, which reflects how a particular retweeting action affects both the originator and the receiver of the retweet. To address the individual and social challenges, we analyze the effect of social similarity on retweet behavior based on a real dataset. Moreover, we cast our predicting problem as a multi-task learning problem. Combining the social and individual understanding, we then propose a novel model for predicting individual retweet behavior. We conduct extensive experiments on a Weibo [1] dataset to validate the effectiveness of the proposed model. Our results demonstrate the superior performance of the proposed model, compared with several alternative classification methods.

Keywords: Online social network, Retweet prediction, Multi-task learning,

[*] Corresponding author
Email address: qgmiao@126.com (Qiguang Miao)

[1] http://weibo.com, the largest microblogging service in China

User similarity, Microblog service, Individual behavior

1. Introduction

The rapid development of online social networks and media, such as Facebook, Twitter and Weibo, have attracted much research attention. In microblogging services, a user shares her/his latest status and an opinion in the form of short message (called a "tweet" in Twitter and Weibo), with a limit of 140 or fewer characters. Her/his followers (those users who are following her/him) can read this message and choose to forward (retweet) it, to allow their followers to be aware of it.(We use the "tweet" and "retweet" terminology for all social media in this paper, for the sake of simplicity, although, for example, Facebook messages are not usually referred to as "tweets".)

Tweets are a rich source of communication messages between users. Their popularity has resulted in information propagation becoming a prominent fundamental function of online social networks. From the perspective of information propagation, "retweeting" is viewed as an atomic behavior [1, 2]. Specifically, retweeting action diffuses information carried in the original message (tweet) to one's followers and again to more followers if more users choose to retweet a retweeted message. Prior information propagation models treats the retweeting behavior as having constant retweet probability, or as following a certain probability distribution. Moreover, statistics show that 1% of Twitter users produce 50% of its content [3] and control 25% of its information diffusion [4]. Therefore, individual retweet behavior is the key to model information propagation in such online social networks.

Moreover, retweeting is regarded as an indicator whether a user is interested in a particular tweet. In practice, a user receives tweets from his followees; they are listed in chronological order, regardless of their potential to interest the user. The individual retweet-predicting behavior model provides a method for users to find interesting tweets. In addition, tweets from followees who have a close relationship with the user are more likely to be retweeted, which is the role of social interaction.

Several existing methods tried to model the retweet behavior as a binary classification problem. They focus on global models, i.e., predicting popular tweets in the whole social network [5] or from the perspective of information spreading [6, 7]. Such these methods lead to the phenomenon of homogeneity. However, individual retweet behavior is hard to capture because of the difficulty in identifying individual retweeting behaviors. In fact, many users either do not retweet at all, or else all their posted tweets are totally retweeted from others. Therefore, when considering individual behavior, both the global and local views are helpful in overcoming the data sparsity and exploring the effects of individuality on retweeting. It has been suggested that a user's behavior is affected greatly by his friends who have the same tastes as he does [8]. As a result, in order to model the individual retweet behavior accurately, it is necessary to take the user similarity into consideration.

In this paper, we adopt a microeconomics approach to study individual retweet behavior, and aim to discuss the effects of social user similarity on the retweet behavior. User similarity is defined as a measurement combining user structural similarity and profile similarity [9]. In particular, the relation between user similarity and some social features has been explored. Based on the data analysis, we further give a user similarity measure based on features extracted from a real dataset. Then our proposed solution starts with a simple predicting model based on logistic regression. We then extend our model casting the individual prediction problem as a multi-task learning problem, in which each task corresponds to a user-predicting task. In a sum, we treat every user's retweeting prediction problem as a task, and combine all tasks with social similarity in our model, which leads to an individual and similarity-related predicting model. Specifically, we decompose the model into two parts: the common part, which is for global optimization over all users' retweet history, and the individual part, which is based on specific user's retweet history. As discussed above, we also further deploy a user-similarity regularization term for smooth predictions, to model the effects of social similarity. We apply our model to a Weibo dataset. We conduct extensive experiments to compare our approach with other binary classification algorithms, and analyze the parameter sensitivity of the proposed model. The results show that our model is useful and effective.

The rest of this paper is organized as follows. Section 2 introduces related work. Section 3 presents a data analysis of user similarity and extracted social features. We formalize the problem and describe the details of our model in Section 4. After that, we describe the dataset and experiments. Based on the experiments, we offer some analysis in Section 5. Our conclusion is in Section 6.

2. Related Work

Research on the retweet behavior is multifaceted. With the extracted features as a starting point, Yang et al. [10] analyzed the effects of different factors on the retweet behavior. Based on their statistical observations, they proposed a factor graph method to solve the prediction problem by emphasizing the impact of factor on retweet behavior. Most of the work tries to solve the information propagation issues by considering the retweet behavior from different aspects. Pezzoni et al.[11] discussed how structure factors and retweet behavior affect information diffusion. Considering the retweet behavior as atomic behavior, they proposed an agent-based information propagation model to generate a cascade. Petrovic et al. [12] also proposed retweet behavior prediction as the key to understanding information propagation. They verified that prediction was indeed possible by conducting human experiments, and proposed a Passive-Aggressive algorithm as the basis of a global predicting model. Yang et al. [13] proposed an influence model in implicit networks. They treated each node's retweet behavior as the capability to propagate information. Due to different properties of different datasets, many works try to determine the key factors for the retweet behavior [14, 15]. Liu et al.[15] presented detailed heuristic cues and systematic cues, which determines the retweet behavior in microblogging under emergency conditions. Spiro et al.[16] discussed the effects of waiting time on the retweet behavior, using statistical methods to understand it. Unlike our model, these works only studied useful factors that predict whether a tweet will be retweeted, regardless of who will do it. To the best of our knowledge, the individual retweeting behavior question has not yet been discussed.

