A Cluster-Based Epidemic Model for Retweeting Trend Prediction on Micro-blog

Zhuonan Feng(✉), Yiping Li, Li Jin, and Ling Feng

Department of Computer Science and Technology, Tsinghua University, Beijing, China
fzn0302@163.com
{liyp09,l-jin12}@mails.tsinghua.edu.cn, fengling@tsinghua.edu.cn

Abstract. Tweets spread on social micro-blog bears some similarity to epidemic spread. Based on the findings from a user study on tweets' short-term retweeting characteristics, we extend the classic Susceptible-Infected-Susceptible (SIS) epidemic model for tweet's retweeting trend prediction, featured by the multiple retweeting peaks, retweeting life-time, and total retweeting amount. We cluster micro-blog users with similar retweeting influence together, and train the model using the least square method on the historic retweeting datato obtain different groups' retweeting rates. We demonstrate its effectiveness on a real micro-blog platform.

Keywords: Micro-blog · Tweet · Retweet · Cluster-based epidemic model

1 Introduction

The emergence of social networks are changing people's communication styles. At the same time, the style of information propagation is undergoing profound changes. Some new media such as Twitter, Facebook, and WhatsApp are widely used and become new information carriers and communication tools of social networks. Compared with the traditional media like TV and newspaper, these new media have several significant characteristics.

1. *Different information propagation modes.* Traditional media adopt "broadcast" mode to propagate information, i.e., "one information center, thousands of listeners" in short. New media such as micro-blog is "we-media" whose information propagation is "mouth to mouth"-like, that is, everyone can become a radio station, transmitting information to his/her listeners.
2. *Different information capacity and user groups.* Traditional media can transmit massive information with a large capacity each time. But ordinary people have no chance to release news on these media. New media is lightweight and easy to publish messages. A micro-blog only requires 140 words or less. Someone who wants to announce something only needs a mobile phone and a client, typing some words and pressing "ok", a message could be sent all over the world.

Q. Chen et al. (Eds.): DEXA 2015, Part I, LNCS 9261, pp. 558–573, 2015.
DOI: 10.1007/978-3-319-22849-5_39

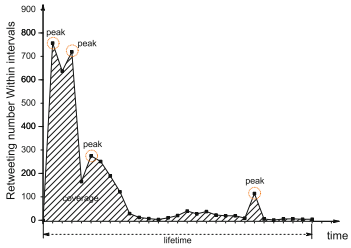


Fig. 1. Retweeting trend for a tweet posted by a real estate celebrity Ren, Z.Q

3. *Different audiences.* Traditional media look like "official voice", and are often influenced by politics. New media are often regarded as the voice of public and are very popular to young generation.

Like traditional media's (such as TV programs) rating surveys, the trend of information diffusion on social networks is also people's concern, and benefits many applications. One important field is micro-blog based advertising and marketing. With the help of microblog's trend analysis and prediction, merchants could know when, where, and who to promote their commodities to. Another important field is fake messages prevention. Network supervisors could prevent rumors diffusion just like disease prevention. For social network tool micro-blogs, the important quota to reflect information diffusion is tweets retweeting. If someone A published a tweet, his/her follower B could see the tweet and retweet it, and B's follower C could also do the same thing. So for micro-blog, information's diffusion trend could be simplified to its retweeting number's variation trend.

In the literature, [7] used a triad (Intensity, Coverage, Duration) to describe and predict information trend via a dynamic activeness (DA) model on the DBLP academic exchange platform. The work reported in this paper differs from this closely related work in the following two aspects. First, the DA model fits the relatively long-term trend of scholars' research interests, whose time granularity is about one year. But in micro-blog, the trend of information dissemination is a short-term trend. Most messages on micro-blog only last several days from their birth to the death. The time granularity is around an hour or minute. Compared with long-term trends, a tweet's diffusion trend fluctuates much drastically, and does not exhibit exponential distribution, which is the theoretical base of the DA model. Our data analysis result shows that DA model is not suitable for short-term trends. Figure 1 shows multiple retweeting peaks during the spread of a tweet posted by a real estate celebrity Ren, Z.Q. Second, based on the characteristics of short-term tweets spread obtained through a user study on real micro-blog data, we compare the similarity and difference between tweets diffusion and epidemic disease spread, and propose a cluster-based SIS (Susceptible-Infected-Susceptible) method for multiple retweeting peaks, lifetime, and coverage prediction. We compare the performance of the method with the existing ones on a micro-blog platform, and the result shows our method is effective and could achieve better performance on each prediction item.

The remainder of the paper is organized as follows. We review related work in Sect. 2. We analyze the characteristics of tweets retweeting trends in Sect. 3, and provide the problem statement in Sect. 4. Comparing the similarity and differences between tweets spread and epidemic spread in Sect. 5, we present a cluster-based SIS epidemic model for predicting tweets retweeting trends in Sect. 6, and evaluate its performance in Sect. 7. We conclude the paper in Sect. 8.

