Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 604294, 10 pages
http://dx.doi.org/10.1155/2014/604294

Research Article

Predicting the Times of Retweeting in Microblogs

Li Kuang,[1] Xiang Tang,[2] and Kehua Guo[3]

[1] *School of Software, Central South University, Changsha 410075, China*
[2] *Hangzhou Institute of Services Engineering, Hangzhou Normal University, Hangzhou 310012, China*
[3] *School of Information Science and Engineering, Central South University, Changsha 410075, China*

Correspondence should be addressed to Kehua Guo; guokehua@csu.edu.cn

Received 8 August 2013; Accepted 20 August 2013; Published 11 February 2014

Academic Editor: Zhongmei Zhou

Recently, microblog services accelerate the information propagation among peoples, leaving the traditional media like newspaper, TV, forum, blogs, and web portals far behind. Various messages are spread quickly and widely by retweeting in microblogs. In this paper, we take Sina microblog as an example, aiming to predict the possible number of retweets of an original tweet in one month according to the time series distribution of its top n retweets. In order to address the problem, we propose the concept of a tweet's lifecycle, which is mainly decided by three factors, namely, the response time, the importance of content, and the interval time distribution, and then the given time series distribution curve of its top n retweets is fitted by a two-phase function, so as to predict the number of its retweets in one month. The phases in the function are divided by the lifecycle of the original tweet and different functions are used in the two phases. Experiment results show that our solution can address the problem of predicting the times of retweeting in microblogs with a satisfying precision.

1. Introduction

Microblog is a social network based platform where information can be shared, propagated, and obtained. Users can publish their tweets through SMS, instant messenger, email, web sites, or third-party applications by inputting at most 140 words [1]. Microblog bloomed rapidly due to its numerous advantages such as real-time and high interaction. The number of Sina microblog users in China has reached up to 250 million during 2 years [2], and it has become a very important Internet application for nearly half of Chinese netizens.

Retweeting is a very important user behavior in microblogs. Users can forward the tweets which they are interested in, so that the followers of the users can see the tweets as well. The tweet publishing pattern and propagation form, as well as its concise presentation with multimedia added such as music, video, and pictures, make the information spreading faster in microblog than that in traditional media, with the content and form being more diverse. Therefore, how to predict the times of retweeting in microblogs by analyzing the features of tweets propagation becomes a hot research topic.

The result of the research can be applied in many areas: a tweet that is retweeted largely represents a hot topic, so the prediction on the times of retweeting can help find hot topics in microblog. Second, a hot tweet can represent the focus that most people are concerned about so we can monitor public opinions in a better fashion by predicting the times of retweeting. Moreover, microblog reacts more rapidly compared to traditional media, especially on social emergency, so traditional media like newspaper can draft news based on the latest hot tweets in microblog.

The 13th International Conference on Web Information System Engineering (WISE 2012) [3] organized a challenge on Sina microblog. The organizers collected a number of retweets related to 33 original tweets from Sina microblog. There are about 100 retweeting records corresponding to each original tweet. One of the proposed challenges is to predict the times of retweeting of the 33 original tweets in one month. Motivated by the challenge proposed in WISE 2012, we addressed the significant problem by three steps:

first, the primitive data are divided into 33 groups, where the data in one group correspond to the retweets of an original tweet. For each group, the primitive data are parsed by extracting the values of property tags, so that the time series distribution of top 100 retweets for each original tweet can be derived. Second, calculate the lifecycle of each original tweet according to its content and the characteristic of the time series distribution of top 100 retweets including response time and interval. Third, in order to predict the times of retweeting of the 33 original tweets in one month, the derived time series distribution curves of top 100 retweets are fitted by a two-phase function, where the first phase is the calculated lifecycle of the original tweet and the second phase is the remainder time in one month. The value in the 1st phase is derived by fitting the curve by a lineal function, while the value in the 2nd phase is by a logarithm function. The final predicted value of retweeting times is the sum of the values of two phases. The experiments show that the proposed solution in this paper can greatly address the problem of predicting the times of retweeting in microblogs, and the average error is controlled within 20%.

The paper is organized as follows. Related work is introduced in Section 2. The form and volume of collected microblog data are introduced in Section 3. The detailed solution to predicting the times of retweeting is illustrated in Section 4. The experiment results are presented in Section 5. And finally the conclusions and future work are given.

2. Related Work

The blossom of microblog aroused wide attention of many researchers. Presently, they begin to conduct research on the problems related to microblogs, including analyzing the contents of microblogs, mining the association relation between microblogs and real society [4–11], and predicting whether a tweet will be retweeted as well as the characteristic of retweeting behavior [12–21].

In the related work on the analysis of microblog contents, researchers found that microblog plays an important role in many areas, for example, political elections, earthquake disaster, marketing management, and various kinds of information spreading [4–11]. Tumasjan et al. [6] find that the political emotion of tweet users has close relation with election and tweets can reflect voters' inclination in real society by using LIWC text analysis software. Bollen et al. [7] find that society, culture, politics, and economy have a great influence on public sentiment through extended emotional analysis. Sakaki et al. [8] successfully find out the earthquake epicenter from Twitter messages through time probability model, and Qu et al. [9] pointed out that microblogs play an important and positive role in disaster by comparing the content of microblogs before and after Yushu earthquake in 2010. Achananuparp et al. [10] proposed a model for describing users' originating and promoting behaviors so as to detect interesting events from sudden changes in aggregated information propagation behavior of Twitter users.