Some studies use different approaches to predict retweet behavior. Fei et al.[17] were first to adopt microeconomics methods for social media behavior prediction. They mainly discussed how information contents and user interests are related to social media behaviors, employing a multi-task learning method to study the matter. Zhang et al. [18, 19] discussed the effect of social influence locality on the retweet behavior. Feng et al.[20] first proposed tweet re-ranking as an approach. With their recommendation method, they introduced a matrix factorization based on the features to solve the individual problem, which had the same goal as our work. However, they did not make use of the user similarity to model shared user preferences.

Aiming to utilize previously-acquired knowledge to solve new but similar problems, transfer learning has a wide range of real-world applications by using computational intelligence [21]. As one type of transfer learning, multi-task learning has been also widely investigated by different researchers. Multi-task learning [22, 23, 24] assumes all the tasks under consideration are uniformly related, and aims to learn a common low-dimensional representation without actually learning the task relationships. A similar individual behavior model has been used in predicting user activity level[25]. Li et al. [26] proposed a collaborative online multitask learning method and showed an application to microblog sentiment detection.

Although previous researches have applied many models to retweet behavior prediction, our work is very different, and its main contributions are summarized as follows. First, we investigate the relation between extracted features and user similarity. Based on the data analysis, we propose a method for evaluating user similarity. Second, our work proposes a multi-task learning model that encodes individual behavior and user similarity into a unified framework for predicting individual retweet behavior. Last, our dataset is extracted from Weibo in China, and we validate our model on the dataset, comparing it with existing models.

3. A User-Similarity Measure

3.1. Data analysis

Our discussions here have been mainly designed to measure the impact of user similarity on the retweet behavior.

We firstly detail several observations as follows. In Weibo platform, suppose user i is a follower of user j, then user j's update tweets will notify user i in the form of timelines. Notice that any update tweets should be either an original tweet or a retweet message from one's follows. As a result, user i will be aware of user j's retweet behavior. Based on the observations, it suggests that follows' retweet action has directed impact on one's retweet behavior, we therefore restricted our focus on the directed relationships between follow pair. Furthermore, we give the definition of retweet behavior.

DEFINITION 1. **Retweet Behavior** We represent follow relationship between i and j as follow pair $i \to j$. And we use the triple (j, t, m) to denote that user j retweets a microblog m at time t. For every triple in j's retweet list $list\_j$, it may be inserted into i's retweet list $list\_i$ according to the i's retweet behavior. Furthermore, $y_{i,m}$ indicates the retweet behavior of user i for the given microblog m.

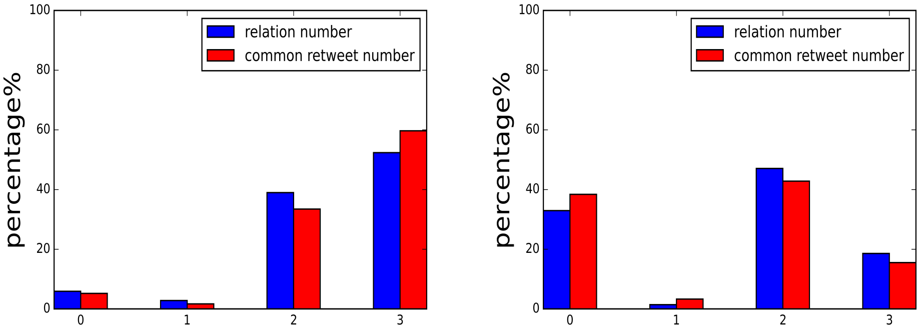
Without loss of generality, for a particular microblog m, we consider the binary action, i.e, $y_{i,m} \in \{0, 1\}$, where $y_{i,m} = 1$ means user i retweet the microblog and $y_{i,m} = 0$ indicates user i ignore it and did not take retweet action. From the definition, it is clear that user can be aware of his follows' retweet behaviors and decide which tweet to retweet. Based on above discussions, we consider the directed similarity between follow relation pair $i \to j$. From a real dataset, we have extracted three kinds of related features, which are respectively structural features, profile features, and tweet features. In our dataset (5.1), the number of tweets retweeted by both i and j , i.e. $|list\_i \cap list\_j|$, is used as an indicator of how much two users share the same retweet behavior affinity.

First, structural features. In our dataset, the "mutual follow" feature belongs to this category. "Mutual follow" refers to a reciprocal relationship between two users. Figure2a shows the average number of common retweets per user, where dissimilarity represents a unidirectional relationship, and similarity represents a reciprocal relationship. Presumably, a "mutual follow" will make two users retweet more of the same tweets than users who are not in a "mutual follow" relationship.

As to user profiles, we extract name, gender, location, birthday, and experience. However, we find that a large number of users are unwilling to fill out birthday and experience fields, because of privacy concerns. There are high "missing value" rates in these features for our active users dataset(5.1), 37% and 43% respectively. To investigate the impact of missing features, we have moreover tried to fill the missing value with a null value and conduct the similarity analysis in our network. Specially, when both of the pair have the same birthday and experience, they are considered as similarity. While if either of follow pair has null value, we consider it as dissimilarity. Then, the results is shown in Figure1.