2 Related Work

2.1 Analysis and Prediction of Retweeting Behaviors

[1] examined conversional practices of retweeting in Twitter, such as how people retweet, why they retweet, and what they retweet. A number of features which may influence tweets retweetability are studied. [11] pointed out that two content features (URLs and hashtags) and two context features (number of followers/followees and the account age) have strong relationship with the retweetability of tweets, but the number of past tweets shows little correlation with the retweeting rate. [15] found that the most important features for retweeting prediction are the identity between the tweet and the retweeter. It trained a probabilistic collaborative filtering model called Matchbox to predict the probability whether a follower would retweet a tweet or not. [14] found that almost 25.5 % of tweets were retweeted from friends' blog spaces, and presented a factor graph model to predict users' retweeting behaviors. [9] used a passive-aggressive algorithm in machine learning to predict retweeting. The performance of the algorithm is dominated by such social features as number of followers and users interests, as well as some tweet-related features like hashtag, URL, trending words, and so on. Besides the features mentioned above, some researchers noticed the "celebrity effect" in social networks and tried to find influential nodes, which contribute to information's diffusion. [4] developed a decentralized version of the influential maximization problem by influencing k neighbors rather than arbitrary users in an entire network. It presented several reasonable neighbor selection schemes to find influential spreaders on twitter. [6] proved in general directed graphs finding k-effectors is a NP-hard problem. By transforming the graph to the most probable active tree, the problem could be solved optimally in polynomial time.

2.2 Retweeting Trend Prediction

[7] defined the information's diffusion trend in social networks as a triad (*intensity, coverage* and *duration*). where trend intensity is the volume of actions in general during a fixed length of time, trend coverage is the number of people taking the given action during a fixed length of time, and the trend duration is the time span that coverage is above a given threshold. Based on the three elements the author designed a Dynamic Activeness (DA) model for information's future trend prediction. Different from [7,13] predicted information diffusion from three

major aspects *speed* (how quickly a tweet will produce an offspring tweet), *scale* (number of child nodes the tweet will produce), and *range* (number of hops in the diffusion chain). Its Cox proportional hazards regression result showed that the rate with which a user is mentioned historically is a strong predictor of information diffusion in Twitter. [2] proposed two novel trend definitions called coordinated and uncoordinated trends to detect popular topics. The author also introduced a novel information diffusion model called ITFM to distinguish viral diffusion of information from diffusion through external entities such as news media. Besides above studies, there are some researches mainly about an aspect of information's diffusion trend. [5] predicted a tweet's lifespan by generating a time series based on tweet's first-hour retweeting number, and comparing it with those of historic tweets of the similar author and post time. Then top-k historic similar tweets were identified, whose mean lifespan was computed as the predicted value of the new tweet.

3 Characteristics of Tweets' Retweeting on Micro-Blog

We focus on three types of tweets which may usually incur large-scale retweeting. (a) Headline news tweets, whose contents are about breaking news around us, can be reached through various media portals. (b) Celebrities' tweets, which have great influence on social networks. The tweets of the estate agent Ren, Z.Q. belong to this category. They are often controversial and give rise to discussion. (c) Entertainment stars' tweets, like the tweets of the popular Hunan TV host Xie, N., who has over 50 million direct followers (also called fans).

To examine the characteristics of these tweets' spread on micro-blog, we study 4000 tweets and their 5 million retweeting activities on sina micro-blog (the biggest social networking platform in China with billions of users and up to 60 million active users per day[1]). Among them, 2300 are headline news tweets, 1000 are celebrities' tweets, and 500 are entertainment stars' tweets. We trace the users who retweet these tweets, and crawl about 370 thousand tweets posted by the retweeting users from November 2013 to December 2014. From the millions of retweeting records, we have some interesting findings.

(1) *Tweets spread on micro-blog through three media - direct followers, indirect followers, and external visitors.* Due to the microblog's transmission mechanism, a tweet is always firstly seen and possibly retweeted by the direct followers of its author. The followers of these retweeting users (also called indirect followers of the original author) can then see and may also retweet the same tweet to their respective followers. Through such a direct/indirect following relationship, the tweet spreads on micro-blog. Besides, as some hot tweets may be reprinted by certain portal sites and media channels, users who are not in the direct/indirect following chains could also come across the tweet and become a part of disseminators. We call them external visitors.

(2) *Direct followers constitute the main force of tweet's propagation at an early stage, while indirect followers and external visitors contribute more at the*

[1] http://blog.sina.com.

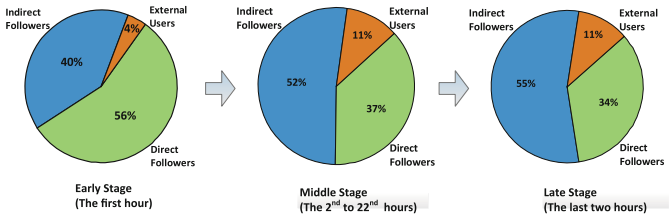


Fig. 2. Proportion of the three retweeting disseminators at different stages after a news tweet is posted

late stage. Figure 2 plots the proportion of the three retweeting disseminators at different stages since the post of a news tweet on average. The majority of retweeting happens within the next 24 h. At the beginning, direct followers are the main disseminators, indirect followers and external visitors' proportions are small. As time goes on, more and more indirect followers and external visitors take part in the propagation. Their influence becomes greater, and direct followers' participation gradually drops off. For most tweets, the proportion of external visitors fluctuates from 4 % to 11 %, except for some very hot news tweets.