In the related work in retweeting tweets, many researchers study and analyze what contents and features of a tweet make it be retweeted more easily. For example, Chen and Zhang [12] predict whether a tweet will be retweeted based on its emotional or content keywords, user tags, and historical retweeting frequency. Xiong et al. [13] studied information diffusion on microblogs based on retweeting mechanism and proposed a diffusion model (SCIR) which contains four states, two of which are absorbing. Zhang et al. [14] predict whether a tweet will be retweeted by ranking tweets based on weighted feature model. Hong et al. [15] discuss why and how people retweet messages, as well as what messages will be retweeted by making use of TF-IDF points. Zaman et al. [16] predict the information spreading in Twitter through collaborative filtering algorithm. Petrovic et al. [1] decide whether a tweet will be retweeted by manual experiments and then predict it by improved passive progressing algorithm. However, few works on predicting the times that a message is retweeted are published.

Zhang et al. [22] propose to compute the probability that a user retweets a tweet by considering several features first and then build a retweet model with the probability to predict the number of possible views of a tweet. Unankard et al. [23] compare four different methods, of which the first one is discovering a regression function based on the popularity of messages and network connectivity, the second one is learning a classification model based on users' preferences in different fields of topics, the third one is simulating retweeting paths starting from a root message by employing Monte Carlo method, and the fourth is building a recommendation model based on collaborative filtering. Luo et al. [24] propose to identify most similar message from training data based on the similarity between their time series values in the same length period and then fit the ARMA models over the whole time series of the identified message, and finally the fitted model is applied to the test tweet to predict future values. Compared with their work, in this paper, we propose a new perspective to differentiate the time period when a tweet may be largely retweeted and that when the possibility of retweeting becomes small and propose a new concept, a tweet's lifecycle, which is determined by analyzing the content of the tweet as well as the time series distribution of its top n retweets. Based on the calculated lifecycle, different functions are fitted within and out of its lifecycle, so as to predict the number of retweets of a tweet in one month.

3. Dataset

In this paper, we take the Sina microblog data as an example to study the prediction on the times of retweeting. This section will introduce the form and volume of the collected raw data.

3.1. Data Form. The basic form of each datum in the collected dataset is as follows:

Tweet:time:A|#|mid:B|#|uid:C\tD\tE...|#|isContain Link:F|#|eventList:G|#|rtTime:H|#|rtMid:I|#|rtUid: J|#|rtIsContainLink:K|#|rtEventList:L.

In which the detailed meaning of each property tag is shown as Table 1.

Table 1: Data tags and their meanings.

time	The time when a re-tweeting message is issued, whose form is yyyy-mm-ddhh:mm:ss
mid	The unique identification ID of the re-tweeting message
uid	The user ID who publishes the re-tweeting message
isContainLink	Whether the re-tweeting message contains a link, whose value is of kind Boolean (true or false)
eventList	The event tags of the re-tweeting messages, that is, its keywords
rtTime	The time when the re-tweeted original tweet is published, whose form is yyyy-mm-ddhh:mm:ss
rtMid	The message ID of the original tweet
rtUid	The user ID who publishes the original tweet
rtIsContainLink	Whether the original message contains a link, whose value is of kind Boolean (true or false)
rtEventList	The event tags of the original messages

In order to illustrate the detailed meaning of every property more clearly, we take the following datum as an example:

time:2011-06-0511:26:56|#|mid:270926510254626223
8|#|uid:6701001061010001018429227021838|#|is
ContainLink:false|#|rtTime:2011-06-05 08:19:59|#|rt
Mid:2709258383303085289|#|rtUid:9256021720209
2828482|#|rtIsContainLink:false|#|rtEventList:Li Na win French Open in tennis$Francesca Schiavone.

The datum shows the following: the original tweet ID (rtMid) is 2709258383303085289, it was created and published by a user with ID 92560217202092828482 (rtUid) at 2011-06-05 08:19:59 (rtTime), it does not contain a link (rtIsContainLink: false), and it is about Li Na winning French Open in tennis with event tags "rtEventList:Li Na win French Open in tennis$Francesca Schiavone." The original tweet is retweeted by a user with uid 6701001061010001018429227021838 at 2011-06-05 11:26:56 (Time), its message ID (mid) is 2709265102546262238, and it does not contain a link (isContainLink:false).

Each primitive datum is constructed by such property-value pairs. We can find the retweeting time, retweeting message ID, the original tweet ID, event tags, and so forth from each datum, so as to understand and use each datum.

3.2. Data Volume. We eliminate repeated messages and finally got 3292 valid messages by preprocessing data based on integrity constraints. The 33 original tweets are annotated with event tags, and the 33 groups of data are mainly involved in 6 events, including the death of Steve Jobs, the earthquake in Japan, Li Na winning French Open tennis contest, Yao Jiaxin's murder case, bombing in Fuzhou, and the publishing of Xiaomi phones. Each of the 33 groups contains about 100 retweeting messages. The original tweet ID and corresponding number of collected retweeting messages for each group are shown in Table 4.

4. Predicting the Times of Retweeting

Given the time series distribution of top n retweets of an original tweet, we aim to predict the number of retweets in the future one month. In order to get a more accurate predicted value, we propose to fit the given time series distribution curve by a two-phase function, whose phases are divided according to the lifecycle of the original tweet.

