Social Media Communication Model Research Bases on Sina-weibo

Ming Wu, Jun Guo, Chuang Zhang, and Jianjun Xie

Pattern Recognition and Intelligent System Lab
Beijing University of Posts and Telecommunications, Beijing China
wuming@sina.com, zhangchuang@bupt.edu.cn

Abstract. The popularity of microblog brings new characters to information diffusion in social networks. Facing new challenges of understanding information propagation in microblog, the framework of information producing and receiving was proposed. A general model named competing-window is also presented based on human behavior. The detailed composition of the model and its basal mathematical description are also given. In addition, a parameter called information lost as a supplement to measure dynamics of information diffusion. Meanwhile, the further application of our model to information processing and propagating was pointed out. All those work is based on the studies on human dynamics. Finally, to verify applicability, the model was applied to empirical data crawled from Sina-weibo. The interesting patterns extracted from empirical data indicate that microblog in deed is fundamentally characterized by human dynamics.

Keywords: communication model, social media, human behavior, competing window, microblog, Sina-weibo.

1 Introduction

Sina-weibo is the most popular microblog service in China. After its launch on July 2009, Sina-weibo users have increased rapidly. They are currently estimated as two thousand million users worldwide. Like twitter, Sina-weibo is an online social network used by millions of people, but which only shares Chinese information around the world to remain socially connected to their friends, family members and coworkers through their computers and mobile handset.

Microblog has many new characteristics. In traditional communicating ways, such as letter, email, phone calls and short messages, users are almost equal in message publishing and replying. Thus communications are mainly peer-to-peer (P2P). Since it would take efforts to deliver messages, each message has its own destination. For example, you would surely not text all your friends what you have visited through short messages. Instead, you may publish a tweet (the common form of messages in microblog) wishing some of your friends would read it. As to the receiving end, letter or email readers may often be expected to reply in some time. However, tweets readers may choose to ignore any message and reply nothing at all. User relations in

Y. Wang and T. Li (Eds.): Knowledge Engineering and Management, AISC 123, pp. 445–454.
springerlink.com

microblog are mostly asymmetric, where a user can have many people following without a need for reciprocity. As the impact of asymmetric relations, microblog is a broadcast communication medium where information dissemination is in large scale involving multi-node interactions. Users have full autonomy to decide or choose how to behave rather than being forced to act. The instantly updated contents are pushed to related users, which advanced the ease of information publishing and disseminating. User relations mainly are asymmetric, namely Asymmetric Follow.

Previous studies of social networks paid intensive attention to structure-based research [1] [2] in network evolution [3] [4], information diffusion [5] [6] and data mining [7] [8]. The studies on communications, e.g. letter and email, short message, mobile phone calls, web information access, blog posting and other social networks, have shown that human activities have the characteristics of non-Poisson distribution (mostly Power Law distribution) with heavy tails. However, microblog has not been covered until recent researches on Twitter, Facebook, etc., including network property analysis [9], prediction of information diffusion [10] [11] and spam detection [12]. However, behaviors of communication in microblog haven't been covered yet.

In this paper, after review previous researches on human dynamics, we model basic user behaviors including tweets publishing, browsing, replying and retweeting. By introducing interest-driven hypothesis, we explain the process of broadcast communication in microblog, which provides a possible explanation to the origin of heavy-tailed Power Law distribution in collective communicating behaviors. Finally, as verifications to the model, empirical statistics are presented.

2 Related Studies

Traditionally, human actions are modeled as Poisson process [13] [14], where events independently occur at a constant rate. Thus the time interval of two consecutive events obeys negative exponential distribution $P(\tau) = \lambda e^{-\lambda\tau}$. Recent studies have shown that human activities are non-Poisson in various fields [1]–[8], where human activities are characterized by bursts of rapidly occurring events separated by long periods of inactivity. Interevent time obeys heavy-tailed power law distribution [15]. Several models were proposed to explain the origin of bursts and heavy tails in human dynamics. Priority-queuing model shows the burst nature of human activities is a consequence of a decision-based queuing process [1] [16] [17]. When individuals execute tasks based on some perceived priority, the timing of the tasks will be heavy tailed. Most tasks being rapidly executed, whereas a few experience very long waiting times. In contrast, priority blind execution is well approximated by uniform interevent statistics. Further development of this model introduced limitations and variations to the queue length [16].