It should be noticed that the correlations between the both two features and retweet similarity are weak because of the great amount of missing values. With respect to gender and location, there are no missing value for both two features. As a result, we use gender and location to measure similarity. Notice that both the two features take binary values. For gender feature, similarity means two users have the same gender, while dissimilarity refers to reverse gender. With respect to location feature, we define the value as similarity when both of users are in the same city and as the dissimilarity when they are in the different cities. Figure2b and Figure2c show how these features affect retweet similarity.

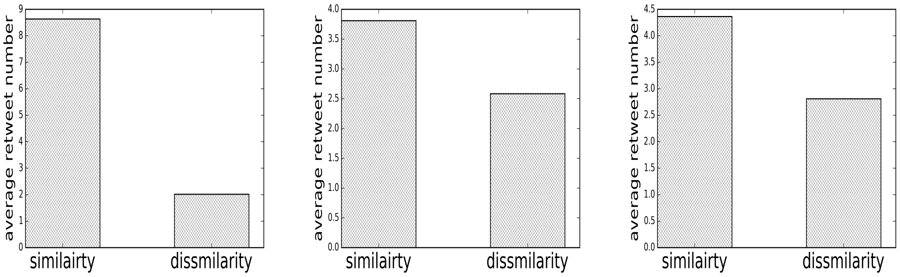
Last, the tweet feature is considered. Because the tweet is usually a short text block–no more than 140 characters–we combine all the tweets posted by a particular user into a single large text container. Then we apply a Latent



(a) birthday feature

(b) experience feature

Figure 1: The statistics of birthday and experience.(0:dissimilarity for real value;1:similarity for real value;2:dissimilarity for real value and miss value;3:dissimilarity for miss value)



(a) mutual follow

(b) gender similarity

(c) location similarity

Figure 2: The effects of structural and profile similarity on retweet number

Dirichlet Allocation topic model [27] to the text containers. After obtaining each user's topic distribution vector $Topic\_i$ and $Topic\_j$, cosine similarity is used to calculate the tweet similarity between users.

$$TopicSim(i,j) = \frac{Topic\_i \bullet Topic\_j}{\|Topic\_i\| \, \|Topic\_j\|} \tag{1}$$

The trend is plotted in Figure3. We can see that the common retweet number increases with the topic similarity. Dispite of the outliers,the slope is nearly at 0.5 with regression.

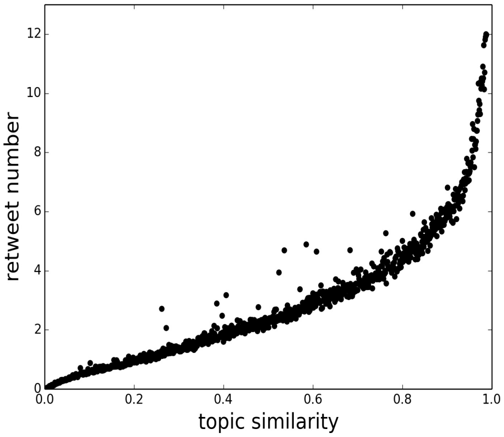
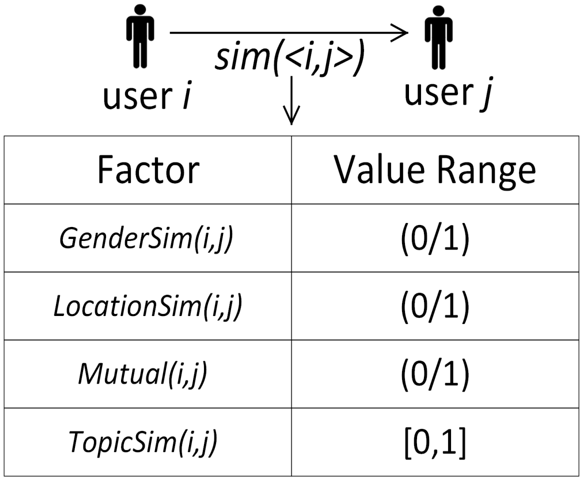


Figure 3: Tweet similarity and retweet number

3.2. User similarity

From the data analysis, we observe that the three kinds of feature affect user retweet behavior similarity to varying degrees. To define user similarity, we combine all features by addition. Figure4 illustrates directed user similarity between two users.



Factor	Value Range
GenderSim(i,j)	(0/1)
LocationSim(i,j)	(0/1)
Mutual(i,j)	(0/1)
TopicSim(i,j)	[0,1]

Figure 4: Directed similarity representation

In Figure4, $GenderSim(i,j)$ is the gender similarity between i and j,*Location*$-$ $Sim(i,j)$ represents the location similarity between them, and $Mutual(i,j)$ indicates whether they are mutual friends. All three features are binary attributes. In addition, $TopicSim(i,j)$ is the application of cosine similarity, to compare the tweets of i and j in a given topic space. Then, we collect users' specific features, and quantify their similarity from all the aspects discussed. More formally, for each user, we define pair-wise user similarity as:

DEFINITION 2. Suppose user i and j have the follow relation $i \rightarrow j$. Then similarity along the direction of the follow is defined as the weighted sum according to the different similarity factors:

$$\begin{aligned} sim(<i,j>) \quad &= \quad \alpha_1 GenderSim(i,j) + \alpha_2 LocationSim(i,j) \\ &\quad + \alpha_3 Mutual(i,j) + \alpha_4 TopicSim(i,j) \end{aligned} \tag{2}$$

where $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1$ and $\alpha_1, \alpha_2, \alpha_3, \alpha_4 \in R^+$. The selection of these four weights is based on the data analysis and ensures the similarity value is in the range $[0,1]$.