4.1. Lifecycle of a Tweet. Every creature in the earth has its own lifecycle. We think that every tweet has its lifecycle like the creatures on the earth as well. We find that the lifecycle of a tweet plays an important role in predicting the times of retweeting. If the contents of two tweets are similar, the retweeting numbers per day of the two are nearly the same, and meanwhile their publishing time points are close, the tweet with a longer lifecycle will have a larger number of retweets. Hence, in order to predict the retweeting times more accurately, we propose the concept of the lifecycle of a tweet, that is, the time duration when a tweet can be retweeted in a large number.

We find that the lifecycle of a tweet is related to the response time of the first retweet, the importance of the content, and the interval distribution of retweets, and we will illustrate the three factors in the following part.

4.1.1. The Response Time of the First Retweet. The response time of the first retweet means the time difference between the time of the first retweet and that of the origin tweet.

Generally speaking, the faster the first retweet is posted, the more attention is paid to the original one. And the more popular the original tweet is, the more likely it will be retweeted. Thus, correspondingly, an original tweet which is retweeted in a short time may get more attention and thus have a longer lifecycle.

According to the 33 groups of retweeting records, we design a formula to calculate the score with respect to response time. We divide them into four levels according to different intervals of response time, and each level corresponds to different functions on the response time. In general, the shorter time the first retweet is posted, the higher score will the original one get. The response time in the high speed group is less than 10 seconds, and the corresponding score in this group is assigned a full score of 10 points. The response time in the 2nd group is between 10 and 100 seconds, and the range of corresponding score in this group is [6, 10] points, and the score declines with a $(\lg x)^{-1}$ speed. The response time in the 3rd group is between 100 and 10000 seconds, and the range of corresponding score in this group is [0.6, 6] points; the score declines with $x^{-1/2}$ speed. The slow ones are over 10000 seconds, some are even more than 70000 seconds, and the range of corresponding score in this group is (0, 0.6]

Table 2: The importance of content.

Content	Rank	$S_{\text{importance of content}}$
High	T3	7–9
Middle	T2	4–6
Low	T1	1–3

Table 3: The interval time distribution.

Interval time distribution	Rank	$S_{\text{interval time distribution}}$
Separate and uneven	T3	3–5
Even	T2	1–3
Grow up highly	T1	0.1-0.2

points; the score declines slower than the 3rd group with $x^{-1/4}$ speed. The score on response time is proportional to the length of its lifecycle. The score with respect to response time is shown as

$$S_{\text{response time}} = \begin{cases} 10, & 0 < x \ll 10 \\ 2 + 8 \cdot (\lg x)^{-1}, & 10 < x \leq 100 \\ 60 \cdot x^{-1/2}, & 100 < x \leq 10000 \\ 6 \cdot x^{-1/4}, & x \geq 10000. \end{cases} \quad (1)$$

4.1.2. The Importance of the Content. The vast amount of retweeting happens only when the content is attractive, which is named as the importance of content. People tend to pay more attention to those tweets with attractive contents, that is, with high grade of importance of content.

The contents of tweets involve all aspects of our lives. According to Sina microblog, tweets can be classified to the categories such as lifestyle, love, entertainment, film, television, sports, finance, science, art, fashion, culture, and media. A tweet will be retweeted by a large number of times only when there is something attractive enough in its content, such as being about a pop star's affair or some big emergency. Take some pieces of news as examples.

(1) Before the death of American singer Michael Jackson was published, there were numerous fans coming into the hospital of the University of California in Los Angeles, where Michael Jackson had been, since they got the news from Facebook and Twitter. Moreover, only one hour later after the announcement of death, there were more than 65000 reply messages and retweets in Twitter; over 5000 of them came out within one minute.

(2) In February 2010, a 93-year-old Mrs. Xiao, who was from Chengdu, needed RH-AB blood because of the fracture. Lacking blood, she was in danger at that time. In that case, her daughter came to send a tweet to ask for help. Only within 12 hours, there were more than 3000 people that helped to retweet it. Fortunately, 3 friends from the Internet donated their blood and she was saved.

To conclude the cases above, the tweet about the death of Michael Jackson received more than 65000 comments and retweets within one hour, and the tweet about seeking RH-AB blood received more than 3000 people's attention within half a day; therefore, we guess that the more attractive the content is, the more chances it would be retweeted.

But what kind of content would be attractive? We believe that if the content is related to the hot issue recently, such as Olympic Games, disaster, or a pop star's affair and big social case, it would be attractive. And moreover, if the time of the tweet issued is close to the time of the occurrence of the event, the tweet would attract much attention and the level of importance of content is high. In comparison, if the tweet is posted in a relatively long time later, or the content is attractive only to some professional people in some specific field, the level of importance of content is in the middle. Finally, if there are few people concentrating on it or the tweet is posted very long time after the event happens, the level of importance of content is low. The rank and corresponding score on the importance of content with respect to different kinds of contents are shown in Table 2. The higher the importance of content is, the more scores the tweet will get on the $S_{\text{importance of content}}$.

For instance, the case of Michel Jackson is about a pop star, and the tweet is issued on time, so that the content of tweet is very attractive, the rank is identified as T3, and the score on the importance of content $S_{\text{importance of content}}$ would be 9.

4.1.3. The Interval Time Distribution of Retweets. According to the observation of data, if the number of retweets grows up very fast, for example, the tweet is retweeted for thousands of times in a short time, the retweeting will be in saturation soon; therefore, the lifecycle of the original tweet is relatively short; if the interval time distribution curve is even, that is, the number of retweeting grows up in a peace way, the life cycle of the original tweet would be relatively long; if the distribution curve of retweets is scatter and discrete, the tweet needs more time to get saturation and the lifecycle would be very long. The rank and corresponding score on the interval time distribution with respect to different type of curve are shown in Table 3.