(3) *Different types of tweets have different main propagators.* For headline news tweets which are externally visible, For news tweets, indirect followers and external visitors contribute a lot in the retweeting compared to other types (Fig. 3 (a)). In comparison, the spread of entertainment stars tweets relies on the direct followers of the author (Fig. 3 (b)). For celebrities who may also have important celebrity followers, both direct and indirect followers play important roles in the tweets spread (Fig. 3(c)).

(4) *If retweeting a tweet is mainly caused by direct followers, the tweet's diffusion trend exhibits an exponential distribution. Otherwise, the tweet's diffusion trend is a superposed result by direct/indirect followers and external visitors.* We make a statistics for about three thousand unnoticed tweets whose retweeting numbers are less than one thousand. We found the messages are mainly retweeted by direct followers. Proportion of external visitors is small enough that could be igored. These tweets' diffusion trends obey exponential distribution as we could see from Fig. 4 (a). The reason might be that direct followers'

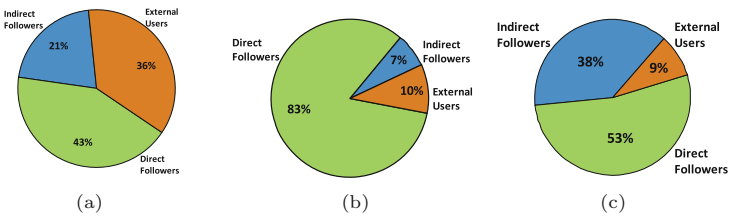
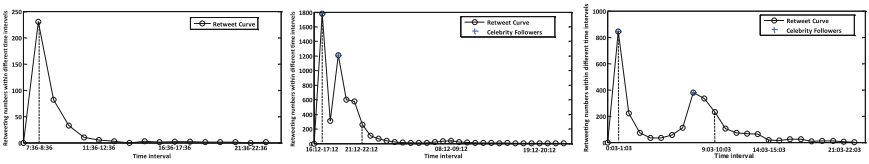


Fig. 3. Proportion of the three retweeting disseminators for (a) a news tweet on MH370 (b) an entertainment star Xie, N.'s tweet (c) a celebrity Ren, Z.Q.'s tweet

participation decays with time. However, we also observe that some notable tweets, such as MH370 whose total retweeting amount is over ten thousand, have an irregular trend rather than an exponential distribution, as many external factors may influence the propagation, not just because of time, as shown in Fig. 4 (b) and (c).

(5) *Retweeting by celebrity followers who have millions of followers will always bring a burst in the process of tweet's retweeting, as their participation could again attract a lot of respective followers to retweet.* The example tweet by Ren, Z.Q. has such a celebrity follower Pan, S.Y., whose retweeting causes another burst in the trend curve in Fig. 4 (c).



(a) (b) (c)

Fig. 4. Retweeting trend for (a) an unnoticed news tweet (b) a hot news an MH370 (c) celebrity Ren, Z.Q.'s tweet

4 Problem Definition

A tweet's retweeting trend can be characterized by multiple *peak*s, *coverage*, and *lifetime*, defined as follows. Let T=$[T.s, T.e]$ denote a time interval T, whose start time is $T.s$ and end time is $T.e$. The temporal length of T is $|T|$=$T.e$-$T.s$, which can be an hour, two hours, etc. Let $Ret(T)$ denote the retweeting number of a tweet within time interval T.

Definition 1. *Given a list of equal-length consecutive time intervals* $T_1, T_2, \cdots, T_n$, *where* $\forall i$ $(1 \leq i \leq n-1)$ $(T_i.e$=$T_{i+1}.s)$ *and* $(|T_1|$=$\cdots$=$|T_n|)$. $T_1, T_2, \cdots, T_n$ *constitute* ***a complete retweeting period of a tweet****, if and only if* $T_1.s$ *is the post time of the tweet, and for the last* l *consecutive time intervals, the retweeting number remains consistently lower than the minimal retweeting amount times a coefficient, i.e.,* $Ret(T_{n-l}), Ret(T_{n-l+1}), \cdots, Ret(T_n) < g \cdot min_{1 \leq i \leq n-l-1}\ Ret(T_i)$, *where* $(1 \leq l \leq n-1) \wedge (0 < g \leq 1)$. *In this study,* $l = 4$ *and* $g = 0.1$. □

Definition 2. *Let* $T_1, T_2, \cdots, T_n$ *be a complete retweeting period of a tweet.*
(1) The retweeting ***peaks*** *are a set of pairs* $\{(T_i, Ret(T_i) \mid (1 < i < n) \wedge (Ret(T_{i-1}), Ret(T_{i+1}) < Ret(T_i)) \wedge (Ret(T_i) > p \cdot max_{1 \leq j \leq n}\ Ret(T_j))\}$, *where* $(0 < p \leq 1)$. *In this study,* p=0.5.
(2) The retweeting ***lifetime*** *of the tweet is* $n \cdot |T_1|$.
(3) The retweeting ***coverage*** *of the tweet is* $\sum_{i=1}^{n} Ret(T_i)$. □

Figure 1 illustrates the multiple peaks, lifetime, and coverage in a tweet's retweeting trend.

Given a tweet, the problem of its retweeting trend prediction is to predict the tweet's multiple peaks, lifetime, and coverage after its first launch on micro-blog.