Interest adjustment model [18] [19] was based on interest variation. Facing a new thing, people often show strong interests. As time goes by and repetition of actions, the interests would gradually descend and finally disappear. Activities would then stop. But after a period of idleness, interests would recover and drive activities again. The automatic adjustment mechanism of interests will produce heavy tails of human behaviors in interevent time distribution. Other models, such as Poisson processes modulated by circadian and weekly cycles [20] [21], preferential linking [18] and

memory-based activity adjustment [22] also give possible explanations to bursts and heavy tails in human activities. Previous studies have mainly focused on separated individuals and the communication is P2P model. However, study of human dynamics in Web2.0 instant broadcast medium, such as microblog, is still insufficient.

3 Communication Models of Social Media

3.1 The Framework of Microblog Communication Model

Basically, there are two kinds of relations: *unidirectional* and *bidirectional* relation. See Fig 1. If user *A* follows user *B*, user *A* is called the follower of *B*. Here the word "follow" means user *A* subscribes to user *B* and *A* will receive tweets from *B*.

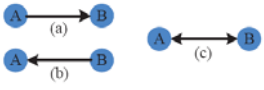

Fig. 1. User relations in microblog: (a) unidirectional relation. User $\mathcal{A}$ is following user $\mathcal{B}$ and $\mathcal{B}$ is followed by $\mathcal{A}$. $\mathcal{A}$ will receive tweets from $\mathcal{B}$, but the inverse is not. (b) $\mathcal{A}$ is followed by $\mathcal{B}$. (c) bidirectional relation. $\mathcal{A}$ and $\mathcal{B}$ are following each other and receive tweets from each other.

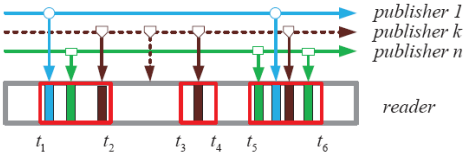

Fig. 2. Visualization of Competing-window model. The specific reader has *n* followers labeled as *publisher 1*, *publisher 2* to *publisher n*. Through the time line, each publisher independently publishes tweets which are instantly pushed to the reader (denoted by down arrows). In microblogs, all tweets are received but wheather to be read completely denpends on the readers. The time period of reading tweets is called time window (denoted by (t_1, t_2), (t_3, t_4) and (t_5, t_6)).

Note that publisher and reader relation is relative. We focus on unidirectional relation when studying information diffusion and on bidirectional relation when comments, replies and retweets are concerned.

Now we can isolate one specific reader and all of his or her friends to get a clear observation of information producing and receiving, which is also the micro node of information dissemination. The whole process is visualized by Fig 2.

From the reader's perspective, individual publishers form relations of competing without even noticing that themselves. The general picture of competing process can be literally described as: information produced by publishers, stretching out on the time line like a stream, is crowding into the reader's limited processing time periods, namely *time windows*.

From the microscopic view, the whole process of microblog can be generally divided into four stages in our model: information publishing, receiving, processing and propagating. Next we will give fundamental mathematical definition and description of those stages and define information lost in microblog.

3.2 The Models for the Different Stages in Microblog Communication

Stage A: Information Publishing

The production of information in microblog is extensively broad participation involving nearly every user of it, which represents one of the remarkable differences between microblog and traditional social media, e.g. blog. We can observe it in two dimensions.