For detailed values, we may make judgments based on the following standards. Divide the interval time distribution of all retweets according to the time equally. (1) If the number of retweets is growing fast, appearing as a linear with high slope (over 60 degrees), or an exponential curve, as Figure 1(a) shows, the curve is of the type dense rise. In general, the score on the interval distribution for this type is [0.1, 0.2]. (2) If the growth of retweets is steady as Figure 1(b) shows, the curve is of the type general steady and the score is [1, 3]. (3) If the growth of the retweets is small and flat, as Figure 1(c) shows, the curve is of the type scatter, and the score is [3, 5]. In addition, if the number of retweets increases sharply at early stage but becomes more and more slow afterwards, which means the trend is subsequent fatigue, the rank for this type of curve is deemed as T1, and the lifecycle would not be long, so the score is set around [0.2,1]. Despite all the criteria, the accurate values need further studies. According to the above

Table 4: 33 original messages and corresponding number of collected re-tweeting messages.

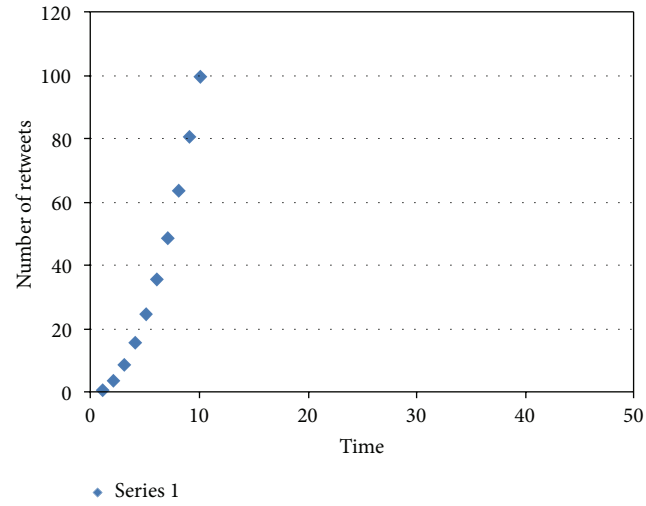
rtMid	Count(*)
2243526721410152330	100
2243578214587694822	100
2700059958269443492	99
2700117991448817596	96
2700176673306864228	100
2701374467440601577	97
2701431322360449433	100
2709258383303085289	100
2709864654666932643	100
2709870697693881414	100
2709871713230486085	100
2709893077170155796	100
51000180083282169	100
51000180083492814	100
51000180091104384	100
510001856830842390	100
510001856834367317	100
510001904903643837	100
510001908564754698	100
5100019107401880	100
55000180091534860	100
55000180527027036	100
550001906873838396	100
58000180083553705	100
8872263516485596	100
8872961090747701	100
8872983825828431	100
8872990233170214	100
8896800636296312	100
8896822338137478	100
8896858839607761	100
8896889634186199	100
8896952812610010	100

discussion, we design the rank and corresponding scores of interval time $S_{\text{interval time distribution}}$ as Table 3 shows.

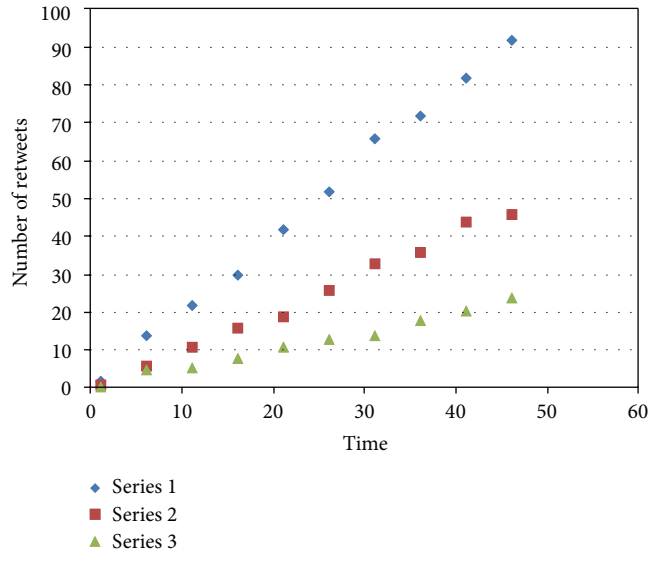
In summary, we make a calculation formula to compute the lifecycle of a tweet considering the above three factors:

$$\begin{aligned}\text{Lifecycle} &= S_{\text{interval time distribution}} \\ &\quad * \left(0.6 * S_{\text{importance of content}} + 0.4 * S_{\text{response time}}\right).\end{aligned} \tag{2}$$

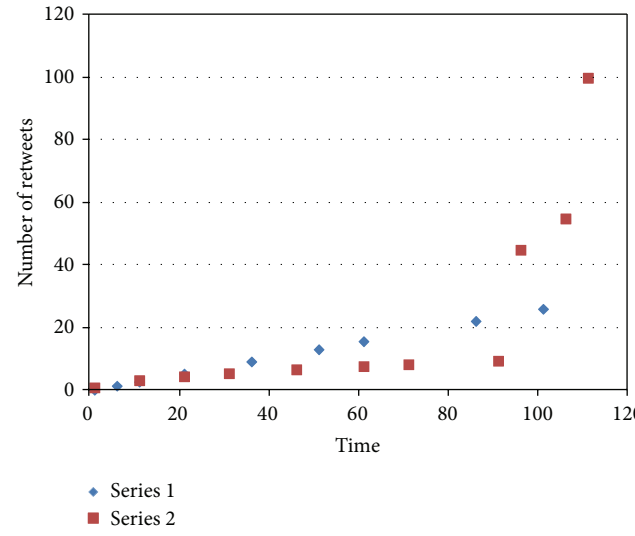
In the formula, the coefficients of the importance of content and response time are 0.6 and 0.4 separately, which are achieved by experiments on training data. The interval time distribution has a direct impact on the whole fitting of function curve, so the score on this part is worked as a product factor.



