Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Measuring the spreadability of users in microblogs*

Zhao-yun DING†[1,2], Yan JIA[2], Bin ZHOU[2], Yi HAN[2], Li HE[2], Jian-feng ZHANG[2]

([1]*College of Information Systems and Management, National University of Defense Technology, Changsha 410073, China*)

([2]*School of Computer Science, National University of Defense Technology, Changsha 410073, China*)

†E-mail: zyding@nudt.edu.cn

Received Mar. 13, 2013; Revision accepted June 14, 2013; Crosschecked Aug. 7, 2013

Abstract: Message forwarding (e.g., retweeting on Twitter.com) is one of the most popular functions in many existing microblogs, and a large number of users participate in the propagation of information, for any given messages. While this large number can generate notable diversity and not all users have the same ability to diffuse the messages, this also makes it challenging to find the true users with higher spreadability, those generally rated as interesting and authoritative to diffuse the messages. In this paper, a novel method called SpreadRank is proposed to measure the spreadability of users in microblogs, considering both the time interval of retweets and the location of users in information cascades. Experiments were conducted on a real dataset from Twitter containing about 0.26 million users and 10 million tweets, and the results showed that our method is consistently better than the PageRank method with the network of retweets and the method of retweetNum which measures the spreadability according to the number of retweets. Moreover, we find that a user with more tweets or followers does not always have stronger spreadability in microblogs.

Key words: Spreadability, Influence, PageRank, Microblogs, Social media, Social network, SpreadRank
doi:10.1631/jzus.CIIP1302 **Document code:** A **CLC number:** TP391

1 Introduction

Microblogs such as Twitter have rapidly become significant means for people to communicate with the world and each other. Message forwarding (e.g., retweeting on Twitter.com) is one of the most popular functions in many existing microblogs. For example, people can choose to retweet messages on their blog space in twitter. In this way, the information carried by the message can be quickly spread in microblogs. There have been a large number of studies about information spread surrounding microblogs, focusing on areas such as inferring and modeling information diffusion (Yang and Leskovec, 2010), finding patterns of spread (Romero *et al.*, 2011b), and various centrality measures (Bakshy *et al.*, 2011).

However, one area that has not received much attention is trying to better understand the spreadability of users. Traditional methods to measure the spreadability of users consider only the link out such as the number of retweets, and neglect the time interval of retweets and the location of users in information cascades.

The time interval of retweets, which stands for the diffused rate of each user, is an important feature to measure the spreadability of users. The lower is the time interval of retweets, the higher is the diffused rate. Fig. 1 shows an example of the importance of the time interval of retweets. Although the numbers of users who are activated by user C and user D are the same, user C spreads the information much faster than user D and we can infer that the spreadability of user C is higher.

Also, the location of users in information cascades, which stands for the ability to drive the propagation of information, is an important feature to

* Project supported by the National Natural Science Foundation of China (Nos. 60933005 and 91124002), the National High-Tech R&D Program (863) of China (Nos. 012505, 2011AA010702, 2012AA01A401, and 2012AA01A402), the 242 Information Security Program (No. 2011A010), and the National Science and Technology Support Program (Nos. 2012BAH38B04 and 2012BAH38B06), China

measure the spreadability of users. The earlier is the time of users in information cascades, the higher is the ability to drive the propagation of information. The assumption is reasonable, as the source of information cascades is actually important in microblogs and usually plays an important role in spreading the information. Fig. 2 shows an example of an information cascade whose depth is four. The spreadability of upper nodes is higher than that of lower nodes in information cascades.

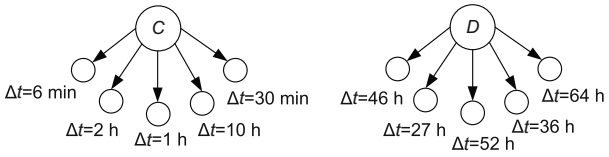

Fig. 1 An example of the importance of the time interval—user *C* spreads the information much faster than user *D*

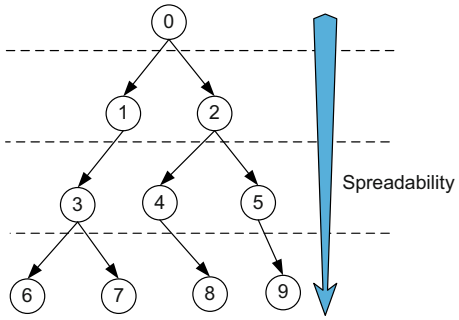

Fig. 2 An example of the information cascade

Moreover, there is a transitive relation of the spreadability. If a user with high spreadability retweets user B, we can infer that user B also has a high spreadability. The transitive relation of the spreadability is similar to the method of PageRank. The factor of restart in PageRank is $1/n$, on the assumption that each node has the same probability of being visited by all other nodes. However, in practice, if a user usually receives the information late, other users will have a lower probability of retweeting the information of this user. So, we cannot take advantage of the traditional Markov processes to measure the spreadability of users.

In this paper, we propose a novel variant PageRank method, SpreadRank, to measure the spreadability of users in microblogs, taking advantage of both the time interval of retweets and the location of users in information cascades. Experimental results indicate that the method of SpreadRank is consistently better than other methods for measuring the spreadability of users in microblogs.

2 Related work

With the popularity of microblogs, there have been a large number of studies about microblogs, focusing on the influence of users and the information diffusion.

