Information Sciences 187 (2012) 15–32

Contents lists available at SciVerse ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A time-varying propagation model of hot topic on BBS sites and Blog networks

Beibei Zhang [a,*], Xiaohong Guan [a,b], Muhammad Junaid Khan [c], Yadong Zhou [a]

[a] Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University, PR China
[b] Center for Intelligent and Networked Systems and TNLIST Laboratory, Tsinghua University, PR China
[c] National University of Science and Technology, Karachi, Pakistan

ARTICLE INFO

Article history:
Received 26 February 2009
Received in revised form 22 October 2010
Accepted 20 September 2011
Available online 22 October 2011

Keywords:
Collective behavior
Uniformly weakly persistent
Single-peak
Multi-peak
Propagation model

ABSTRACT

Modeling the propagation of hot online topic is a preliminary requirement of predicting the trend of hot online topic. We propose a time-varying hot topic propagation model in online discussion context based upon the collective behavior of users who are in different social subgroups on blog networks and bulletin board system (BBS) sites. By analyzing the stability of the equilibrium of our model, we search for the threshold to be watershed of the trend of hot online topic and generalize about two theorems from the results of analysis, they exposit two sufficient conditions under which the trend of hot online topic will die out or remain uniformly weakly persistent. Furthermore, we propose methods to predict the trend of hot online topic on the strength of our model and theorems. For different motivation, we design two methods: Method (I) is mainly served as a way of theoretical research for predicting long trend of single-peak hot online topic by the thresholds of theorems; and for application, we design method (II) to predict the number of users writing or commenting upon article posts with respect to multi-peak hot online topic and single-peak one in the following two days with the help of Method (I). Experiments of two methods are performed on widely-discussed topics on the Sina Blog and the famous Liang Quan Qi Mei (LQQM) BBS and Xi'an Jiaotong University (BMY) BBS in China. The experimental results show that our methods predict the trend of hot online topic efficiently not only for theoretical motivation but also for applicable motivation, and reduce the computational complexity. Hence, our model can serve as basis for predicting trends in hot online topic propagation.

© 2011 Elsevier Inc. All rights reserved.

1. Introduction

With the rapid development of Internet technology, particularly the emergence of Web 2.0 sites [20] such as blog networks, BBS sites and social networks, people are not only exchanging information online but also expressing their ideas and opinions openly. Generally speaking, blogs and BBS are two kinds of web 2.0 sites which are extremely popular in China, much like twitter and Facebook in Europe and the US. Essentially BBS site serves as electronic information center and emerging media, which is used to offer public information releases, chatting, mail and other services as an alternative to physical bulletin platform. In 2005, there were at least 123 BBS sites at 78 universities and colleges in China, and these BBS sites have several million registered users. Popular BBS sites in China include the LQQM BBS established in Tsing University and the

* Corresponding author.
E-mail addresses: bbzhang@sei.xjtu.edu.cn (B. Zhang), xhguan@sei.xjtu.edu.cn (X. Guan), contactjunaid@yahoo.com (M.J. Khan), ydzhou@sei.xjtu.edu.cn (Y. Zhou).

0020-0255/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2011.09.025

BMY BBS established at Xi'an Jiaotong University. For the Sina blog in China, there are almost 230 million users totally all over the world. Many people like expressing their opinion about hot topics on blog and BBS in recent years. For example, "08 Olympic" and "Melamine contaminated milk powder incident" are both hot topic on the BMY BBS and LQQM BBS, whereas "Roh Moo-Hyun's death" and "Influenza A[H1N1]" are hot topics on the Sina blog.

Factually online topics are playing significant roles in public life undeniably. To understand the evolutionary trend of different hot online topic, it is necessary to study the propagation mechanism of hot online topic.

Theoretically if we can analyze each single user's behavior for individuals participating in online discussion, maybe we can give a clear explanation for all users' behavior trend as to hot online topic. However, the information of users' behavior is often embedded in a vast size of network information; for example, mining web information from Sina blog will produce several millions MB data per day, because of web sites updating every day, thus this way is very time-consuming and is not feasible. In another way, we can model the propagation of hot online topics by studying users' group collective behavior. This understanding of users' collective behavior helps us to establish the propagation process of hot online topic as well as effectively let us tackle the disorientation problem in modeling.