(a) Dense rise



(b) General steady



(c) Scatter

Figure 1: (a) Dense rise, (b) general steady, and (c) scatter.

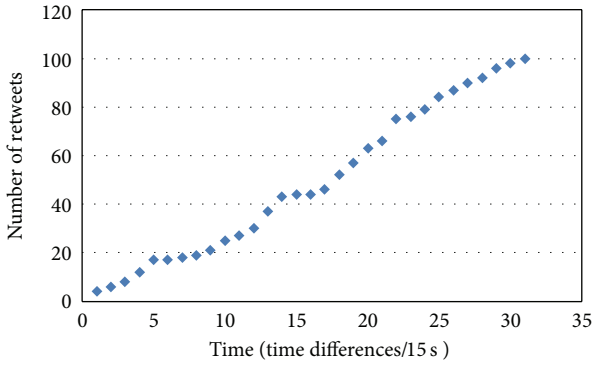


Figure 2: The time distribution scatter diagram of top 100 retweets of an original tweet which is related to the Steven Jobs' death.

Take the retweeting of an original tweet related to Steven Jobs' death issued at 12:07:52 2011/10/6 as an example. First, the event of Jobs' death belonged to the category of a star's affair, so the rank of the importance of the content is T3. Steven Jobs is the ex-CEO and one of the founders of Apple, who has a significant impact on the public, so we set $S_{\text{importance of content}}$ as 9. Second, the response time of the first retweet is 22 seconds, and according to formula 1 we have $S_{\text{response time}}$ as 8. Last, the number of retweets is increasing steady as Figure 2 shows, at the pace of 10 more retweets per minute, and the retweeting saturates within 460 seconds. The interval time distribution is like Figure 1(b), which belongs to general steady type, so $S_{\text{interval time distribution}}$ is set to 1. Therefore, the lifecycle of the original tweet is $1 * (9 * 0.6 + 8 * 0.4) = 8.6$ days.

4.2. Two-Phase Function Curve Fitting. The given time series distribution curve of top 100 retweets of an original tweet is then fitted by a two-phase function whose phases are divided according to the lifecycle of the original tweet. Main steps are illustrated as follows.

(1) We make use of Matlab, a mathematical analysis tool, for the purpose of function curve fitting. We need first to make a connection between *mysql* andMatlab and then execute sql statements through *exec* function, so as to import data from *mysql* to Matlab.

(2) Take preliminary analysis and draw scatter diagram based on the imported data. In the diagram, the x-axis data item "time" is not accurate time points but calculated by the time difference. In order to make the result more intuitive, we make the points in the scatter diagram more concentrated by dividing time slots. Figure 2 shows the time distribution scatter diagram of top 100 retweets of an original tweet which is related to Steven Jobs' death mentioned in Section 3.1.

In the following part, we will calculate the prediction value by fitting the curve with a two-phase function. In the first phase, that is, within the calculated lifecycle of the original tweet, a linear function is used to fit the curve. Most of the retweets occur within the lifecycle of the tweet, and the remainder appears as slow growing, so a logarithmic function like $a * \lg(x - b) + c$ is used to fit the curve in the 2nd phase. The detailed processes in the 3rd and 4th steps are shown as follows.

(3) In order to minimize error, we select a linear function which has the highest matching degree with the scatter points to fit the curve in the 1st phase. The line passes through as much points as possible. For every two points (x_1, y_1) and (x_2, y_2), a liner function $[(y_2 - y_1)/(x_2 - x_1)](x - x_1) + y_1$ is used to link them, and the whole curve is fitted from the relation among points. The detailed slope and intercept are decided based on the model of double moving average [25] in *Matlab*. It can avoid the lag deviation of single moving average method. The double moving average method adjusts the single one by adding a second moving average and then builds a linear model based on both average values.

The average of first moving is

$$M_t^{(1)} = \frac{1}{N}\left(y_t + y_{t-1} + \cdots + y_{t-N+1}\right). \tag{3}$$

Double moving average is making another moving average based on the first moving average, and the corresponding formula is

$$M_t^{(2)} = \frac{1}{N}\left(M_t^{(1)} + M_{t-1}^{(1)} + \cdots + M_{t-N+1}^{(1)}\right). \tag{4}$$

Since we have analyzed the growth of retweets in the 1st phase which appears as a liner function, we suppose the prediction model in the 1st phase is

$$y_{t+m} = a_t x + b_t, \quad m = 1, 2 \ldots, \tag{5}$$

in which t is the current time and m is the time slots from t to the lifecycle of the tweet; a_t is the slope and b_t is the intercept, and the two are called smooth coefficients.