One is from the distribution of time intervals between two successive messages. Barabasi [23] and Vazquez's [16] works pointed out that in email, mobile communications and web browsing, timing of individual human actions are characterized by bursts of rapidly occurring events separated by long periods of inactivity. Tweet publishing and browsing are no exception but obey *Power Law* [15] distribution with heavy tails. $C = e^{c}$, α is called the *exponent*.

$$P(x) = Cx^{-\alpha} \tag{1}$$

The other dimension is from the information density distribution of time, which is as individual as the person but will show statistical stability as a whole. Though previous studies have addressed several models to simulate user behaviors in online network or web site , not much work has been done in microblogs. We suggest using statistical analysis to extract patterns of user behaviors in microblogs whose mathematical form can be written as

$$F_u(t) = \int_{\Delta t} f_u(\tau) d\tau \tag{2}$$

where $F_u(t)$ is the *information entropy* during time period Δt and $f_u(\tau)$ is its density of time.

Stage B: Information Receiving

Microblogs adopt mechanism of pushing friends' messages (or say tweets) to their followers automatically. Since the amount of friends of one specific reader may range from zero to hundreds, thousands or even more, and his or her friends post tweets in a particular way which obeys power law distribution, the time interval of received tweets will form a new distribution. Theoretically, when a reader has enough friends to satisfying the assumption of mathematical derivation of a typical Poisson distribution, we can prove the quantity of messages (denotes as k) arrived in time duration t is Poisson process

$$p_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \tag{3}$$

Let τ be the successive time interval of received messages, then the distribution of τ can be derived from (3). We have

$$p\{\mathrm{T} < t\} = 1 - p\{\mathrm{T} \geq t\} = 1 - p_0(t) = 1 - e^{-\lambda t} \tag{4}$$

where $t \geq 0$. Thus probability density function (PDF) would be negative exponential distribution

$$f_i(\tau) = \lambda e^{-\lambda \tau} \tag{5}$$

Stage C: Information Processing

Email or mobile phone message readers almost always read the emails or messages from their friends even though they don't need to reply. But in social networks like microblog network, reader may ignore the message, which, unlike email or phone message reading will not upset his or her friends. Such phenomenon is characterized by human dynamics. microblog is user centric communication, which means reader is no longer a mere and passive information container or processer. Instead, not only publishers can freely choose to post messages, but readers also can freely choose whether to read the messages. People can enjoy this kind of freedom only after the emergence of new social network forms.

It is relatively easier to model or do statistical analysis on publisher's behaviors than understand reader's, which means precisely deciding whether messages or which message being read is rather difficult. Traditional webpage browsing can be modeled as Random Walk [24]. However, in microblog, there is neither web links directing user to the next webpage, nor web logs indicating when and how the page is read. Besides, message reading is interests-related and affected by the strength of relationship between publisher and reader.

Here we address two possible methods as solutions. One is doing surveys among microblog users, from which empirical model can be created. Another one is building a support vector (denoted as V_s) and a weight vector (denoted as V_w) based on reading habits and relation strength, which can be applied to predicting the reading and reposting behavior of a reader. Each dimension of the vector is an initially normalized parameter representing one factor that attributes to the probability of message reading. Then what we need to do is dynamically adjusting V_w according to the algorithms we take. The algorithms can be borrowed from related subjects, e.g. neural network, pattern recognition and machine learning, and revised if needed. The ultimate result would be presented as formula (6), where R is the predicting factor.

$$R = V_s \cdot V_w^T \tag{6}$$

Stage D: Information Propagating

Information diffusion in blogs and microblogs has been studied in some aspects. In blogosphere, dynamics of information propagation in environments of low-overhead personal publishing is studied in both macro and micro scope. But it's far from

enough to fully understand information diffusion in microblog, especially when facing heavy information lost. Investigation on Twitter [10] also gives empirical conclusion on the speed, scale and range of topic diffusion. But those studies only deliver an overall prospect of topic diffusion and encounter difficulties when answering what topic will be propagated.

By adopting methods much like those mentioned in Stage C, we can build a support vector model of information diffusion, which would further explain when and what information propagates by introducing analysis of human dynamics.