5 Analogy Between Tweets Spread and Epidemic Spread

5.1 Subjects

A classic epidemic model for the spread of infectious diseases is the Susceptible-Infected-Susceptible (SIS) model [3]. It divides the population into two counterparts: (1) $S(t)$: the susceptible counterpart at time t, and (2) $I(t)$: the infectious counterpart at time t. A susceptible person may become infectious by contacting with the infectious people. An infections person may also be cured without immunity as a susceptible one and may become infectious again, such as influenza and enteritis. Let β be the infectious rate, and let α be the recovery rate in unit time. The SIS model has the following expressions.

$$\begin{aligned} &S(t+1) - S(t) = -\beta \cdot S(t) \cdot I(t) + \alpha \cdot I(t); \\ &I(t+1) - I(t) = \beta \cdot S(t) \cdot I(t) - \alpha \cdot I(t); \\ &N(t) = S(t) + I(t) \equiv K. \end{aligned} \tag{1}$$

Here K is the total population remaining unchanged.

If we view a tweet message as an infectious disease, action "*retweet*" as "infect", then *retweeter*s are like infectors. In the SIS model, becoming a member of susceptible crowd $S(t)$ is subject to the following conditions: (1) A susceptible person has the chance of contacting infectors; (2) A susceptible person has poor immunity to this kind of disease; (3) An infectious person can recover later and become a susceptible one, possibly being infected again.

Accordingly, we could determine the susceptible crowd of tweets propagation on micro-blog: (1) A susceptible person should be a direct follower of the tweet's author or its retweeter, and should be interested in the tweet; (2) An infectious person who has retweeted the tweet may possibly retweet it again, just like a patient's repeated infection. So some infectors may become susceptible again.

5.2 Influence Factors

The reason why a disease becomes an epidemic has two factors: the gene of disease and the infectivity of infectors. If a disease contains a highly pathogenic gene such as H5N1, the disease will be aggressive and easy to spread. If a virus carrier is active and has a lot of contact with others, s/he will increase the disease's spread. Tweets spread in social network obeys the same principle. Two aspepcts contribute to a tweet's propagation: *tweet's gene* and *infector's gene*.

Tweet's Gene. After data analysis, we observe some important features closely relate to the tweet's heat.

- *Hashtag.* Hashtag is the symbol following "#" in a tweet. The content between two "#"s is a topic. If the topic is popular and hot, the tweet will most probably get more attentions.
- *URL.* URL is the link in a tweet. The URL may arouse followers' interest to the unknown content.
- *Multimedia.* Compared with shriveled words, pictures and videos are visual and easy to attract followers' attentions.
- *Post-time.* Tweets published at daytime will be retweeted more than those at wee hours.
- *Content-keywords.* The most important thing determining a tweet's influence is its content. Our experiments show if a tweet contains hot keywords, it will most probably get popular.

Retweeter's Gene. Retweeter's gene includes tweet's author and tweet's retweeters who are retweeting infectors. The tweet propagation from its author to the author's direct followers is viewed as a kind of retweeting action by the author. Different types of users have different influence power, and celebrity's influence is obviously larger than ordinary users'. Below is a list of features about an infector's influence.

- *Number of direct followers.* A retweeter with a large amount of direct followers is always popular in social network.
- *Number of being mentioned.* If retweeter is frequently mentioned, s/he appears to be more popular and active than those less mentioned.
- *Account verification.* A retweeter verified by microblog authentication mechanism is trustworthy and is well received by the public.
- *Account age.* A senior (re)tweeter, who is on micro-blog for a long time, gains lots of popularity and is thus easy to be focused.
- *Retweeting stimulation ratio.* Assume a direct follower A retweets n number of tweets by the same author. Via A, m number of tweets are retweeted by A's direct followers. The retweeting stimulation ratio is the ratio between m and n, i.e., m/n.

5.3 Spread Mechanisms

Differences exist between tweet's retweeting and disease spread mechanisms.

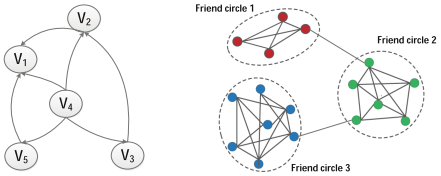


Fig. 5. Strongly-connected graph vs. social network

(1) *The classic SIS epidemic model is based on a strongly connected graph in theory, but social network micro-blog is a weakly connected graph.* In SIS, every infectious person has the chance of infecting others. While on microblog, everyone has his/her own friend circle (Fig. 5). People in different circles may not contact each other, and thus have no chance to infect the other sides. Moreover, in social network, information's transmission is often unidirectional and irreversible. This does not meet the strongly connected graph condition.

(2) *The SIS model assumes the size of the population is a constant, but in social network this is almost impossible.* SIS requires a closed counterpart $N(t)$. However, due to the social network's openness (external visitors), finding a completely closed environment is impossible. Direct application of the SIS model to the social network field is not appropriate.