In this paper, we propose a time-varying hot topic propagation model (THTPM for abbr.) on BBS or blog to describe the collective behavior of users who are in different online social subgroups. The greatest merit of our model is that it can be applied to reflect the user's state transition process efficiently and does not depend on any empirical parameters. For accessing to the threshold by which we can prejudge the trend of hot topic discussion, we analyze the stability of equilibrium of THTPM and conclude two theorems. They reflect: when threshold $Q^* < 1$, hot topic will die out, and when threshold $Q_* > 1$, hot topic will keep uniformly weakly persistent. Furthermore, we propose method (I) and method (II) to predict the trend of hot online topic based on our model for the consideration of theoretical motivation and practical application respectively. By Method (I) we can predict the trend of single- peak hot topic mainly for theoretical motivation, however, in real internet circumstance, there are many multi-peak hot topics, and so we design method (II) to predict the size of discusser- users writing or commenting upon article posts mainly with respect to multi-peak hot online topic in the following two days but it also can predict single-peak hot-topic trend.

Our main contribution lies in that we firstly put forward a novel time-varying state model to depict the dissemination mechanism of hot online topic. Meantime, we bring forward the first universal trend prediction theory of hot topic based on moving time windows which can be applied to distinct hot topics effectively and our results show the proposed prediction method is validate for both single-peak and multi-peak hot topics.

The rest of the paper is organized as follows. Section 2 discusses related work; Section 3 shows problem formation. Section 4 describes definition about the dynamic propagation model of hot online topic. In Section 5, we analyze the stability of the equilibrium of the topic in system (2.1). Experiments are discussed in Section 6. Section 7 presents the conclusions and future work.

2. Related work

To the best of our knowledge, scholarly interest in social media analysis has increased due to the growing use of tools such as weblogs in past several years. Many earlier studies tended to engage in different aspect of the following five stages in blog networks from topic data acquisition to propagation modeling of hot online topic. First, semantic analysis was used to develop data mining and topic detection techniques to research online topics on inception. Zheng [29] proposed a document representation methodology to take into account both noun phrases and various semantic relationships, as there were a number of semantic relationships that could relate a pair of words. Ginter [6] proposed an unsupervised method, based on hidden Markov models, which was combined with latent semantic analysis to freely define topics of interest without necessarily data annotation; this method could also be used to identify short segments. Second, data mining [2,9,10,24,26,27] has been developed to study social media to identify textual keywords that refer to important events or topics. For instance, the so-called "trigger segment detection" method [9] was used to identify important segments of conversations by tracking changes in the classifier designed to distinguish between business outcomes; then the study used data mining to extract important linkages between key entities e.g., insights. Meanwhile, Hristidis [10] presented a model that took into consideration the continuous flow of text in streams and used efficient pipelined algorithms so that the results could be readily made available. Compact pattern stream tree (CPS-tree) [24] was developed to identify the complete set of recent frequent patterns from a high-speed data stream over a sliding window. A novel approach [27] was derived for exposing a universal representation for blogs and the structural similarities among blogs and blog posts to make them available for reuse. A third focus involved topic detection in blogs [21,22]; Sekiguchi [22] studied topic detection based on similarities among user interests. In the context, a general probabilistic algorithm [21] could effectively identify correlated patterns and their periods across text streams, even if the streams had completely different vocabularies. The fourth emphasis focused on the analysis of the social facets of blog communities [3,4,7,11–16,18,23,28], including content analyses of blogs and studies of social structures [11,13]. Content analysis and topic detection were typically performed on blogs by tracking the diffusion of information through the blogosphere [7] or by analyzing the sentiments expressed on blogs [16,18]. Content analysis was performed on medical question and answer portals [4], medical weblogs, medical reviews and Wikis to develop an overview of the medical content available online. The last major area of research was in modeling the propagation of hot topic discussions to describe the mechanism of hot topics propagation in online social network. For example, Zhou [30] put forward a dynamic

probability model to predict trends in topic discussions on online social networks. Zhao [28] used susceptible-infection (SI) model based on individual fitness to describe the propagation of incidental topics.