Stage E: Information Lost

We can define *information density* and *processing ability*, of which the distribution of time is $f_i(t)$ and $f_p(t)$ respectively. Information density and processing ability refers to the entropy density of information received and processed in per unit time. Comparing $f_i(t)$ and $f_p(t)$, we notice that *information lost*, which is defined as the entropy of information being ignored by the reader, exists if the integration of $f_i(t)$ is greater than that of $f_p(t)$. Thus information lost during time period $\Delta\tau$, denoted as L, can be calculated as follows

$$L = \begin{cases} \int_{\Delta\tau} f_i(t) - f_p(t)dt & , \int_{\Delta\tau} f_i(t)dt > \int_{\Delta\tau} f_p(t)dt \\ 0 & , otherwise \end{cases} \tag{7}$$

Note that sometimes it could be difficult to decide the entropy of information, so for simplification we assume that each message has the same entropy $H(X_i)$, where X_i is the i_{th} message being received and get

$$LM = \begin{cases} (M-N)\cdot H(X_i) & , M > N \\ 0 & , M \le N \end{cases} \tag{8}$$

where M and N are the count of messages being received and processed by the reader, respectively.

4 Empirical

4.1 Data Source

In our study, we choose Sina-weibo as main empirical data source, which now is the largest microblog community in China with over 200 million users. A standard data set of about 2,000 typical users and 750,000 tweets has been built. Tweets from 3/1/2011 to 4/20/2011 were crawled. Due to privacy and limits of the Sina-weibo API, we can't obtain users' all tweets or whole follower list.

4.2 Individual Information Source

The interevent time of consecutive tweets obeys heavy-tailed Power Law distribution. We adopted the power law fitting methods. The mean exponent α is 1.8, see Fig 3.

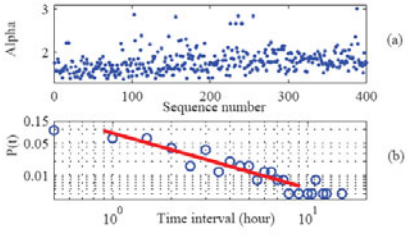

Fig. 3. User behavior of tweets publishing, fitted by Power Law. (a) value of power law exponent α. The mean value is α =1.78 with standard deviation σ =0.25. The goodness of fit, namely *p-value* is p = 0.52, σ =0.25. Note that if p is large (close to 1), the difference between the empirical data and the model can be attributed to statistical fluctuations alone; if it is small, the model is not a plausible fit. (b) plots a user's fitting (ID: 1463172125) with α =1.79 and p = 0.81. Heavy tails widely exist in time interval distributions. Overall tests show power law fit is applicable (p > 0.5 with support of 38.0% and p > 0.1 with support of 94.1%).

To understand the dynamics of individual publisher more comprehensively, we perform observations on message density through time line. The results indicate that no one curve can fit most distributions. Reversely, user behavior of information producing differs from one to another.

4.3 Information Receiving Patterns

Time interval of consecutive messages at the receiving end is our first concern. We have modeled this process as negative exponential distribution theoretically, which in fact matches empirical results pretty well. See Fig 4.

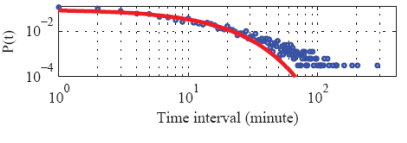

Fig. 4. Time interval distribution of consecutive tweets at individual follower. Tweets received in one month time duration were analyzed. Fitted by $P(t) = ae^{bt}$, axis x and y both are in logarithm. The figure shows a fit for a follower (ID: 1912337980, follows 50 publishers, 3,255 tweets received and counted.) where a = 0.089, b = -0.102 and R^2 = 0.905.