(3) *In SIS, the infectious rate λ is the same for all infectors. This is not the case for tweet's retweeting: some users have more direct/indirect followers, and thus higher infectious rates than those with less followers.*

6 A Cluster-Based SIS Model for Retweeting Trend Prediction

6.1 The Model

Based on the above analysis, we extend the SIS model for tweets retweeting trend prediction. Each counterpart in the epidemic model is divided into several friend circles. Although the entire social contact graph is not fully connected, each friend circle still meets the connection condition. To highlight celebrity effect of social network, people with different influence power have different infectious rates. We cluster infectious crowd $I(T)$ into small groups according to infectors' infectious rates, i.e., influence power. We also consider the influence of external retweeting visitors. Given a tweet message at time t, we use $I(T)$, $S(T)$, and $E(T)$ to denote the number of retweeters, number of direct followers of the retweeters during time interval T, and number of external retweeting visitors, respectively. Once an external visitor retweets the tweet, s/he becomes a member of the retweeting crowd $I(t)$. Initially, $I(T_1)$ is 1, signifying the tweet's author, $S(T_1)$ is the number of author's direct followers, and $E(T_1)$ is 0.

$$\begin{aligned} S(T_{i+1}) - S(T_i) &= -\sum_{i=1}^{\#c} \beta_i \cdot S_i(T_i) + \alpha \cdot I(T_i) + \gamma \cdot I(T_i); \\ I(T_{i+1}) - I(T_i) &= \sum_{i=1}^{\#c} \beta_i \cdot S_i(T_i) + E(T_{i+1}) - E(T_i); \\ E(T_{i+1}) - E(T_i) &= \omega(T_i) \cdot (I(T_{i+1}) - I(T_i)). \end{aligned} \tag{2}$$

Here, $\#c$ is the number of clusters, β_i is the infection rate of the i-th cluster crowd, α is the infectors' reinfection (multiple retweeting) rate, γ is the direct fan-out of the retweeters, and $\omega(T_i)$ is the ratio of the increase of external retweeting visitors versus the increase of the infectious crowd $I(T_i)$ from time interval T_i to T_{i+1}.

6.2 Clustering of Infectious and Susceptible Crowds

Clustering infectious and susceptible crowds into $\#c$ groups is based on micro-blog users' influence power (i.e., infectious rates), determined by tweets' and retweeters' genes, described in Sect. 5. It proceeds in the following three steps.

Step-1:Tweets Clustering. A tweet's gene has two types of attributes: textual attribute (*content-keywords*) and numerical attributes (*hashtag, URL, multimedia, post-time*). As tweets on different topics attract different users, for example, some may be interested in politics, while some in entertainment news, and the two infectious/susceptible crowds are independent and have minor intersection. Hence, we first cluster tweets into groups based on their textual contents. We divide tweeter's total tweets by day. On each day's tweets, we extract keywords from their textual contents, and gather similar tweets together by the similarity of their keywords. We merge and sort these similar keywords by their $tf * idf$ values, and take the top-k keywords as the cluster's topic keywords. We gather different days' tweets together by the similarity of their top-k topic keywords, and combine the keywords of similar topics. The process repeats until the clustering result is stable. The advantage of this textual cluster approach is its simplicity and high efficiency compared with classic topic model LDA. It can bring the tweets with the same themes together, such as the air crashes in 2014: MH370's missing, MH17's being shot down, and QZ8501's fatal accident.

After content-based clustering, we further split tweets within one group into several sub-groups using a multi-dimensional KNN (K-Nearest Neighbor) method based on their numerical attributes values. Here, we assign value 1 or 0 to the attribute *hashtag, URL*, and *multimedia* according to their existence in the tweet. Attribute *post-time* is mapped to 0 if the tweet was posted at midnight when most micro-blog users are inactive, and 1 otherwise.

Step-2: Retweeters (Infectors) Clustering. From each tweets cluster obtained after Step-1, we can obtain all the tweets' retweeters (infectors). Based on these retweeters' genes (*number of direct followers, number of being mentioned, account verification, account age,* and *retweeting stimulation ratio*), we cluster these retweeters into several groups via the multi-dimensional KNN method.

Step-3: Susceptible Crowd Clustering. Each retweeters (infectious) cluster leads to a corresponding susceptible crowd cluster, whose members are the direct followers of at lease one retweeter in the former cluster, and are interested in the tweets by either posting/retweeting similar tweets before, or having similar self-description as the tweeter.

6.3 Predicting a Tweet's Retweeting Trend

After splitting the susceptible population into $\#c$ clusters for a set of tweets in each category (headline news, celebrity, or entertainment stars), we can apply the least square method to learn parameters $\beta_1, \cdots, \beta_{\#c}, \alpha, \gamma$ in Formula 2. Let Γ be the set of tweets belonging to the same tweet category. Let $T_1, T_2, \cdots, T_n$

be the complete retweeting period of the tweets in Γ, starting from the birth to the end. For a tweet $\tau \in \Gamma$, the prediction error between the real values $S(T_{i+1})$, $I(T_{i+1})$, $E(T_{i+1})$ and the predicted values $S^*(T_{i+1})$, $I^*(T_{i+1})$, $E^*(T_{i+1})$ at time interval T_{i+1} derived from Formula 2 can be expressed through function

$$f = \sum_{\tau \in \Gamma} \sum_{i=1}^{n} [\ (S(T_{i+1}) - S^*(T_{i+1}))^2 + (I(T_{i+1}) - I^*(T_{i+1}))^2 + (E(T_{i+1}) - E^*(T_{i+1}))^2]$$

where

$$S^*(T_{i+1}) = S(T_i) - \sum_{i=1}^{\#c} \beta_i \cdot S_i(T_i) + \alpha \cdot I(T_i) + \gamma \cdot I(T_i);$$
$$I^*(T_{i+1}) = I(T_i) + \sum_{i=1}^{\#c} \beta_i \cdot S_i(T_i) + E(T_{i+1}) - E(T_i);$$
$$E^*(T_{i+1}) = E(T_i) + \omega(T_i) \cdot (I(T_{i+1}) - I(T_i)).$$