Tunkelang (2009) originally proposed a method analog to PageRank, later named TunkRank, to measure the influence of users in microblogs. Weng *et al.* (2010) proposed an algorithm called TwitterRank to measure the influence taking both the topical similarity between users and the link structure into account. Cha *et al.* (2010) presented an in-depth comparison of three measures of influence: indegree, retweets, and mentions. Lee *et al.* (2010) found influential individuals based on the temporal order of information adoption in Twitter. Pal and Counts (2011) categorized tweets into three categories—original tweet (OT), conversational tweet (CT), and repeated tweet (RT)—to identify topical authorities in microblogs. Bakshy *et al.* (2011) referred narrowly to the influencer as the ability to consistently seed the cascades that spread further than others, in which seed nodes had higher influence. Romero *et al.* (2011a) proposed an algorithm that determines the influence and passivity of users based on their information forwarding activity; however, they neglected the time interval of retweets and the location of users in information cascades.

Yang and Counts (2010) constructed a novel model to capture the three major properties of information diffusion—speed, scale, and range—by analyzing information diffusion on Twitter, via users' ongoing social interactions as denoted by mentions. Yang and Leskovec (2010) developed a linear influence model to model information diffusion on Twitter. Ye and Wu (2010) analyzed the propagation patterns of general messages and showed how breaking news (e.g., Michael Jackson's death) spread through Twitter. Furthermore, they evaluated different social influences by examining their stabilities, assessments, and correlations. Romero *et al.* (2011b) analyzed the ways in which tokens known as hashtags spread on a network defined by the interactions among Twitter users. They found significant

variation in the ways in which widely-used hashtags on different topics spread. Wu *et al.* (2011) studied several longstanding questions in media communications research, in the context of the microblogging service Twitter, regarding the production, flow, and consumption of information. Sadikov *et al.* (2011) addressed the problem of missing data in information cascades on Twitter. Yang and Leskovec (2011) studied temporal patterns associated with online content and how the content's popularity grew and faded over time. Macskassy and Michelson (2011) developed retweet behavior models with the sign 'RT @user' and analyzed what information was being spread and why it was being spread. Kwak *et al.* (2010) constructed retweet trees with the sign 'RT @user' and demonstrated how retweets spread and how many got involved. Zaman *et al.* (2010) found retweets by looking for the string 'RT @' in the body of the tweet and presented a new methodology for predicting the spread of information in microblogs. Letierce *et al.* (2010) conducted a preliminary analysis of the retweets with the sign 'RT @user' to figure out how messages were spread. Myers *et al.* (2012) presented a model in which information could reach a node via the links of the social network or through the influence of external sources. ver Steeg and Galstyan (2012) studied information transfer in microblogs. Tsur and Rappoport (2012) presented an efficient hybrid approach based on a linear regression for predicting the spread of an idea in a given time frame on Twitter. Ding *et al.* (2013) proposed a novel method to mine topical influencers based on the multi-relational network in micro-blogging sites. However, how to measure the spreadability of users was neglected in their works.

Another interesting line of related research aims to combine PageRank with the temporal information. Berberich *et al.* (2004) developed the T-Rank algorithm, a link analysis method that takes into account the temporal aspects: freshness and activity of pages and links. Yu *et al.* (2005) proposed a time-weighted PageRank, in which the inlinks of a page are weighted according to their timestamps. Liu *et al.* (2008) proposed a method called BrowseRank to compute the page importance, considering the lengths of staying time spent on the pages by users. However, the aims of their methods were to measure the page importance, whereas the location of nodes in information cascades was neglected in their works. We combine the time interval of retweets and the location of users in information cascades.

3 Methods

3.1 Data collection

For the purpose of this study, a set of Twitter data about Chinese-based twitters who have published at least one Chinese tweet was prepared. About 0.26 million users and 10 million tweets were collected through the application programming interface (API) of Twitter.com. The method of getting the dataset is described in detail as follows:

1. We first get the top 100 Chinese-based twitters and their tweets from Twitter.com, and denote the aggregate of these users as seed set Set_0.

2. We then crawl all followers and friends of each seed twitter $s \in \mathrm{Set}_0$. At the same time, all tweets of these users are crawled.

3. If followers and friends are Chinese users and they were not processed before, we add them to the new seed set $\mathrm{Set}_{\mathrm{seed}}$.

4. Loop 1–3.

We first extract from the dataset information cascades which correspond to distinct diffusion messages, where each message comprises a single initiator, or 'seed' retweeted by other users. Then we begin by describing the information cascades that we are trying to mine. As illustrated in Fig. 3, the distribution of information cascades' sizes is approximately power-law, implying that the vast majority of information does not spread at all, while a small fraction is retweeted many times. The depth of the information cascades is right skewed, where the deepest information cascades can propagate as far as nine generations from their origin.

Then we combine all information cascades to construct a directed weighted graph $G{=}(V, E, W(E))$ and measure the spreadability in graph G according to our method named SpreadRank.

3.2 Weight computation

The weights of the diffused graph represent how frequently the information is diffused from a user to another user. The naive method to determine the weights is to count the total number of retweets. However, the number of tweets published by users is

ignored. For example, in Fig. 4, although the number of retweets from user B is larger than that from user C, user A retweeted most of the tweets of user C and only a small number of the tweets of user B. We can infer that it is easier to spread the information from user C to user A.