4.4 Information Processing and Diffusion in One Node

Firstly, we try to understand when one specific reader process messages received. The time window distribution of readers is shown in Fig 5. We draw the time window distributions indirectly using the data of posting and reposting time. Since there is no direct evidence of user online, we assume that posting and reposting (denoted as random variable X) closely correlate with message reading (denoted as random variable Y). The confidence of this assumption depends on the correlation coefficient

$$\rho_{XY} = \frac{E[X - E(X)][X - E(Y)]}{\sqrt{D(X)}\sqrt{D(Y)}} \tag{9}$$

where $E(X)$ and $D(X)$ represent expectation and variance of X respectively.

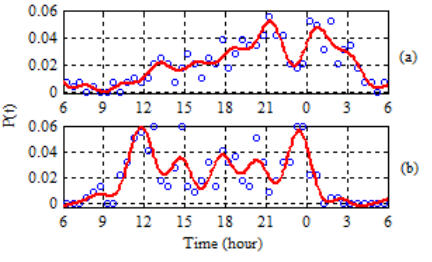

Fig. 5. Time window of readers: (a) time window of Yao Chen. (b) time window of Xiao S.We assume a day begins at 6:00 which corresponds to the time table of Chinese people. We found that most users' time window has two peaks indicating availabily of inforamtion processing.

Till now, we are no able to give a detailed description of information propagating through one node, or reader more precisely, which involves in-depth study of human dynamics. However, based on our data, we can determine the distribution of reposting time, illustrated in Fig 6.

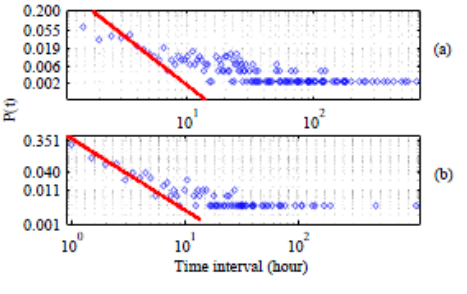

Fig. 6. The distribution of time interval between the original message being posted and being reposted by the reader. We performed power law fit on the data: (a) reader Yao Chen, where $p = 0.88$, $\alpha = 2.78$, fit area is $t < 7.0$ hours. (b) reader Xiao S, where $p = 0.56$, $\alpha = 2.17$, fit area is $t < 12.5$ hours. In the fit area, reposting time interval well matches the power law, and outside distribution tends to be the heavy tail of power law distribution.

The rudimentary conclusion of information diffusion in one node would be as follows: reposting is directly characterized by human dynamics, or more specifically, bursts of rapidly occurring events separated by long periods of inactivity [23]. In application, the "fit area" would be the golden time of information propagating through one node.

4.5 Information Lost Predictions

Here we assume the amount of information being processed is proportional to the time when the reader is online. Due to the diversity in time window distribution of different users, we can only deal with specific reader using statistical methods.

Now we apply our model to the prediction of information lost of user Yao Chen. We have already extracted the information receiving distribution (see Fig 5) and time window distribution (see Fig 6). We define three parameters: m, the average receiving messages per minute; v, the messages can be read by reader per minute time; $\Delta t_n = t_{n2} - t_{n1}$, the n_{th} online duration. Thus formula (8) can be written as

$$LM = \sum_n \{m \int_{t_{n1}}^{t_{n2}} f_i(t)dt - v\Delta t_n \int_{t_{n1}}^{t_{n2}} f_r(t)dt\} \tag{10}$$

5 Conclusions

Our work aims at building a framework to explain interactions between users and information diffusion between publishers and readers. Thus we have introduced the competing window model, which provides the fundamental framework to answer questions about information producing, receiving, processing and propagating. The model applies to microscopic perspective, which is also the foundation of macro network and information diffusion.

To verify this framework, we applied it to real social network, Sina-weibo. The empirical data provides detailed observation of microblog and realistic evidence of human dynamics and proves the feasibility and robustness of our general framework.

A framework can't solve every concrete problem. This is our rudimentary work and further studies are needed to better understand human dynamics involved information diffusion.