4. Individual Logistic Regression with User Similarity

As stated in the previous section, we formulate retweet behavior prediction as a binary classification problem in the usual way. In this section, we describe an individual and social interaction model based on logistic regression.

Generally, logistic regression builds a linear function on input features, and predicts target labels with a sigmoid function as follows:

$$y_i = \sigma(\mathbf{w^T x_i}) = \frac{1}{1 + \exp(-\mathbf{w^T x_i})} \tag{3}$$

where $\mathbf{x_i}$ is a data instance, y_i is corresponding prediction, and $\mathbf{w}$ is the coefficient vector that is to be learned from the data. Moreover, the prediction problem can be formulated as learning an optimal solution $\mathbf{w}$ by solving the minimization problem on the basis of the global retweet data:

$$\min_{\mathbf{w}} F = \min_{\mathbf{w}} (\sum_{i=1}^{N} \ell(y_i, \mathbf{w^T x_i}) + \gamma_0 \left\| \mathbf{w} \right\|_2^2) \tag{4}$$

where N is the entire number of retweet history and γ_0 is a parameter of the regularization term $\left\| \mathbf{w} \right\|_2^2$. The loss function $\ell(y_i, \mathbf{w^T x_i})$ is defined as

$$\ell(y_i, \mathbf{w^T x_i}) = -y_i \log(\sigma(\mathbf{w^T x_i})) - (1 - y_i) \log(1 - \sigma(\mathbf{w^T x_i})) \tag{5}$$

An important advantage of logistic regression, used here as the base classifier, is that the output is probability, which is very useful in the re-ranking problem as stated previously. Furthermore, previous work [14] has shown that logistic regression can apply to retweeting prediction. However, this base model fails to model the individual and social interaction factors in the real world. Therefore, in the following, we show how to extend this model to capture these two important factors.

4.1. Individual factors

As mentioned above, accounting for individual behavior requires that for different users, the prediction model regarding individual retweet behavior should depend on the individual under consideration. A global model, such as **SVM**, will fail to make precise predictions on individual behavior. One way to address this challenge is create and apply numerous models to the different users, to learn their $\mathbf{w_i}$ based on their individual retweet data. However, there are some users who never retweet others, and other users who always retweet others, over the period of the dataset. Moreover, some personal retweet data is very sparse and cannot model behavior accurately. Therefore, inspired by multi-task learning, which aims to learn a set of different but related tasks jointly by exploring the commonality relativeness across tasks, we build an individual model for an individual user as a task.

The proposed model has two parts, common parts shared by multiple tasks and specific part for individual user tasks. In particular, we assume that all individual prediction task $\mathbf{V_i}$ can be written as:

$$\mathbf{V_i} = \mathbf{w_g} + \mathbf{w_i} \tag{6}$$

where the components $\mathbf{w_i}$ are learned based on the local historical retweet data.

Therefore, we presents our individual model as follows,

$$\begin{aligned} \min_{\mathbf{w_g}, \{\mathbf{w_i}\}_{i=1}^{M}} F = \min_{\mathbf{w_g}, \{\mathbf{w_i}\}_{i=1}^{M}} (\sum_{i} \sum_{j=1}^{N} \ell(y_{ij}, (\mathbf{w_g} + \mathbf{w_i})^T \mathbf{x}_{ij}) \\ + \gamma_0 \left\| \mathbf{w_g} \right\|_2^2 + \gamma_1 \sum_{i=1}^{M} \left\| \mathbf{w_i} \right\|_2^2) \end{aligned} \tag{7}$$

where, M is the number of users, $\{\mathbf{w_i}\}$ is the individual part, and γ_1 is a parameter for indvidual users. $\mathbf{x}_{ij}$ is the $j - th$ data instance that belongs to user i tweet history, and y_i is the corresponding target value. In this model, we can set γ_1 small to allow more individual consideration, although if we do so, it will suffer from the sparse data.

4.2. Social interaction

The most notable factor in social networks is the social interaction. In social networks, users tend to retweet tweets that are retweeted by his followers who share similar tastes. To model retweeting influence further, we take social interaction into consideration. Formally, based on the user similarity, we introduce a regularization term, to smooth the prediction and make his prediction similar to that of his friends who share more similar tastes.

$$\beta \sum_{i=1}^{M} \sum_{j \in F(i)} sim(< i, j >) \bullet (\mathbf{w_i}^T \mathbf{X} - \mathbf{w_j}^T \mathbf{X})^2 \tag{8}$$

where $sim(< i, j >)$ is the similarity measure between i and j according to definition 2. $F(i)$ shows the followers of user i , and β is a parameter to adjust how much similarity to apply. We take $\mathbf{w_i}^T \mathbf{X}$ as $\sum_{k}^{N} \mathbf{w_i}^T \mathbf{x}_k$ to keep the notation uncluttered, where N is the total number of tweet data instances that belong to both user i and user j. As the regularization term, social interaction plays a varying role in user's friends according to different similarity values. In particular, the greater value will lead to more similar retweet behavior, which is consistent with real-world data analysis.