Our research focuses on the fifth area-modeling the propagation of hot online topics. However, due to the nature of this problem, the previous works suffered two major issues:

(1) Serious dependence on empirical parameters: Zhao [28] proposed susceptible-infection (SI) model to model the propagation of incidentals topics based on empirical parameter-individual fitness and analyze the propagation velocity, author found out the propagation velocity of incidental topics depended on the individual fitness. But it is very difficult to estimate individual fitness in precision. Zhou [30] proposed the dynamic probability model based on three factors including individual interest, group behavior and time lapse but the model depended on certain empirical parameters, for example, individual interest by human experience, which limited the suitability of the model.

(2) Incapability to work well for both single-peak and multi-peak hot topic: It is well known that the single-peak and multi-peak phenomena exist in common. Basically the dynamic probability model [30] predict the user's behavior, i.e. attending the topic discussion or not, and then obtain the number of the attending users and prediction method based on the probability model was just adapted to single-peak topic. It was not available as a method of predicting the trend of multi-peak hot topic discussion though the work did show the single-peak topic spread and rich variation on a daily basis. Many researchers have realized this issue, but still there is no evidence on existing works to cope with this situation until now.

Because of these issues, the previous works can hardly be applied to predict the trend in both single-peak and multi-peak hot online topic discussion though their works enlighten us on understanding how to model the propagation of hot online topic. To resolve this dilemma, we develop the THTPM model which utilizes the features of the collective behavior of users who are in different social subgroups.

3. Problem formulations

Now we formalize the problem of modeling the propagation of hot online topic.

On the beginning, we assume target hot topic as T_D, other topics which belong to the same category as target hot topic as T_R, topics category as T.

User subset $D = \{d_k(t)\}(1 \leqslant k \leqslant m, m < n)$ denotes the Discussed Group, where $d_k(t)$ represents the individual who participates in writing or commenting upon article posts with respect to target hot online topic T_D at time t (we can deem $d_k(t)$ as discusser of topics for convenience); And a user subset $R = \{r_l(t)\}(1 \leqslant l \leqslant m_1, m_1 < n)$ denotes the Related Group, where $r_l(t)$ represents the individual who writes or comments upon article posts with respect to other topics T_R which belong to same topic category T at time t; And a user subset $E = \{e_l(t)\}(1 \leqslant l \leqslant m_2, m_2 < n)$ indicates the Exited Group, where $e_l(t)$ represents the individual who wrote or commented upon article posts with respect to target hot online topic T_D at time $t-1$ but quit from the Discussion Group at time t. We regard $R(t), D(t)$ and $E(t)$ as the state variable of three different user group respectively, and the propagation of target hot online topic modeling can be expressed as differential equations of user's state vector $(R(t), D(t), E(t))$.

The following Fig. 3.1 represents a concise block diagram that illustrates user's state transition among all subgroups $R(t)$, $D(t)$, and $E(t)$.

For better to understand the correlation among $D(t)$, $R(t)$ and $E(t)$, we analogize the propagation of hot online topic to the epidemics. Actually through the observation and analysis of the propagation of hot online topic in our sample data, we found some helpful features for modeling the propagation of hot online topic. First, the propagation mechanism of hot topic in blog or on BBS is indeed similar with the mechanism of the epidemics [1,5,28]. Second, the propagation of a hot topic on blog network or BBS site is not up to only one individual or several individuals but up to the collective effort of many users just

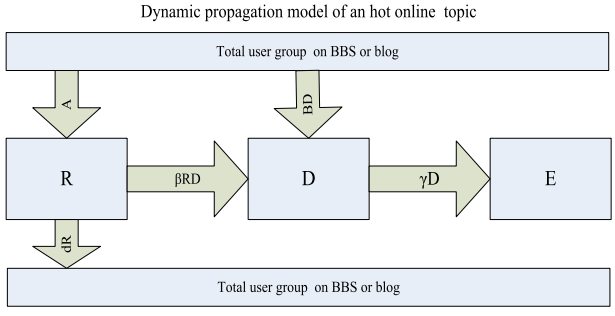

Fig. 3.1. Dynamic propagation model of an hot online topic.

like the spreading process of an epidemic. Because the propagation of hot online topic is influenced by many delicate factors and the equations will become very complicated, it is reasonable that we choose time-varying SIR model based on the users-sharing scheme [17].

4. Model definitions

We introduce the time-varying state equation to describe the transmission mechanism of hot online topic at now.