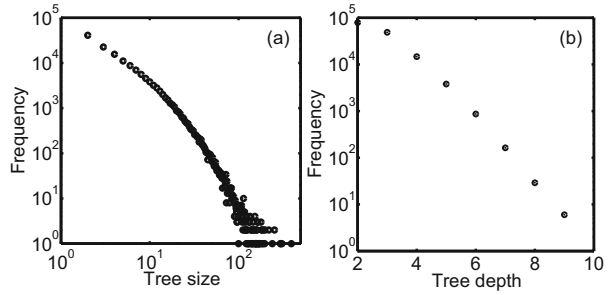

Fig. 3 Distribution of information cascades' size (a) and depth (b)

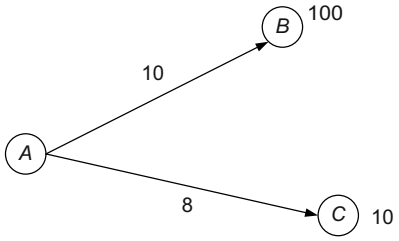

Fig. 4 An example of retweets and tweets. The number of retweets from user *B* is larger than that from user *C*; however, user *A* retweets most of the tweets of user *C* and only a small proportion of the tweets of user *B*

Next, we consider the weights measure r/T, called DivT, where r stands for the number of retweets and T stands for the number of tweets. The rationale behind this measure is that it counts the proportion (rather than the actual number) of one's tweets which are retweeted. Intuitively, the higher the proportion of a user's tweets which are retweeted, the higher the spreadability of the user. In practice, however, we find that in general this measure overpenalizes users with a large number of tweets, underpenalizes users with a small quantity of tweets, and is overly sensitive to spuriously large values of r when T is small. Then, to strike a balance, we consider the weights measure $r/\log T$, called DivLogT, to measure the spreadability from a user to another user in microblogs.

3.3 Analysis of location of users in information cascades

In this subsection, we analyze the impact of the location of users in information cascades. Two key factors may be explored:

1. The spreadability of different users in an information cascade is different. We refer to the spreadability of user A in an information cascade as the size of the cascaded subtree whose root node is user A, i.e., $\mathrm{sp}_A=|\mathrm{set}_A|$, where $\mathrm{set}_A \in \mathrm{children}(A)$ (i.e., set_A is the set of user A's children).

2. A user may be in many information cascades. Intuitively, the more often a user is in information cascades, the higher the spreadability of the user.

However, it is difficult to combine these two factors to synthetically measure the spreadability of each user. For example, in Fig. 5, user A is in three information cascades and user B only in one information cascade, but the information of user B is diffused to more users.

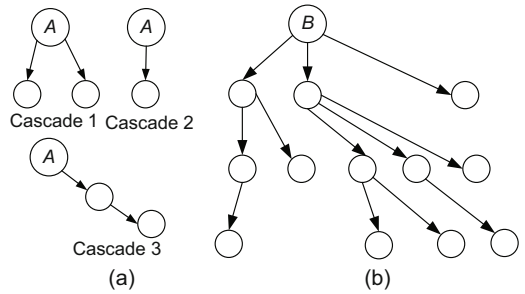

Fig. 5 A sample of information cascades of user *A* (a) and user *B* (b). User *A* is in three information cascades and user *B* only in one information cascade, but the information of user *B* is diffused to more users

To combine these two factors to measure the spreadability of each user synthetically, we count all children nodes in each information cascade for the user who is the root node of each cascaded subtree. Thus, the impact of the location of users in information cascades can be defined as follows:

$$\mathrm{scorel}_A=\sum\nolimits_{0<a\le n}|\mathrm{set}_A^a|,\ \mathrm{set}_A^a\in\mathrm{children}(A^a), \quad (1)$$

where n is the number of cascaded subtrees whose root node is user A.

3.4 Analysis of the time interval of retweets

The lower is the time interval of retweets, the higher is the diffused rate. In this subsection, we analyze the impact of the time interval of retweets. First,

we analyze the distribution of sample time intervals of retweets with $\text{size}_{\text{sample}}$=71 000 (Figs. 6a and 6b). Experimental results indicate that the distribution of time intervals of retweets is approximately power-law, implying that the vast majority of information is retweeted in a short duration, while a small fraction is retweeted for a long time (We can estimate the parameter ∂=1.63 of the power-law distribution).

However, the tail of this distribution is smaller than those of traditional power-law distributions, because a topic does not usually go on for a long time. Fig. 7 shows the distribution of five topics' durations, implying that a topic generally was discussed actively for 10 days at most, and then died out rapidly or gradually. In Fig. 7, if the topics went on for a very long time, we will ignore these timestamps. So, we can infer that if a user spreads the information after 10 days, it will have few contributions for the propagation of the information and this diffusion will be ignored.

After removing the time intervals that are longer than 10 days (1230 sample time intervals are removed), we find that the time intervals satisfy a negative exponential distribution (Figs. 6c and 6d). Then we can infer the parameter of the negative exponential distribution λ=1.9768×10^4 according to the rest of the sample time intervals. We give the probability density function of the negative exponential distribution as follows:

$$f(x)=\begin{cases}\dfrac{1}{19\,768}\mathrm{e}^{-x/19\,768}, & 0\le x\le 192\text{ h},\\ 0, & x>192\text{ h}.\end{cases} \tag{2}$$

Next, we can measure the impact of the time interval of retweets according to the probability density function of the negative exponential distribution. The lower is the time interval of retweets, the higher the diffused rate becomes.

$$f(\Delta t)=\begin{cases}\dfrac{1}{\lambda}\mathrm{e}^{-\Delta t/\lambda}, & 0\le \Delta t\le 192\text{ h},\\ 0, & \Delta t>192\text{ h},\end{cases} \tag{3}$$

where Δt is the time interval of retweets.