4.3. Overall model

Combining all the factors discussed above, we get the complete model. With the simple logistic regression as the basis for this model, we model the retweeting prediction as a multi-task learning problem as follows:

$$\begin{aligned}\min_{\mathbf{w_g},\{\mathbf{w_i}\}_{i=1}^{M}} F = \min_{\mathbf{w_g},\{\mathbf{w_i}\}_{i=1}^{M}} & \{(\sum_{i}\sum_{j=1}^{N}\ell(y_{ij},(\mathbf{w_g}+\mathbf{w_i})^{T}\mathbf{x}_{ij}) \\ & + \gamma_0 \left\|\mathbf{w}_g\right\|_2^2 + \gamma_1 \sum_{i=1}^{M}\left\|\mathbf{w_i}\right\|_2^2 \\ & + \beta \sum_{i=1}\sum_{j\in F(i)} sim(<i,j>) \bullet (\mathbf{w_i}^{T}\mathbf{X} - \mathbf{w_j}^{T}\mathbf{X})^{2}\}\end{aligned} \tag{9}$$

We call this model Individual Retweet Behavior Logistic Regression with User Similarity (*IRBLRUS* for short). To learn this model, we propose to use gradient descent method, learned as follows:

$$\frac{\partial F}{\partial \mathbf{w}_g} = \sum_{i}^{M}\sum_{j=1}^{N}\mathbf{x}_{ij}(\sigma((\mathbf{w}_g+\mathbf{w}_i)^{T}\mathbf{x}_{ij}) - y_{ij}) + \gamma_0\mathbf{w}_0 \tag{10}$$

$$\begin{aligned}\frac{\partial F}{\partial \mathbf{w}_i} = & \sum_{j=1}^{N}\mathbf{x}_{ij}(\sigma((\mathbf{w}_g+\mathbf{w}_i)^{T}\mathbf{x}_{ij}) - y_{ij}) + \gamma_1\mathbf{w}_i \\ & + \beta\sum_{j\in F(i)} sim(<i,j>) \bullet \mathbf{X}(\mathbf{w}_i^{T}\mathbf{X} - \mathbf{w}_j^{T}\mathbf{X})\end{aligned} \tag{11}$$

Based on the above derivatives, we update $\mathbf{w}_g$ and $\mathbf{w}_i$ respectively by the following rules until the solution converges.

$$\mathbf{w}_g \leftarrow \mathbf{w}_g - \eta\frac{\partial F}{\partial \mathbf{w}_g} \tag{12}$$

$$\mathbf{w}_i \leftarrow \mathbf{w}_i - \eta \frac{\partial F}{\partial \mathbf{w}_i} \tag{13}$$

where η is the learning rate. The overall algorithm for *IRBLRUS* is listed in Algorithm 1.

Algorithm 1 Gradient Descent Algorithm for *IRBLRUS*

```
Input user features X, labeled data Y, similarity value $sim(< i, j >)$ , $i$ and $j$ are users, regularization parameters $\gamma_0$ , $\gamma_1$ and $\beta$ , learning rate $\eta$ and maximal number of iterations $I$
Output Global parameter $\mathbf{w}_g$ and personalized parameters $\mathbf{w}_i$ .
set $\mathbf{w}_g$ and $\mathbf{w}_i$ randomly
for $i = 1$ to $I$ do
    fix $\{\mathbf{w}_i\}_{i=1}^{M}$ , and keep updating $\mathbf{w}_g \leftarrow \mathbf{w}_g - \eta \frac{\partial F}{\partial \mathbf{w}_g}$
    for $n = 1$ to $M$ do
        fix other parameters, and keep updating $\mathbf{w}_i \leftarrow \mathbf{w}_i - \eta \frac{\partial F}{\partial \mathbf{w}_i}$
    end for
    if satisfy convergence condition then
        break
    end if
end for
return $\mathbf{w}_g$ and $\{\mathbf{w}_i\}_{i=1}^{M}$
```

5. Experiments

In this section, we compare our method with some baselines on the real dataset with Spark clusters[28]. The extensive experimental results demonstrate the effectiveness of our method. Moreover, we discuss the importance of social interaction in the individual model. Before that, we report how we design the real dataset and the features designed for retweet behavior prediction.

5.1. Dataset description

The dataset used for our experiments is collected from Weibo, the largest microblogging platform in China. We started by randomly choosing active seed users. In particular, we determined that an active user should satisfy the following attributes: (1) number of followers and followees > 50; (2) number of tweets per week > 10. As a result, we obtained 12,013 active users. Using these active users as seed users, we crawled a network with about 92,034 users, and extracted all the microblogs posted by them from June 1st to August 30th in 2013, which gave rise to 9,913,495 tweets. Moreover, we divide the tweets into retweets which are retweeted messages and original tweets which are non-retweeted messages according to the retweet behavior definition. In this data set, there are 716,178 retweets and 9,197,317 original tweets respectively. Moreover, we ensure that the original tweets used here are not retweeted in 30 days after the specified time window. The statistics of this dataset are shown in Table 1.