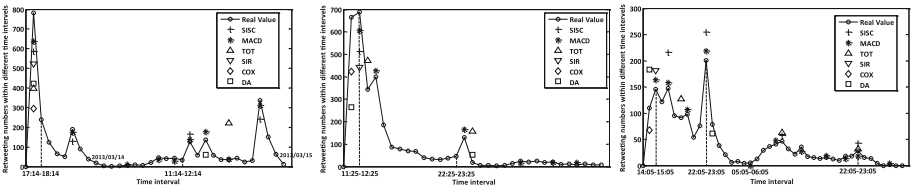
To minimize the prediction error $min\ f$, we compute f's partial derivatives $\partial f/\partial \beta_i = 0\ (i = 1, \cdots, \#c)$, $\partial f/\partial \alpha = 0$, $\partial f/\partial \gamma = 0$, $\partial f/\partial \omega = 0$, and obtain the values of parameters $\beta_1 \beta_2 \cdots \beta_{\#c}$, α, γ in Formula 2. We can then estimate the retweeting amounts of a coming tweet at different future time intervals using the materialized Formula 2 obtained in the corresponding category, and finally derive the retweeting peaks, lifetime, and coverage of the tweet according to definition 2. The detail procedure is shown as follows:

1. Initialize the variables of $peaks, coverage$ and $lifetime$. Here $coverage = 0$, $peaks$ is an empty list and $lifetime$ is the tweets start time.
2. Set threshold $\theta_1 = I(T_1), \theta_2 = I(T_1).counter = 0$.
3. While(true)
   i get retweeting number $I(T_i)$ at T_i, $coverage = coverage + I(Ti)$.
   ii Compute $I(T_{i+1})$ with Formula 2, and get retweeting number $I(T_{i-1})$, at T_{i-1}. If $I(T_i) > I(T_{i+1})$ and $I(T_i) < I(T_{i-1})$ and $I(T_i) > 0.5 * \theta_1$, add$(T_i, I(T_i))$ into peaks, update $\theta_1 = I(T_i)$.
   iii If $I(T_i) < 0.1 * \theta_2$, Set $counter = counter + 1$.Else if $I(T_i) < \theta_2$, Update $\theta_2 = I(T_i)$.
   iv when $counter = 4; lifetime = T_i - T_0$ break.
4. End while
5. Get current $peaks, coverage, lifetime$ as the final result.

7 Evaluation

7.1 Set-Up

We crawl three kinds of tweets from sina micro-blog, including 2300 headline news tweets, 1000 celebrities' tweets, and 500 entertainment stars' tweets. We randomly pick up 100 tweets from each category as the test data, use the rest to train and obtain three cluster-based SIS models, each corresponding to one tweet category. The prediction performance is measured by MAPE (Mean Absolute Percentage Error), defined as $MAPE = \sum_{\tau \in \Gamma} \frac{|RealVal - PredictVal|}{RealVal} / |\Gamma|$. Besides, we use recall, precision, and F1-measure to measure multiple peaks prediction results. We compare our cluster-based SIS prediction model (SISC) with another



(a) (b) (c)

Fig. 6. Retweeting peaks prediction for (a) a Headline news (b) a celebrity Ren, Z.Q.'s tweet (c) an entertainment star Xie, N.'s tweet

four information diffusion models, which can be used or extended to resolve the retweeting trend prediction problem.

- DA (Dynamic Activeness) model [7] predicts information propagation trends (intensity, coverage, and duration) on a DBLP platform based on the concept of node activeness. It uses the law that the decrease of activeness roughly obeys exponential distribution to predict a tweet's future trend.
- Cox proportional hazards regression model [13] is used to predict whether and when a tweet produces its first offspring node based on the features of users and tweets. With this model, people could predict the speed, scale, and range of information's diffusion trends.
- MACD (Moving Average Convergence-Divergence) model [8], which predicts topic trends on Twitter based on the deviation between short-term moving average and long-term moving average. When the short term average is greater than the long term average, the trend will be upward. Conversely, the trend will be downward.
- TOT (Topics over Time) model [12] is a time dependent topic model based on LDA. It models topic distribution conditioned on time stamps. Then the model could determine the topic's future trend according to the distribution.
- SIR model [10] determines threshold conditions for arbitrary cascading models on arbitrary networks. It utilizes the balance point of infectors and rehabilitees to get the threshold of a tweet's diffusion, which is the sign of peak's coming. But the condition of the threshold's existence requires a closed environment, where the total amount of susceptibles and infectors is a constant.