3.5 SpreadRank

In this subsection, we combine the time interval of retweets and the location of users in information cascades to measure the spreadability of users. We propose a novel variant PageRank method called SpreadRank.

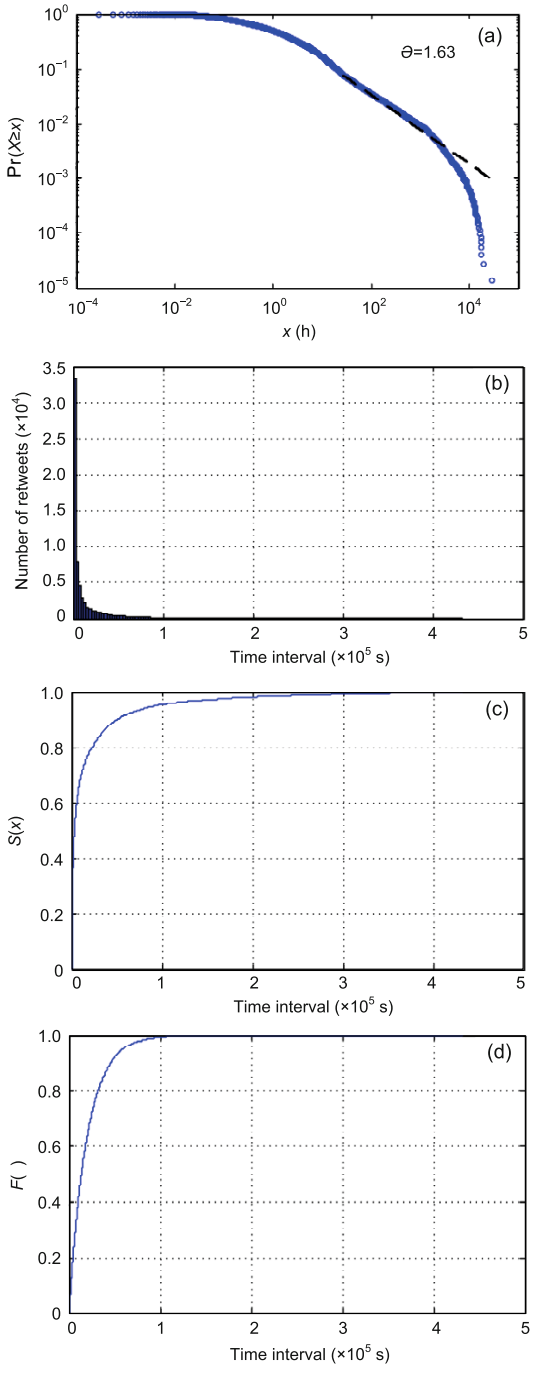

Fig. 6 Negative exponential distribution: (a) distribution of time intervals; (b) histogram of data; (c) fitting empirical distribution of sample $S(x)$; (d) theoretical distribution of hypothesis testing $F(x)$

Two key steps in the technique described above may be explored. One is, for each user who retweets the tweets of his/her friends, how to divide his/her score to his/her friends. The other is, how to measure the factor of restart in our method, that is, how to

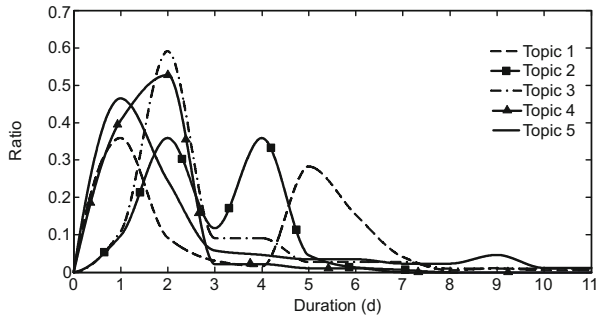

Fig. 7 The distribution of topics' durations

give a jump probability to each user not considering the web part.

For the first step, we compute the transition probability of each user. Two features need to be considered in this process: (1) the weight of retweets, and (2) the time interval of retweets. The higher the weight of retweets and the shorter the time interval of retweets, the higher the transition probability. We combine these two features to compute the transition probability of each user.

The transition ability from user u_i to his/her friend u_j is defined as follows:

$$p(u_j|u_i)=\frac{\sum_{r_{ij}} f(\Delta t_{ij})}{\log T_j}, \tag{4}$$

where Δt_{ij} is the time interval of retweets, and T_j stands for the number of user u_j's tweets.

In the above formula, we treat all retweets respectively considering the time interval, and then sum up all scores to synthetically measure the spreadability.

Next, we take advantage of all transition probabilities from user u_i to all his/her friends $u \in \mathrm{out}(u_i)$ who are retweeted by user u_i to divide his/her score to a friend u_j.