Table 1: Dataset Statistics

Users	Retweets	Original tweets	Relations
92,034	716,178	9,197,317	1,272,871

We split the dataset into a training set and a test set, with July 31, 2013 as the dividing date. Since the number of original tweets is much higher than the number of retweets (the unbalanced ratio is 12:1), we select a number of original tweets of only about the number of retweets in the set. Specially, we sample two random negative instances for each positive instances to ensure the sample ratio. The balanced method is aimed to alleviate the unbalanced data problem and ensure the performance of model [18].The final training set and test set statistics are given in Table 2.

5.2. Feature construction

Feature construction is the key to improving the classification accuracy. In this section, before conducting experiments to verify our proposed model, we

Table 2: Training set and Test Set Statistics

	Retweets	Original tweets
Training	588,002	1,168,838
Test	128,176	240,390

first introduce the features designed for the individual retweet behavior prediction task in microblogs. There are three categories of features in the vector consisting of the retweet history $\mathbf{x}_{ij}$, which are also summarized in Table 3.

- **Structure-based features:** These features reflect the structural characteristics of the tweet poster. In Weibo, there are two explicit relationships, which are “followers” and “followees”. Moreover, because of Weibo’s directed relations, we introduce the PageRank [29] here to evaluate the user’s structural importance.

- **User-based features:** This group of features is referred to as the user attributes. The number of original tweets and retweets represents a user’s activity. Users fall into two types: verified users, usually news media, celebrities etc.; and unverified users. Experience and hobby are the indicators of users’ involvement in the Weibo community.

- **Tweet-based features:** This group of features is referred to as tweet attributes. As to tweet content, the similarity between a tweet and its author’s interests can be obtained by LDA, and the length of a tweet is given after the removal of stop words. Whether the tweet contains the character “@” (which means a user is somehow involved in this tweet) and a URL (which means the tweet includes a picture or video) are set as two features. “@” emphasizes the social function of the tweet, while a URL emphasizes the tweet’s information and media content. Then, the retweet number (indicating the number of times it has been retweeted), the comment number (the number of comments it has accumulated), and

Table 3: Summary of Features

Category	Features of x_{ij} in $\mathbf{X_i}$
Structural features	Number of followers
	Number of followees
	Value of PageRank
User features	User is verified or not
	Number of original tweets
	Number of retweets
	Working experience
	Study experience
	Number of Hobbies
Tweet features	LDA similarity between tweet and author
	Length of Tweet
	If tweet contain URL
	If tweet @ someone
	Number of retweeting counts
	Number of comments
	Number of likes
	Number of topics
	Time of tweet

the "like" number (the number of "likes" it has collected) are the historical characteristics of this tweet. Publishing time and topic number have to do with tweet engagement throughout the social network.

5.3. Evaluation metric and models

We use F1-score, precision, and recall here to measure the performance. These measures are defined for binary classification in this paper:

$$recall = \frac{\#correctly\ classified\ as\ retweet}{\#true\ retweet} \tag{14}$$

$$precision = \frac{\#correctly\ classified\ as\ retweet}{\#classified\ as\ retweet} \tag{15}$$

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall} \tag{16}$$

Logistic Regression (**LR**) and Support Vector Machine (**SVM**) are used as binary classifiers here. These two classifiers, one of which is our base algorithm, are commonly used in predicting retweet behavior. Using these two classifiers, we can explore how retweet prediction can be tackled as a classification problem, and further verify how much improvement our proposed model can lead to.

Non-individual retweet prediction used in[12] is also compared here to prove the need for considering individual behavior. In this algorithm, the trained model is based on the whole dataset, which considers a tweet retweeted by anyone as a positive instance. Moreover, the recommendation method for finding the most interesting tweets [30] is also borrowed here.

We tune the parameters of all models by considering our real dataset. Specifically speaking, based on the data analysis in section 3.1, our similarity parameters here are $\alpha_1 = 0.3,\ \alpha_2 = \alpha_3 = 0.1,\ \alpha_4 = 0.5$.

5.4. Overall results

To verify the effectiveness of our model, the overall results are shown in Figure5 and Table 4. The comparative models are generally divided into individual models and non-individual models. We find that because of its appropriate kernel function, **SVM** may perform slightly better than linear model **LR** but still worse than our model. It can be concluded that the retweet prediction problem can be tackled by classification methods, no matter linear model or nonlinear model. Usually, nonlinear model performs better than linear model. However, our method **IRBLRUS** is better than **SVM**, which shows that introduction of individual behavior and social interaction considerably improve the accuracy of **LR**.

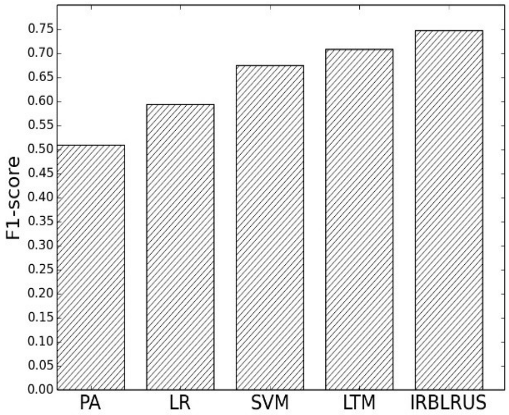


Figure 5: The overall F1-score results

The non-individual prediction applied here is an online learning **PA** algorithm, which is suitable for large amount data. Because of its global quality, **PA** performs worst in these five algorithms. A Latent Topic Model (**LTM**), for finding interesting tweets for a user, can also be used to find the retweets by selecting tweets that are likely to be retweeted. It performs better than **PA**. These two algorithms indicate the need for considering individual user behavior.