According to model (5), we can have

$$\begin{aligned} a_t &= y_t, \\ y_{t-1} &= y_t - b_t, \\ y_{t-2} &= y_t - 2b_t, \\ &\vdots \\ y_{t-N+1} &= y_t - (N-1)\, b_t. \end{aligned} \tag{6}$$

So we have

$$\begin{aligned} M_t^{(1)} &= \frac{1}{N}\left(y_t + y_{t-1} + \cdots + y_{t-N+1}\right) \\ &= -\frac{y_t + \cdots + \left[y_t - (\mathrm{N}-1)\, b_t\right]}{N} \\ &= y_t - \frac{N-1}{2} b_t. \end{aligned} \tag{7}$$

Therefore,

$$y_t - M_t^{(1)} = \frac{N-1}{2} b_t. \tag{8}$$

According to model (5) and to make similar inference as (8), we can have

$$y_{t-1} - M_{t-1}^{(1)} = \frac{N-1}{2} b_t. \tag{9}$$

Therefore,

$$\begin{aligned} y_t - y_{t-1} &= M_t^{(1)} - M_{t-1}^{(1)} = b_t, \\ M_t^{(1)} - M_t^{(2)} &= \frac{N-1}{2} b_t. \end{aligned} \tag{10}$$

Then the smooth coefficients can be calculated by

$$\begin{aligned} a_t &= 2M_t^{(1)} - M_t^{(2)}, \\ b_t &= \frac{2}{N-1}\left(M_t^{(1)} - M_t^{(2)}\right). \end{aligned} \tag{11}$$

According to the fitting curve, the function value when the x-axis value reaches the lifecycle of the original tweet is the predicted number of retweets in the 1st phase. An example scatter diagram and its corresponding fitting curve in the 1st phase are shown in Figure 3.

(4) For the remaining part that is beyond the lifecycle while being within one month, a logarithm function is used to fit the curve. The coefficients in the logarithm function $a * \lg(x - b) + c$ can be achieved by fitting the scatter points, and we can get the predicted value in the 2nd phase by passing the value of rest time into the function.

Take retweeting of the original tweet about Steven Jobs' death issued at 12:07:52 2011/10/6 as an example. Its lifecycle is 8.6 days as calculated in Section 4.1. In the 1st phase, the fitted linear function is $[(y_2 - y_1)/(x_2 - x_1)](x - x_1) + y_1 = (100 - 98/31 - 30)(x - 30) + 98 = 2(x - 30) + 98$, which can be derived from Matlab. We should translate the metric from day to seconds before the following calculation; that is, 8.6 days is equal to 743040 seconds (8.6 day $*$ 86400 sec/day = 743040 seconds). As we mentioned in step 2, the accurate seconds are divided into time slots by every 15 seconds. So here x is equal to 743040/15 = 49536, and then we can get the predicted retweeting number in the 1st phase by passing the value of x into the linear function; that is, $2 * (49536 - 30) + 98 = 99110$. In the 2nd phase, the logarithm function $(a * \lg(x - b) + c)$ is used to predict the retweeting number in the remaining 21.4 days. The coefficients can be achieved directly by *Matlab*; here a is 2432, b is −714, and c is −1.599e + 004, and the value of the 2nd phase by passing x into the logarithm function is 117. Finally, the values of the two phases are summed up and the final result of the prediction on the retweeting number in 30 days is 99110 + 117 = 99227. Compared to actual retweeting number 110904, the deviation of our result is

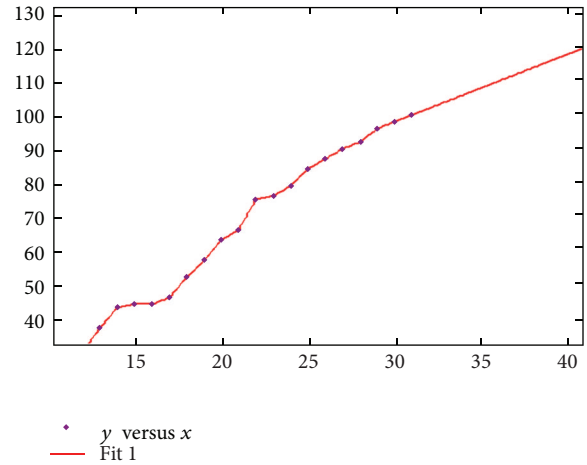


Figure 3: An example diagram and its corresponding fitting curve in the 1st phase.

$$\begin{aligned} &\frac{|\text{real number} - \text{prediction number}|}{\text{real number}} * 100\% \\ &= \frac{110904 - 99227}{110904} * 100\% = 10.52\%. \end{aligned} \tag{12}$$

5. Experiment Analysis

The result of prediction on the times of retweeting of the 33 original tweets is presented in Table 5.

In this table we can find out that the average error is less than 20%; we can conclude that our prediction is almost close to the real number of retweeting. Although different events have different lifecycle, we can get that the prediction values in the 1st phase play a dominate role, while those in the 2nd phase account for a smaller proportion.

6. Conclusions and Future Work

The prediction on the times of retweeting in microblog is to quantize the speed of information spread in microblogs and to find out the focus of public attention at all times, which is the key point of our research. In this paper, we analyze the behavior characteristics of retweeting in microblog and predict the times of retweeting of an original tweet in one month by a two-phase function curve fitting. The experiment shows that our approach can work out the prediction on retweeting times, and the average error is controlled within 20%.