Given the transition matrix $\boldsymbol{P}$, the transition probability of a random walk from user u_i to user u_j is defined as follows:

$$P(u_j|u_i)=\frac{(\sum_{r_{ij}} f(\Delta t_{ij}))/\log T_j}{\sum\limits_{u\in \mathrm{out}(u_i)} (\sum_{r_{iu}} f(\Delta t_{iu}))/\log T_u}. \tag{5}$$

For the second step, we compute the factor of restart in a random walk. The factor of restart in traditional PageRank is $1/n$, on the assumption that each node has the same probability to be visited by all other nodes. However, in practice, if a user usually receives the information late, other users will have a lower probability to retweet the information of this user.

We take advantage of the location of users in information cascades to measure the factor of restart, to ensure that users receiving the information earlier have higher scores.

Given teleport vector (the factor of restart) $\boldsymbol{e}$, each element e_i can be defined as

$$e_i=\frac{\mathrm{scorel}_i+\mu}{\sum\limits_{A\in V}(\mathrm{scorel}_A+\mu)}, \tag{6}$$

where μ is a smoothing factor to avoid the zero of the scorel. If a user has never been retweeted by others, the location of users' scorel in information cascades will be equal to 0. Also, it will ensure the convergence of the Markov processes; that is, each node in graph G can be visited by a random walk.

Finally, we can measure the spreadability of user u_j according to the web-graph part and the teleport vector:

$$\mathrm{SR}(u_j)=\alpha \sum_{u_i\in \mathrm{in}(u_j)} P(u_j|u_i)\mathrm{SR}(u_i)+(1-\alpha)e_j, \tag{7}$$

where $\mathrm{in}(u_j)$ represents all followers of user u_j, α is a jump factor, and $\mathrm{SR}(u_i)$ stands for the spreadability of user u_i.

Given the score vector $\boldsymbol{r}$ which measures the spreadability of all users and the transition matrix $\boldsymbol{P}$ which measures the random walk in graph G, the above expression can be rewritten as

$$\boldsymbol{r}=\alpha \boldsymbol{P}\boldsymbol{r}+(1-\alpha)\boldsymbol{e}. \tag{8}$$

Although the above expression is similar to PageRank, the transition probabilities and the teleport vector are different from those of PageRank. In the above expression named SpreadRank, we take advantage of the time interval of retweets and the location of users in information cascades, and the spreadability of users can be measured more accurately.

SpreadRank can be formulated in matrix notation as follows:

$$\boldsymbol{r}_k=\boldsymbol{r}_{k-1}\cdot \boldsymbol{M}. \tag{9}$$

The SpreadRank matrix $\boldsymbol{M}$ is computed as

$$\boldsymbol{M}=\alpha\cdot \boldsymbol{P}+(1-\alpha)\cdot \boldsymbol{e}\cdot \boldsymbol{1}, \tag{10}$$

where $\boldsymbol{1}$ is a vector with each element being 1.

Because the smoothing factor μ ensures that each node in graph G can be visited by a random walk, we can infer that the SpreadRank matrix $\boldsymbol{M}$ is a stochastic matrix, aperiodic and irreducible. Thus, it is also easy to see that this Markov chain is ergodic, and the stationary probabilities can be found as r_i^n $(n \to \infty)$, for any initial vector $\boldsymbol{r}^0$.

4 Experiments

4.1 Parameter setting

This subsection evaluates the choice of parameter α in SpreadRank. Usually, the damping factor α is set to 0.85 and experimental results (Boldi *et al.*, 2005) have proven that any minor change of α does not have a huge impact on final results in traditional PageRank. However, how the damping factor α influences the results of SpreadRank is unknown.

4.1.1 Influence on the rankings

First, we rank users according to the scores of SpreadRank. Next, we evaluate the impact of the damping factor α by comparing the rankings of users with different damping factors α. The Spearman rank correlation, denoted as ρ, is used to analyze the correlation of two rankings:

$$\rho = 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n-1} (r_i^1 - r_i^2)^2. \tag{11}$$

r^1 and r^2 are two rankings of SpreadRank under different damping factors α. The higher is the score of the Spearman rank correlation, the more correlative are the two rankings with different damping factors α.

Given a damping factor α and the corresponding ranking r, as well as its four neighboring damping factors α_1=α−0.01, α_2=α−0.02, α_3=α+0.01, α_4=α+0.02 and the corresponding rankings r_1, r_2, r_3, r_4, we compute the Spearman rank correlation between the ranking r and the rankings r_1, r_2, r_3, r_4:

$$\bar{\rho} = \frac{\rho_1 + \rho_2 + \rho_3 + \rho_4}{4}. \tag{12}$$

The higher is the score of the Spearman rank correlation $\bar{\rho}$, the more correlative are the ranking r and its neighboring rankings r_1, r_2, r_3, r_4. Fig. 8 gives the experimental results.

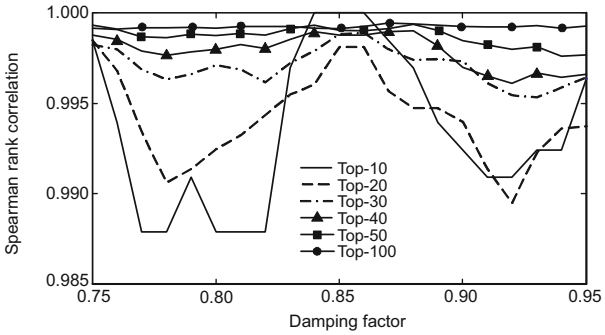

Fig. 8 The impact of the damping factor on the correlation of two rankings

Experimental results also show that any minor change of α does not have a huge impact on final results. With the increase of α, the Spearman rank correlation drops first, then goes up, next drops, and finally goes up. Because the impact of the web-graph part and the location of users are determined by α, when α is 0, the web-graph part of the process is annihilated, resulting in the final results dominated by the location of users scorel. As α approaches 1, the web part becomes more and more important, leading to the final results dominated by the network of retweets.