Finally, we explore the effectiveness of user similarity, by considering the reduction of our model: **IRBLR**, which only takes the individual behavior component into account. The result of the comparison is shown in Table4. The final result shows that **IRBLR** performs better than basic **LR** (by more than 0.062), while still below the **IRBLRUS**. On the other hand, **LTM** is the comparative method without user similarity, which is also below our proposed method. Both two confirm our inclusion of user similarity which explores the contribution of every individual tasks in the whole model.

Table 4: Comparison results

	PA	LR	SVM	LTM	IRBLR	IRBLRUS
Recall	0.384	0.461	0.548	0.584	0.521	0.620
Precision	0.753	0.832	0.876	0.898	0.882	0.951

5.5. Sensitivity analysis

In the following, we discuss how the model parameters affect the performance of **IRBLRUS**. Mainly, to study the effect of the social interaction weight on the final result, we set the β from 0.1 to 1.0, and show the precision/recall/F1-score trend in Figure6 . To ensure that only the β is affecting the outcome, we set the γ_0 and γ_1 at the optimum values with every β values. From the result, we see that a too-large β will hurt performance, while a value that is too small will fail to function in the model, and also hurt the performance. Meanwhile, we observe that recall value drops only slightly while precision climb greatly, which indicates that social interaction is able to support multiple tasks in predicting retweets precisely.

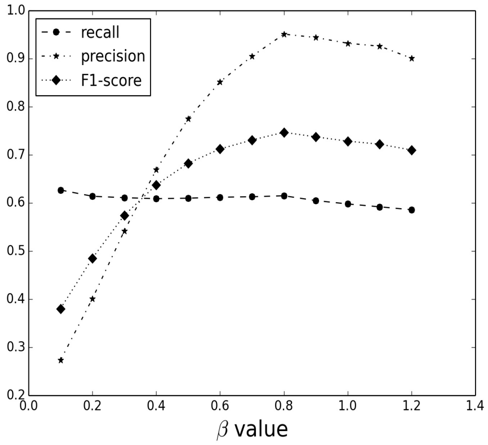


Figure 6: Change the socialization parameter β

We have also found how the imbalance weight in the training dataset will affect the final performance. We set the ratio between original tweets and retweets as b_{weight}. We iteratively change b_{weight} from 1 to 5, and show the trend in Figure7. We can find that although the F1-score and precision will climb slightly, recall is stunted with the increase of b_{weight} . However, the F1-score is always better than with basic **LR**, which is considered as acceptable.

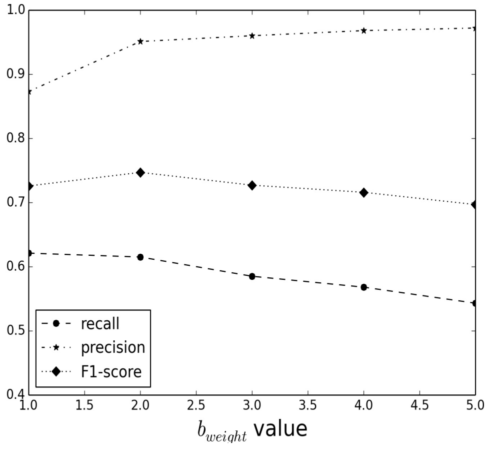


Figure 7: Change the class ratio b_{weight}

6. Conclusion and Future Work

In this paper, we proposed a novel individual user behavior model with user similarity, by incorporating multi-task learning. Based on a Weibo dataset, our extensive experiments demonstrate that the model is effective and necessary in predicting individual user retweet behavior, which is seldom discussed in other works.

In the future, we will explore new methods for defining the user similarity, and also try to introduce other different base algorithms to improve retweet prediction. Aiming to accommodate the streaming nature of tweeting, we will incorporate the online learning to adapt to the data in real time.

7. Acknowledgment

The work was jointly supported by the National Natural Science Foundations of China under grant No. 61472302,61272280 and U1404620; The Program for New Century Excellent Talents in University under grant No. NCET-12-0919; The Fundamental Research Funds for the Central Universities under grant No. K5051203020, K5051303018, JB150313and BDY081422,; Natural Science Foundation of Shaanxi Province, under grant No.2014JM8310.

References

[1] D. Boyd, S. Golder, G. Lotan, Tweet, tweet, retweet: Conversational aspects of retweeting on twitter, in: Proceedings of the 43rd Hawaii International Conference on System Sciences, 2010, pp. 1–10.

[2] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, W. Kellerer, Outtweeting the twitterers - predicting information cascades in microblogs, in: Proceedings of the 3rd conference on Online social networks, Berkeley, CA, USA, 2010.

[3] S. Wu, J. M. Hofman, W. A. Mason, D. J. Watts, Who says what to whom on twitter, in: Proceedings of the 20th international conference on World Wide Web, ACM, 2011, pp. 705–714.

[4] T. Lou, J. Tang, Mining structural hole spanners through information diffusion in social networks, in: Proceedings of the 22nd international conference on World Wide Web, 2013, pp. 825–836.

[5] L. Hong, O. Dan, B. D. Davison, Predicting popular messages in twitter, in: Proceedings of the 20th international conference companion on World wide web, ACM, New York, NY, USA, 2011, pp. 57–58.