References

1. Barabasi, A.-L., Albert, R.: Emergence of Scaling in Random Networks. Science 286(5439), 509–512 (1999)
2. Fu, F., Liu, L.: Empirical Analysis of Online Social Networks in the Age of Web2.0. Physica A (2007)
3. Gross, T., Blasius, B.: Adaptive Coevolutionary Networks- A Review. Journal of the Royal Society – Interface 5, 259–271 (2008)

4. Bringmann, B., Berlingerio, M., Bonchi, F., Gionis, A.: Learning and Predicting the Evolution of Social Networks. IEEE Intelligent Systems 25(4), 26–35 (2010)
5. Xu, B., Liu, L.: Information diffusion through online social networks. In: 2010 IEEE International Conference on Emergency Management and Management Sciences (ICEMMS), August 8-10, pp. 53–56 (2010)
6. Yang, J., Leskovec, J.: Modeling Information Diffusion in Implicit Networks. In: 2010 IEEE 10th International Conference on Data Mining (ICDM), December 13-17, pp. 599–608 (2010)
7. Bird, C., Gourley, A., Devanbud, P., et al.: Mining email social networks. In: Proceedings of the 2006 International Workshop on Mining Software Repositories, Shanghai, China, pp. 137–143 (2006)
8. Yassine, M., Hajj, H.: A Framework for Emotion Mining from Text in Online Social Networks. In: 2010 IEEE International Conference on Data Mining Workshops (ICDMW), December 13, pp. 1136–1142 (2010)
9. Teutle, A.R.M.: Twitter: Network properties analysis. In: 2010 20th International Conference on Electronics, Communications and Computer (CONIELECOMP), February 22-24, pp. 180–186 (2010)
10. Yang, J., Counts, S.: Predicting the Speed, Scale, and Range of Information Diffusion in Twitter. In: Proc. ICWSM (2010)
11. Boyd, D., Golder, S., Lotan, G.: Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In: 2010 43rd Hawaii International Conference on System Sciences (HICSS), January 5-8, pp. 1–10 (2010)
12. Wang, A.H.: Don't follow me- Spam detection in Twitter. In: Proceedings of the International Conference on Security and Cryptography, SECRYPT (2010)
13. Haight, F.A.: Handbook of the Poisson distribution. Wiley, New York (1967)
14. Reynolds, P.: Call center staffing. The call Center School Press, Lebanon (2003)
15. Newman, M.E.J.: Power laws, Pareto distributions and Zipf's law. Contemporary Physics 46(5), 323–351 (2005)
16. Vazquez, A., Oliveira, J.G., DezsöGoh, K.I., Kondor, I., Barabasi, A.L.: Modeling bursts and heavy tails in human dynamics. Phys. Rev. E 73, 36127 (2006)
17. Gabrielli, A., Caldarelli, G.: Invasion percolation and critical transient in the Barabasi model of human dynamics. Phys. Rev. Lett. 98, 208701 (2007)
18. Goncalves, B., Ramasco, J.: Human dynamics revealed through Web analytics. Phys. Rev. E 78, 26123 (2008)
19. Han, X.P., Zhou, T., Wang, B.H.: Modeling human dynamics with adaptive interest. New. J. Phys. 7, 73010–73017 (2008)
20. Malmgen, R.D., Stouffer, D.B., Motter, A.E., Amaral, L.A.N.: A Poissonian explanation for heavy tails in e-mail communication. Proc. Natl. Acad. Sci. USA 105, 18153–18158 (2008)
21. Malmgen, R.D., et al.: On universality in human correspondence activity. Science 325, 1696–1700 (2009)
22. V´azquez, A.: Impact of memory on human dynamics. Physica A 373, 747–752 (2007)
23. Barabási, A.L.: The origin of bursts and heavy tails in human dynamics. Nature 435, 207–211 (2005)
24. Pearson, K.A.R.L.: The Problem of the Random Walk. Nature 72(1867), 342–342 (1905)