So, we can infer that it is more reasonable to set α as 0.85. Fig. 8 indicates that the corresponding ranking r is more correlative with α=0.85 than with the four neighboring damping factors.

4.1.2 Influence on time cost

We analyze how the damping factor α influences the time cost of the SpreadRank. Fig. 9 gives the experimental results.

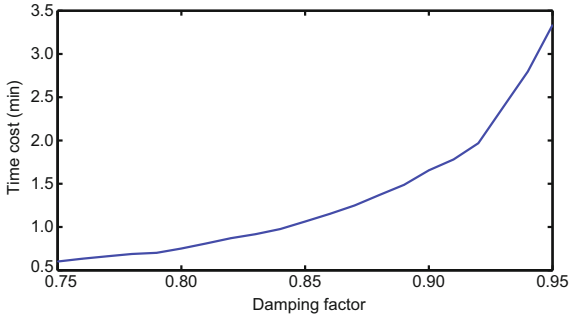

Fig. 9 The time cost at different damping factors

With the increase of α, the time cost of SpreadRank increases. Because the web-graph part of

the process becomes more and more important, the speed of convergence becomes slow with the increase of α. However, the time cost of SpreadRank is not too high. Only about 1 min is spent with $\alpha = 0.85$ by SpreadRank.

4.2 Methods for comparison

We compare the following methods for measuring the spreadability of users:

1. The spreadability of user A is measured by the number of users who retweet the tweets of user A. We call this method retweetNum.

2. The spreadability of users is measured by the method of PageRank in the network of retweets. We call this method retweetRank.

$$\boldsymbol{r} = \alpha \boldsymbol{W} \boldsymbol{r} + (1-\alpha)\boldsymbol{i}, \tag{13}$$

where the transition matrix $\boldsymbol{W}$ is computed considering only the weights of the network and $\boldsymbol{i}$ is a vector with each element being $1/n$.

3. The spreadability of users is measured by the method of variant PageRank in the network of retweets, also considering the time interval of retweets. We call this method retweetRank+T.

$$\boldsymbol{r} = \alpha \boldsymbol{P} \boldsymbol{r} + (1-\alpha)\boldsymbol{i}. \tag{14}$$

4. The spreadability of users is measured by the method of variant PageRank in the network of retweets, also considering the impact of the locations of users in information cascades. We call this method retweetRank+L.

$$\boldsymbol{r} = \alpha \boldsymbol{W} \boldsymbol{r} + (1-\alpha)\boldsymbol{e}. \tag{15}$$

5. Our method is called SpreadRank.

4.3 Effectiveness verificatinon

To evaluate the effectiveness of our algorithm SpreadRank from the perspectives of the link structure and the time interval, we introduce the metric of coverage. Since users in microblogs affect the propagation of information, given a seed user and a period of time, the number of users who can be activated by the seed user should be a good indicator of how good the spreadability is. We introduce the concept of 'coverage' due to this intuition.

Definition (Coverage) Given a seed node in a network and a period of time, the coverage is defined as the number of nodes that are either directly or indirectly activated by this seed node, i.e., how many users are influenced in a period of time by a seed node in the propagation of information.

Fig. 10 gives an example of computing the coverage of seed node 0.

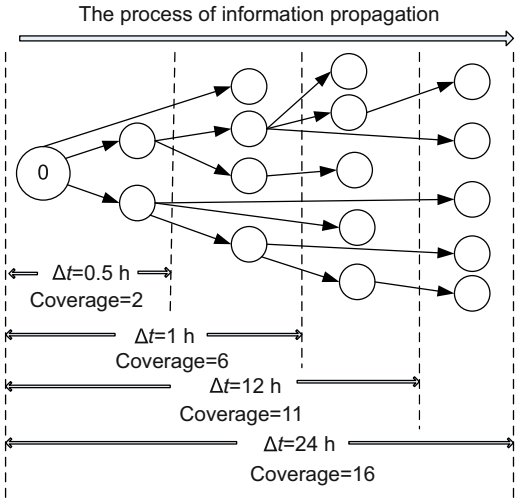

Fig. 10 An example of computing the coverage

In the experiments, we set up periods of time as follows: Δt_1=0.2 h, Δt_2=0.5 h, Δt_3=1 h, Δt_4=3 h, Δt_5=6 h, Δt_6=12 h, Δt_7=18 h, Δt_8=24 h, Δt_9=36 h, Δt_{10}=48 h, Δt_{11}=72 h, and Δt_{12}=96 h. Also, we obtain the top-k users according to each method ranking based on the spreadability of users, where k=7, 9, 15, and 18. The coverage of the top-k users is defined as the sum of all top-k users' coverage. Fig. 11 gives the experimental results.

Experimental results indicate that the method of SpreadRank is consistently better than other methods for measuring the spreadability of users in microblogs. For the method of SpreadRank, the spreadability of users is consistently stronger than other methods and more users are activated in the appointed periods of time with top-k seed users. The methods of retweetRank+T and retweetRank+L are better than the methods of retweetRank and retweetNum, and the method of retweetNum is the worst.