Even so, our work still has some improvement to do, which is the direction in the future. First, the selected function may not be proper in some time, which leads to some exceptional results, so we may try some other function model. Second, we may do experiments on big data in order

Table 5: Result of prediction on re-tweeting of the 33 original tweets and comparison to real values.

rtmid	rttime	Event	Lifecycle	1st value	2nd value	Prediction value	Actual value	Deviation
8872263516485596	2011/10/6 19:17:17	death of Steve Jobs$mourn Steve Jobs	9.6	4969	23	4992	5587	10.65%
8872961090747701	2011/10/6 12:07:52	death of Steve Jobs$mourn Steve Jobs	8.6	99110	117	99227	110904	10.52%
8872983825828431	2011/10/6 10:01:33	death of Steve Jobs$mourn Steve Jobs	10.2	22786	84	22870	27514	16.88%
8872990233170214	2011/10/6 10:14:24	death of Steve Jobs$mourn Steve Jobs	9.8	11560	37	11597	13768	15.77%
8896800636296312	2011/8/17 9:33:03	Xiaomi release$Xiaomi	24.4	67256	42	67298	72021	6.56%
8896822338137478	2011/8/17 12:28:16	Xiaomi release$Xiaomi	24.2	484	17	501	587	14.65%
8896858839607761	2011/8/17 10:19:08	Xiaomi release$Xiaomi	25.2	40506	34	40540	47297	14.29%
8896889634186199	2011/8/17 11:04:27	Xiaomi release$Xiaomi	24.8	39810	46	39856	38017	4.84%
8896952812610010	2011/8/17 17:04:08	Xiaomi release$Xiaomi	26.5	5415	5	5420	6903	21.48%
51000180083282169	2011/3/11 15:09:44	House prices$ houseJapan Earthquake$Miyagi-ken	0.6	7468	2	7470	5972	25.08%
51000180083492814	2011/3/11 15:45:08	House prices$ houseJapan Earthquake$Miyagi-ken	0.64	4495	4	4499	5538	18.76%
51000180091104384	2011/3/11 16:31:18	Japan Earth-quake$magnitude 9.0 earthquake	1.5	9709	26	9735	11699	16.79%
55000180091534860	2011/3/11 16:31:55	Japan Earth-quake$magnitude 9.0 earthquake	0.97	14611	47	14658	16891	13.22%
55000180527027036	2011/3/12 9:19:52	Japan Earth-quake$magnitude 9.0 earthquake	1.6	6888	25	6913	8022	13.82%
58000180083553705	2011/3/11 15:08:16	Japan Earth-quake$magnitude 9.0 earthquake	0.8	25645	52	25697	30174	14.84%
5100019107401880	2011/4/1 12:56:42	Yao Jiaxin murder case$Zhang Miao	2.4	10819	77	10896	12400	12.13%
510001856830842390	2011/3/27 11:54:27	Yao Jiaxin murder case$Yao Jiaxin	8.3	4439	23	4462	4873	8.43%
510001856834367317	2011/3/27 18:55:52	Yao Jiaxin murder case$Yao Jiaxin	11.8	756	6	762	776	1.80%
510001904903643837	2011/4/19 10:19:43	Yao Jiaxin murder case$Yao Jiaxin	1.12	7244	108	7352	9779	24.82%
510001908564754698	2011/4/13 14:36:44	Yao Jiaxin murder case$Yao Jiaxin	12	33846	32	33878	36298	6.67%
550001906873838396	2011/4/17 10:40:19	Yao Jiaxin murder case$Yao Jiaxin	9.2	47524	58	47582	53385	10.87%
2243526721410152330	2011/4/22 12:18:52	Yao Jiaxin murder case$Yao Jiaxin	1.4	24181	92	24273	27906	13.02%

Table 5: Continued.

rtmid	rttime	Event	Lifecycle	1st value	2nd value	Prediction value	Actual value	Deviation
2243578214587694822	2011/4/22 12:41:36	Yao Jiaxin murder case$Yao Jiaxin	10.5	12138	41	12179	14462	15.79%
2700059958269443492	2011/5/27 4:28:31	Fuzhou bombings$Qian Mingqi$Fuzhou	1.5	2451	22	2473	2813	12.09%
2700117991448817596	2011/5/26 20:50:41	Fuzhou bombings$Qian Mingqi$Fuzhou	1.6	6919	14	6933	7979	13.11%
2700176673306864228	2011/5/27 0:48:40	Fuzhou bombings$Qian Mingqi$Fuzhou	1.62	6994	17	7011	7876	10.98%
2701374467440601577	2011/5/26 12:01:08	Fuzhou bombings$Qian Mingqi$Fuzhou	8.1	4806	15	4821	5465	11.78%
2701431322360449433	2011/5/26 19:11:24	Fuzhou bombings$Qian Mingqi$Fuzhou	1.3	8956	34	8990	10772	16.54%
2709258383303085289	2011/6/5 8:19:59	Li Na win French Open in tennis$Francesca Schiavone	10.2	1726	7	1733	1927	10.07%
2709864654666932643	2011/6/4 23:00:46	Li Na win French Open in tennis$Francesca Schiavone	0.7	36267	67	36334	43146	15.79%
2709870697693881414	2011/6/4 21:50:59	Li Na win French Open in tennis$Francesca Schiavone	0.9	116374	112	116486	136544	14.69%
2709871713230486085	2011/6/4 21:46:34	Li Na win French Open in tennis$Francesca Schiavone	1.38	39670	72	39742	48925	18.77%
2709893077170155796	2011/6/4 20:58:17	Li Na win French Open in tennis$Francesca Schiavone	1.6	3493	8	3501	3983	12.10%

to optimize and adjust the curve fitting, so as to reduce the error.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The work is supported in part by the following funds: the National Natural Science Foundation of China under the Grant no. 61202095 and 61173176 and the Scientific Research Project of Central South University under the Grant no. 7608010001.