[6] Q. Yan, L. Wu, L. Zheng, Social network based microblog user behavior analysis, Physica A: Statistical Mechanics and its Applications 392 (7)

(2013) 1712 – 1723. doi:http://dx.doi.org/10.1016/j.physa.2012.12.008.

[7] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, J. Leskovec, Can cascades be predicted?, in: Proceedings of the 23rd international conference on World wide web, Republic and Canton of Geneva, Switzerland, 2014, pp. 925–936.

[8] S. A. Macskassy, M. Michelson, Why do people retweet? antihomophily wins the day, in: Fifth International AAAI Conference on Weblogs and Social Media, AAAI Press, 2011, pp. 209–216.

[9] C. Akcora, B. Carminati, E. Ferrari, User similarities on social networks, Social Network Analysis and Mining 3 (3) (2013) 475–495. doi:10.1007/s13278-012-0090-8.

[10] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, Z. Su, Understanding retweeting behaviors in social networks, in: Proceedings of the 19th ACM international conference on Information and knowledge management, ACM, New York, NY, USA, 2010, pp. 1633–1636.

[11] F. Pezzoni, J. An, A. Passarella, J. Crowcroft, M. Conti, Why do i retweet it? an information propagation model for microblogs, in: Social Informatics, Vol. 8238, Springer, 2013, pp. 360–369.

[12] S. Petrovic, M. Osborne, V. Lavrenko, Rt to win! predicting message propagation in twitter, in: Fifth International AAAI Conference on Weblogs and Social Media, AAAI Press, 2011, pp. 586–589.

[13] J. Yang, J. Leskovec, Modeling information diffusion in implicit networks, in: Data Mining (ICDM), 2010 IEEE 10th International Conference on, 2010, pp. 599–608.

[14] Z. Xu, Q. Yang, Analyzing user retweet behavior on twitter, in: Proceedings of the 2012 International Conference on Advances in Social Networks

Analysis and Mining (ASONAM 2012), IEEE Computer Society, Washington, DC, USA, 2012, pp. 46–50.

[15] Z. Liu, L. Liu, H. Li, Determinants of information retweeting in microblogging, Internet Research 22 (4) (2012) 443–466.

[16] E. Spiro, C. Irvine, C. DuBois, C. Butts, Waiting for a retweet: modeling waiting times in information propagation, in: NIPS'12, NIPS workshop of social networks and social media conference, 2012.

[17] H. Fei, R. Jiang, Y. Yang, B. Luo, J. Huan, Content based social behavior prediction: A multi-task learning approach, in: Proceedings of the 20th ACM international conference on Information and knowledge management, ACM, New York, NY, USA, 2011, pp. 995–1000.

[18] J. Zhang, J. Tang, J. Li, Y. Liu, C. Xing, Who influence you? predicting retweet via social influence locality, ACM Transactions on Knowledge Discovery from Data 9 (3) (2015) 25:1–25:26. doi:10.1145/2700398.

[19] J. Zhang, B. Liu, J. Tang, T. Chen, J. Li, Social influence locality for modeling retweeting behaviors, in: Proceedings of the Twenty-Third international joint conference on Artificial Intelligence, AAAI Press, 2013, pp. 2761–2767.

[20] W. Feng, J. Wang, Retweet or not? personalized tweet re-ranking, in: Proceedings of the sixth ACM international conference on Web search and data mining, ACM, New York, NY, USA, 2013, pp. 577–586.

[21] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, G. Zhang, Transfer learning using computational intelligence: A survey, Knowledge-Based Systems 80 (2015) 14–23. doi:10.1016/j.knosys.2015.01.010.

[22] R. Caruana, Multitask learning, in: S. Thrun, L. Pratt (Eds.), Learning to Learn, Learning to Learn, Springer US, 1998, pp. 95–133.

[23] A. Argyriou, T. Evgeniou, M. Pontil, Multi-task feature learning, in: Advances in Neural Information Processing Systems, MIT Press, 2007, pp. 41–48.

[24] T. Evgeniou, M. Pontil, Regularized multi–task learning, in: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, 2004, pp. 109–117.

[25] Y. Zhu, E. Zhong, S. J. Pan, X. Wang, M. Zhou, Q. Yang, Predicting user activity level in social networks, in: Proceedings of the 22nd ACM international conference on Conference on information and knowledge management, ACM, New York, NY, USA, 2013, pp. 159–168.

[26] G. Li, S. C. Hoi, K. Chang, W. Liu, R. Jain, Collaborative online multitask learning, Knowledge and Data Engineering, IEEE Transactions on 26 (8) (2014) 1866–1876. `doi:10.1109/TKDE.2013.139`.

[27] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, The Journal of Machine Learning Research 3 (2003) 993–1022.

[28] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. Mccauley, M. Franklin, S. Shenker, I. Stoica, Fast and interactive analytics over hadoop data with spark, in: NETWORKED SYSTEMS, Vol. 37, USENIX Association, 2012.

[29] L. Page, S. Brin, R. Motwani, T. Winograd, The pagerank citation ranking: Bringing order to the web., Technical report, Stanford InfoLab (1999).

[30] Z. Xu, Y. Zhang, Y. Wu, Q. Yang, Modeling user posting behavior on social media, in: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, ACM, New York, NY, USA, 2012, pp. 545–554.