Moreover, we find that the spreadability of users for the method of retweetRank+T is stronger than that for the method of retweetRank+L with lower periods of time, but weaker with higher periods of time. Because the time interval is considered in the method of retweetRank+T, the information is diffused faster and more users are activated in lower

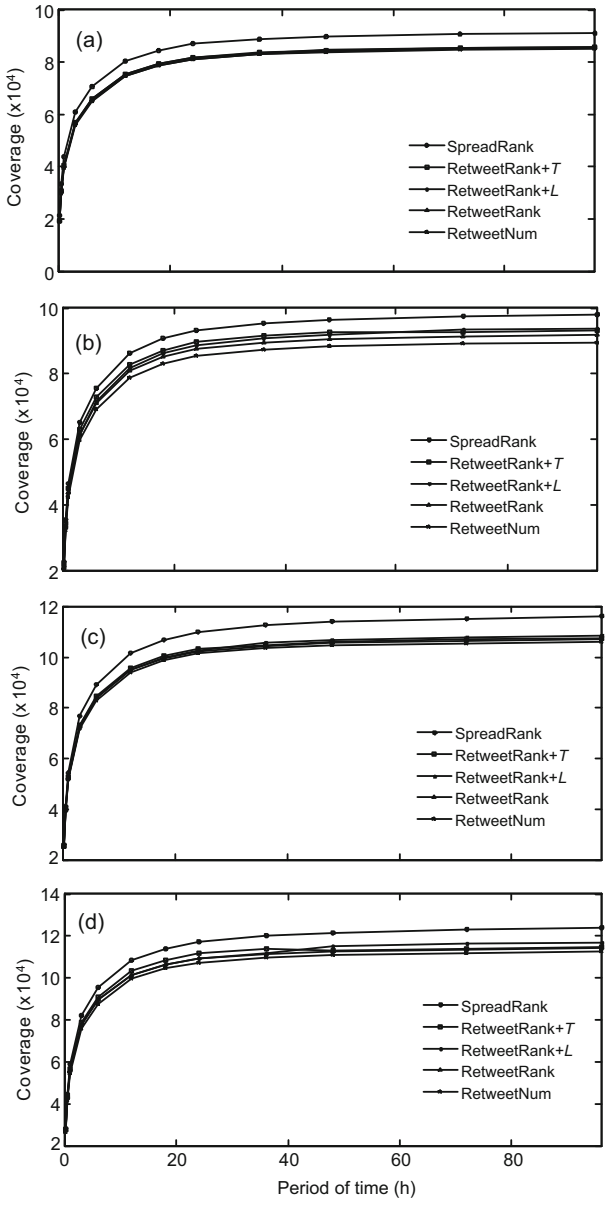

Fig. 11 The coverage with different periods of time for the top-7 (a), top-9 (b), top-15 (c), and top-18 (d) users, at the 12 periods of time of 0.2, 0.5, 1, 3, 6, 12, 18, 24, 36, 48, 72, and 96 h

periods of time. However, the location of users in information cascades is considered in the method of retweetRank+L; the information is diffused further and more users are activated in higher periods of time.

4.4 Characteristics of users with higher spreadability

Here we analyze the characteristics of users with higher spreadability, including a large number of tweets and a large number of followers. We try to answer two questions: (1) Does a user with higher spreadability publish more tweets? (2) Does a user with higher spreadability have more followers?

We get the top-k users according to rankings by the spreadability of users, and then obtain the numbers of tweets or followers of these users. Fig. 12 gives the correlation between the number of tweets or followers and the ranking based on the spreadability of users.

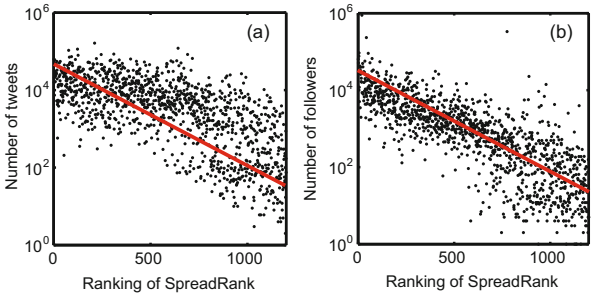

Fig. 12 The correlation between the ranking of SpreadRank and the number of tweets (a) or between the ranking of SpreadRank and the number of followers (b)

The general trend of spreadability and the number of tweets or followers is correlative; moreover, the general trend of spreadability and the number of followers is more correlative. However, users with higher spreadability do not always publish more tweets or have more followers in the local part. So, we can infer that a user with more tweets or followers does not always have stronger spreadability in microblogs.

5 Conclusions

A novel method called SpreadRank is proposed to measure the spreadability of users in microblogs, considering both the time interval of retweets and the location of users in information cascades. Experimental results indicate that this method is consistently better than other methods. Moreover, we find that a user with more tweets or followers does not always have stronger spreadability in microblogs.

In future work, we will combine the spreadability and features of users to measure the influence of users more accurately in microblogs.

References

Bakshy, E., Hofman, J.M., Mason, W.A., Watts, D.J., 2011. Everyone's an Influencer: Quantifying Influence on Twitter. Proc. 4th ACM Int. Conf. on Web Search and Data Mining, p.65-74. [doi:10.1145/1935826.1935845]

Berberich, K., Vazirgiannis, M., Weikum, G., 2004. T-Rank: Time-Aware Authority Ranking. Proc. Algorithms and Models for the Web-Graph: 3rd Int. Workshop, p.131-141. [doi:10.1007/978-3-540-30216-2_11]

Boldi, P., Santini, M., Vigna, S., 2005. PageRank as a Function of the Damping Factor. Proc. 14th Int. Conf. on World Wide Web, p.557-566. [doi:10.1145/1060745.1060827]

Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P., 2010. Measuring User Influence in Twitter: the Million Follower Fallacy. Proc. 4th Int. AAAI Conf. on Weblogs and Social Media, p.10-17.