7.2 Performance of Multiple Retweeting Peaks Prediction

Test 1: We randomly pick a tweet from each category, and examine their retweeting peaks (*peak time interval, peak value*). In Fig. 6(a), among the five peaks occuring on a news tweet's retweeting curve, SISC can predict four peaks of them. However, COX and SIR only can predict the first peak at 2014/3/13 17:24 and miss the other four. Although MACD predicts the five peaks, some non-peak points are also involved in the final results. DA and TOT just predict

Table 1. Average Peak Prediction Performance

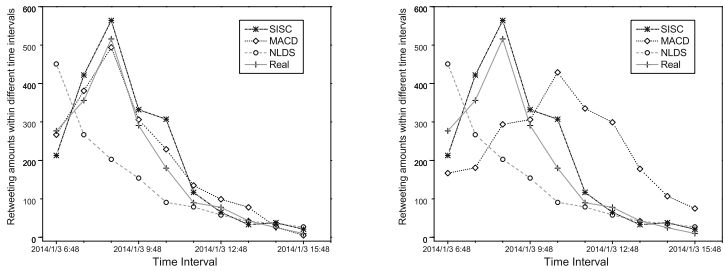
Tweet Category	Method \ Metric	Recall	Precision	F1-Measure	Peak Value MAPE	Peak Time Interval
News	SISC	0.61	0.63	0.62	0.38	(|:0.57; +:0.23; -:0.20)
	MACD	0.73	0.31	0.44	0.35	(|:0.64; +:0.27; -:0.09)
	DA	0.32	0.61	0.42	0.46	(|:0.28; +:0.68; -:0.04)
	TOT	0.34	0.58	0.43	-	(|:0.23; +:0.70; -:0.07)
	SIR	0.21	0.43	0.28	0.52	(|:0.11; +:0.89; -:0.00)
	COX	0.17	0.36	0.23	0.65	(|:0.11; +:0.89; -:0.00)
Celebrity	SISC	0.63	0.66	0.64	0.42	(|:0.61; +:0.33; -:0.06)
	MACD	0.68	0.27	0.39	0.31	(|:0.63; +:0.23; -:0.14)
	DA	0.36	0.54	0.43	0.63	(|:0.22; +:0.68; -:0.10)
	TOT	0.38	0.58	0.46	-	(|:0.20; +:0.69; -:0.11)
	SIR	0.31	0.51	0.39	0.72	(|:0.16; +:0.84; -:0.00)
	COX	0.22	0.32	0.26	0.68	(|:0.14; +:0.86; -:0.00)
Entertainment Star	SISC	0.71	0.78	0.74	0.33	(|:0.68; +:0.20; -:0.12)
	MACD	0.75	0.41	0.53	0.35	(|:0.64; +:0.25; -:0.11)
	DA	0.62	0.67	0.64	0.41	(|:0.59; +:0.20; -:0.21)
	TOT	0.64	0.68	0.66	-	(|:0.56; +:0.32; -:0.12)
	SIR	0.71	0.78	0.74	0.36	(|:0.64; +:0.36; -:0.00)
	COX	0.67	0.55	0.60	0.55	(|:0.61; +:0.39; -:0.00)

|,+,-: the predicted peak time is equal to/ahead of/behind the real value respectively; - inapplicable

the first and the last one. As the time-dependent TOT model can not predict the precise peak values, we position its prediction results randomly in the figure. In general, SISC has a better performance than other algorithms on a tweet's peaks prediction. The results in Fig. 6(b) and (c) are similar to the one in Fig. 6(a).

Test 2: While Fig. 6 only shows the retweeting trend for an individual tweet, we examine the average performance of the methods on 100 tweets per category. We have two observations from the results in Table 1.

First, for the tweets of news and celebrity, the recall, precision, F1-measure, and MAPE values of our SISC method are (61 %, 63 %, 62 %, 38 %), each of which is better than (32 %, 61 %, 42 %, 46 %) of the closely related DA method. COX's performance is even worse than DA. The reason is that DA and COX are both based on simple mathematical assumptions (DA is based on exponential distribution and COX is based on normal distribution). They fit for long-term trend prediction, but ignore the influence of external factors. The extended MACD exhibits the best recall (73 %) and MAPE (35 %) in peaks prediction with the lowest precision (31 %) among the six methods. MACD's difficulty is the choice of time interval. Shorter time interval may make the deviation between the short and long term averages fluctuate frequently and produce many redundant points, thus reducing the prediction precision of MACD. The performance of the extended TOT and SIR in peaks prediction is not good. Because SIR requires a closed compartmental environment, its threshold determining algorithm can only predict the propagation mainly caused by direct followers, and ignore the burst caused by external factors. TOT suits topic-level trend prediction according to a certain distribution, but for fine-grained tweet-level trend prediction, it is insufficient. Our SISC method's recall and MAPE are close to those of MACD, but with a much better F1-measures.