References

[1] S. Petrovic, M. Osborne, and V. Lavrenko, "RT to win! Predicting message propagation in twitter," in *Proceedings of 5th International AAAI Conference on Weblogs and Social Media*, pp. 586–589, 2011.

[2] China Internet Network Information Center (CNNIC), *The 29th Internet Development Statistics Report in China*, 2012.

[3] "WISE 2012 challenge," http://www.wise2012.cs.ucy.ac.cy/challenge.html.

[4] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: tweets as electronic word of mouth," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 11, pp. 2169–2188, 2009.

[5] R. Long, H. F. Wang, Y. Q. Chen, O. Jin, and Y. Yu, "Towards effective event detection, tracking and summarization on microblog data," in *Web-Age Information Management*, H.

Wang, S. Li, S. Oyama, X. Hu, and T. Qian, Eds., vol. 6897 of *Lecture Notes in Computer Science*, pp. 652–663, 2011.

[6] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: what 140 characters reveal about political sentiment," in *Proceedings of 4th International AAAI Conference on Weblogs and Social Media*, pp. 178–185, 2010.

[7] J. Bollen, H. Mao, and A. Pepe, "Determining the public mood state by analysis of microblogging posts," in *Proceedings of the 12th International Conference on the Synthesis and Simulation of Living Systems*, pp. 667–668, 2010.

[8] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th International World Wide Web Conference (WWW '10)*, pp. 851–860, April 2010.

[9] Y. Qu, C. Huang, P. Zhang, and J. Zhang, "Microblogging after a major disaster in China," in *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW '11)*, pp. 25–34, March 2011.

[10] P. Achananuparp, E. P. Lim, J. Jiang, and T. A. Hoang, "Who is retweeting the tweeters? Modeling, originating, and promoting behaviors in the twitter network," *ACM Transactions on Management Information Systems*, vol. 3, no. 3, article 13, 2012.

[11] J. Tang, X. Wang, H. Gao, X. Hu, and H. Liu, "Enriching short text representation in microblog for clustering," *Frontiers of Computer Science in China*, vol. 6, no. 1, pp. 88–101, 2012.

[12] J. Chen and C. Zhang, "Research on prediction of comprehensive forwarding probability based on emotional word content, user tags, historical forward rate in MicroBlogging community," 2012, http://www.paper.edu.cn/releasepaper/content/201111-371.

[13] F. Xiong, Y. Liu, Z. J. Zhang, J. Zhu, and Y. Zhang, "An information diffusion model based on retweeting mechanism for online social media," *Physics Letters A*, vol. 376, no. 30-31, pp. 2103–2108, 2012.

[14] Y. Zhang, R. Lu, and Q. Yang, "Predicting retweeting in microblogs," *Journal of Chinese Information Processing*, vol. 26, no. 4, pp. 109–114, 2012.

[15] L. Hong, O. Dan, and B. D. Davison, "Predicting popular messages in twitter," in *Proceedings of the 20th International Conference Companion on World Wide Web (WWW '11)*, pp. 57–58, April 2011.

[16] T. R. Zaman, R. Herbrich, J. V. Gael, and D. Stern, "Predicting information spreading in twitter," in *Proceedings of the Workshop on Computational Social Science and the Wisdom of Crowds (NIPS '10)*, 2010.

[17] Y. Zhang, R. Lu, and Q. Yang, "Prediction of the micro-blog retweet behavior," in *Proceedings of the National Conference on Information Retrieval*, 2011.

[18] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: conversational aspects of retweeting on twitter," in *Proceedings of the 43rd Annual Hawaii International Conference on System Sciences (HICSS-43 '10)*, January 2010.

[19] R. Lahan, *The Economics of Attention*, University of Chicago Press, 2006.

[20] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, "Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network," in *Proceedings of the 2nd IEEE International Conference on Social Computing (SocialCom '10)*, pp. 177–184, August 2010.

[21] J. Berger and K. L. Milkman, "Social transmission, emotion, and the virality of online content," Wharton Research Paper, 2010.

[22] H. B. Zhang, Q. Zhao, H. Y. Liu, J. He, X. Y. Du, and H. Chen, "Predicting retweet behavior in weibo social network," in *Web Information Systems Engineering—WISE 2012*, X. S. Wang, I. Cruz, A. Delis, and G. Huang, Eds., vol. 7651 of *Lecture Notes in Computer Science*, pp. 737–743, 2012.

[23] S. Unankard, L. Chen, P. Li et al., "On the prediction of retweeting activities in social networks—a report on WISE 2012 challenge," in *Web Information Systems Engineering—WISE 2012*, X. S. Wang, I. Cruz, A. Delis, and G. Huang, Eds., vol. 7651 of *Lecture Notes in Computer Science*, pp. 744–754, 2012.

[24] Z. L. Luo, Y. Wang, and X. T. Wu, "Predicting retweeting behavior based on autoregressive moving average model," in *Web Information Systems Engineering—WISE 2012*, X. S. Wang, I. Cruz, A. Delis, and G. Huang, Eds., vol. 7651 of *Lecture Notes in Computer Science*, pp. 777–782, 2012.

[25] C. T. Ragsdale, *Spreadsheet Modeling and Decision Analysis*, Cengage Learning, 6th edition, 2010.