Ding, Z., Jia, Y., Zhou, B., Han, Y., 2013. Mining topical influencers based on the multi-relational network in micro-blogging sites. *China Commun.*, **10**(1):93-104. [doi:10.1109/CC.2013.6457533]

Kwak, H., Lee, C., Park, H., Moon, S., 2010. What Is Twitter, a Social Network or a News Media? Proc. 19th Int. Conf. on World Wide Web, p.591-600. [doi:10.1145/1772690.1772751]

Lee, C., Kwak, H., Park, H., Moon, S., 2010. Finding Influentials Based on the Temporal Order of Information Adoption in Twitter. Proc. 19th Int. Conf. on World Wide Web, p.1137-1138. [doi:10.1145/1772690.1772842]

Letierce, J., Passant, A., Breslin, J.G., 2010. Understanding How Twitter Is Used to Spread Scientific Messages. Proc. Web Science Conf., p.91-100.

Liu, Y., Gao, B., Liu, T., Zhang, Y., Ma, Z., He, S., Li, H., 2008. BrowseRank: Letting Web Users Vote for Page Importance. Proc. 31st Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, p.451-458. [doi:10.1145/1390334.1390412]

Macskassy, S.A., Michelson, M., 2011. Why Do People Retweet? Anti-homophily Wins the Day! Proc. 5th Int. AAAI Conf. on Weblogs and Social Media, p.209-216.

Myers, S., Zhu, C., Leskovec, J., 2012. Information Diffusion and External Influence in Networks. Proc. 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.33-41. [doi:10.1145/2339530.2339540]

Pal, A., Counts, S., 2011. Identifying Topical Authorities in Microblogs. Proc. 4th ACM Int. Conf. on Web Search and Data Mining, p.45-54. [doi:10.1145/1935826.1935843]

Romero, D.M., Galuba, W., Asur, S., Huberman, B.A., 2011a. Influence and Passivity in Social Media. Proc. 20th Int. Conf. on World Wide Web, p.113-114. [doi:10.1145/1963192.1963250]

Romero, D.M., Meeder, B., Kleinberg, J., 2011b. Differences in the Mechanics of Information Diffusion across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter. Proc. 20th Int. Conf. on World Wide Web, p.695-704. [doi:10.1145/1963405.1963503]

Sadikov, E., Medina, M., Leskovec, J., Garcia-Molina, H., 2011. Correcting for Missing Data in Information Cascades. Proc. 4th ACM Int. Conf. on Web Search and Data Mining, p.55-64. [doi:10.1145/1935826.1935844]

Tsur, O., Rappoport, A., 2012. What's in a Hashtag? Content Based Prediction of the Spread of Ideas in Microblogging Communities. Proc. 5th ACM Int. Conf. on Web Search and Data Mining, p.643-652. [doi:10.1145/2124295.2124320]

Tunkelang, D., 2009. A Twitter Analog to PageRank. Available from http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank/

ver Steeg, G., Galstyan, A., 2012. Information Transfer in Social Media. Proc. 21st Int. Conf. on World Wide Web, p.509-518. [doi:10.1145/2187836.2187906]

Weng, J., Lim, E.P., Jiang, J., He, Q., 2010. TwitterRank: Finding Topic-Sensitive Influential Twitters. Proc. 3rd ACM Int. Conf. on Web Search and Data Mining, p.261-270. [doi:10.1145/1718487.1718520]

Wu, S., Hofman, J.M., Mason, W.A., Watts, D.J., 2011. Who Says What to Whom on Twitter. Proc. 20th Int. Conf. on World Wide Web, p.705-714. [doi:10.1145/1963405.1963504]

Yang, J., Counts, S., 2010. Predicting the Speed, Scale, and Range of Information Diffusion in Twitter. Proc. 4th Int. AAAI Conf. on Weblogs and Social Media, p.355-358.

Yang, J., Leskovec, J., 2010. Modeling Information Diffusion in Implicit Networks. Proc. 10th IEEE Int. Conf. on Data Mining, p.599-608. [doi:10.1109/ICDM.2010.22]

Yang, J., Leskovec, J., 2011. Patterns of Temporal Variation in Online Media. Proc. 4th ACM Int. Conf. on Web Search and Data Mining, p.177-186. [doi:10.1145/1935826.1935863]

Ye, S., Wu, F., 2010. Measuring message propagation and social influence on Twitter.com. *LNCS*, **6430**:216-231. [doi:10.1007/978-3-642-16567-2_16]

Yu, P.S., Li, X., Liu, B., 2005. Adding the Temporal Dimension to Search - a Case Study in Publication Search. Proc. IEEE/WIC/ACM Int. Conf. on Web Intelligence, p.543-549. [doi:10.1109/WI.2005.21]

Zaman, T.R., Herbrich, R., Gael, J., Stern, D., 2010. Predicting Information Spreading in Twitter. Proc. Workshop on Computational Social Science and the Wisdom of Crowds, p.20-29.