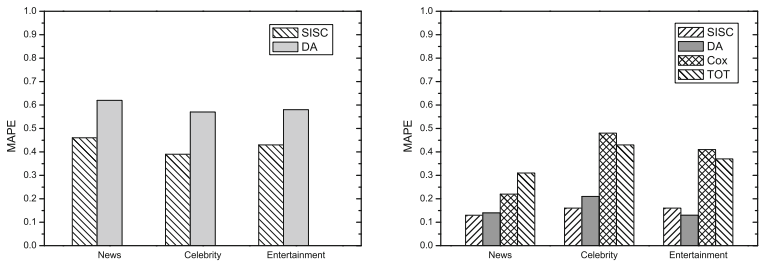


(a) Short time interval (10min) (b) Long time interval (1hr)

Fig. 7. Peak values prediction with different time interval lengths

Second, for the tweets of entertainment star, the recall, precision, F1-measure, and MAPE values between SISC (71 %, 78 %, 74 %, 33 %) and DA (62 %, 67 %, 64 %, 41 %) are significantly reduced. The differences between SISC and other algorithms show the similar phenomena. The reason is the entertainment star's tweets are mainly retweeted by his/her direct fans, indirect fans and exteranl fans seldom participate. The environment of information's diffusion is nearly closed and the tweet's retweeting trend approximates to exponential distribution, which is fit for the application condition of SIR and DA. The situation is similar to COX and TOT. So for this type of tweets, SISC's performance approaches to other algorithms.

Test 3: To illustrate the limitation of MACD, we set different time intervals for MACD model. Figure 7(a) is the predicted result of short time interval (10 min), and Fig. 7(b) is the predicted result of long time interval (1 h). In Fig. 7(a), we can see the peak values of MACD approximate to real values. While in Fig. 7(b), the performance of MACD is not so good, where the predicted result of peak values is far from the real one. Yet other algorithms' predicted results remain unchanged. This test proves that selecting a appropriate time interval is sensitive to the prediction quality of MACD.



(a) Coverage prediction (b) Lifetime prediction

Fig. 8. Average performance on three kinds of tweets

7.3 Performance of Retweeting Coverage Prediction

Coverage prediction can only be done by SISC and DA. Figure 8(a) compares the average MAPE on retweeting coverage prediction between SISC and DA, where the former has a better performance on average than the latter. Because DA makes prediction based on an exponential distribution of retweeting amounts. In most cases, the real data does not obey the distribution strictly.

7.4 Performance of Retweeting Lifetime Prediction

Among the six methods, only SISC, DA, COX, and TOT can predict the retweeting lifetime of a tweet. From Fig. 8(b), we can see that SISC and DA perform similarly in predicting tweet's retweeting lifetime on average. The reason is that the trend's termination conditions are similar for both DA and SISC. Compared to the data-driven methods like SISC and DA, the rough distribution (TOT:beta distribution, COX:log-normal distribution) based COX and TOT methods perform worse than SISC and DA, as real-time result is always more reliable than estimated one.

8 Conclusion

In this paper, we draw inspirations from the epidemic dynamic models developed in the medical field to predict retweeting trends on micro-blog. We extend the classic epidemic model (SIS) by clustering micro-blog users with similar retweeting rates together. Our performance study showed that the extended epidemic model is quite effective compared with the existing methods. This model provides a new way for information's diffusion trend analysis on social networks.

Acknowledgement. The work is supported by National Natural Science Foundation of China (61373022, 61073004), and Chinese Major State Basic Research Development 973 Program (2011CB302203-2).

References

1. Boyd, D., Golder, S., Lotan, G.: Tweet, tweet, retweet: conversational aspects of retweeting on twitter. In: Proceedings of HICSS (2010)
2. Budak, C., Agrawal, D., Abbadi, A.: Structural trend analysis for online social networks. In: Proceedings of VLDB (2011)
3. Hethcote, H.: A thousand and one epidemic models. In: Levin, S.A. (ed.) Frontiers in Mathematical Biology. Lecture notes in Biomathematics, vol. 100, pp. 504–515. Springer, Heidelberg (1984)
4. Kim, H., Wang, K., Yoneki, E.: Finding influential neighbors to maximize informationdiffusion in twitter. In: Proceedings of IW3C2 (2014)
5. Kong, S., Feng, L., Sun, G., Luo, K.: Predicting lifespans of popular tweets in microblog. In: Proceedings of SIGIR (2012)
6. Lappas, T., Terzi, E.: Finding effectors in social networks. In: Proceedings of KDD (2010)

7. Lin, S., Kong, X., Yu, P.: Predicting trends in social networks via dynamic activeness model. In: Proceedings of CIKM (2013)
8. Lu, R., Yang, Q.: Trend analysis of news topics on twitter. In: Proceedings of Machine Learning and Computing (2012)
9. Petrović, S., Osborne, M., Lavrenko, V.: RT to win! predicting message propagation in twitter. In: Proceedings of AAAI (2010)
10. Prakash, B.A., Chakrabarti, D., Faloutsos, M., Valler, N., Faloutsos, C.: Threshold conditions for arbitrary cascade models on arbitrary networks. In: Proceedings of ICDM (2011)
11. Suh, B., Hong, L., Pirolli, P., Chi, E.H.: Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In: Proceedings of SocialCom (2010)
12. Wang, X., McCallum, A.: Topics over time: a non-Markov continuous-time model of topical trends. In: Proceedings of KDD (2006)
13. Yang, J., Counts, S.: Predicting the speed, scale, and range of information diffusion in twitter. In: Proceedings of AAAI (2010)
14. Yang, Z., Guo, J., Tang, J., Li, J., Zhang, L., Su, Z.: Understanding retweeting behaviors in social networks. In: Proceedings of CIKM (2010)
15. Zaman, T., Herbrich, R., van Gael, J., Stern, D.: Predicting information spreading in twitter. In: Proceedings of NIPS Workshop (2010)