By Fig. 3.1, it is reasonable for us to suppose that each user in exited group $E(t)$ loses interest in writing or commenting upon article posts with regard to any topics in the same topics category T as target hot topic T_D. Then we establish the following state equations according to previous discussion and Fig. 3.1:

$$\begin{cases} \frac{dR}{dt} = A - \beta RD - dR \\ \frac{dD}{dt} = BD + \beta RD - \gamma D \\ \frac{dE}{dt} = \gamma D \end{cases} \tag{2.1}$$

We describe all parameters as follows.

(1) $A(t)$ denotes the input rate of the related group R, it represents the number of those who wrote or commented upon article posts as to completely unrelated topics at time $t-1$ but wrote articles about related topics T_R at time t.
(2) $d(t)$ denotes the output rate of the related group R, $d(t) = Z/w$, where Z is the number of those who wrote or commented upon article posts with respect to related topic T_R at time $t-1$ but did not write or comment upon article posts with respect to related topics T_R at time t, and w is the number of those who wrote or commented upon article posts with respect to hot online topic T_D at time $t-1$.
(3) $B(t)$ denotes the input rate of the discussion group D, $B(t) = c/u$, where c is the number of those who wrote or commented upon article posts with respect to completely unrelated topics at time $t-1$ but wrote or commented upon article posts with respect to target hot online topicT_D at time t, and u is the number of those who wrote or commented upon article posts with respect to target hot online topic T_D at time $t-1$.
(4) β denotes the coefficient of the spreading rate, β is the average number of those who have contacted any member of the discussion group after reading an article.
(5) γ denotes the coefficient of overflowing rate, $\gamma = a(t)/l(t-1)$, where $a(t)$ denotes the number of those who wrote or commented upon article posts with respect to target hot online topic T_D at time $t-1$ but quit from the discussion group at time t, and $l(t-1)$ is the number of those who wrote or commented upon article posts with respect to hot online topic T_D at time $t-1$.

5. Analysis of hot online topics propagation model

For the sake of developing the proper method to predict the trend of hot online topic by our time-varying dynamic model effectively, we should understand the stability of the equilibrium of our time-varying dynamic model. Therefore, we first analyze the stability of the equilibrium of the constant dynamic propagation model (CDPM) where all the parameters are constant. It will help us to understand the stability of the equilibrium of our time-varying hot topic propagation model (THTPM). The THTPM model, firstly proposed by us, is the state equations with all time-varying parameters but overflowing rate $\gamma$to depict the transmission mechanism of hot online topic in introduction.

5.1. Constant dynamic propagation model (CDPM)

We will analysis the stability of the equilibrium of constant system (2.1) as following:

In view of the specific feature of third equation of constant system (2.1), the state equations in system (2.1) can be rewritten as

$$\begin{cases} \frac{dR}{dt} = A - \beta RD - dR \\ \frac{dD}{dt} = BD + \beta RD - \gamma D \end{cases} \tag{2.2}$$

For determination of the equilibrium of system (2.2), let

$$\begin{cases} \frac{dR}{dt} = A - \beta RD - dR = 0 \\ \frac{dD}{dt} = BD + \beta RD - \gamma D = 0 \end{cases} \tag{2.3}$$

Solving Eq. (2.3),

$$D = 0, \text{ or, } B + \beta R - \gamma = 0 \tag{2.4}$$

(1) When $D = 0$, solving $A - \beta RD - dR = 0$,

$$R_0 = \frac{A}{d} \tag{2.5}$$

We capture the non-trivial equilibrium $F_0 = (A/d, 0)$.

(2) When $B + \beta R - \gamma = 0$ solving $A - \beta RD - dR = 0$,

$$D = \frac{A}{\gamma - B} - \frac{d}{\beta} \tag{2.6}$$

Therefore, we capture the positive equilibrium $F^+(R^+, D^+) = F^+(\gamma - B/\beta, A/\gamma - B - d/\beta)$ where $\gamma - \beta > 0$.

We have illuminated the existence of the equilibrium of system (2.2), and next, we will elucidate the stability of the equilibrium of system (2.2).

In the first beginning, for the purpose of treating of the stability of the equilibrium of system (2.2), by Eq. (2.6), we obtain threshold Q_0, which is viewed as our basic reproductive number [8].

$$Q_0 = \frac{A\beta}{(\gamma - B)d} \tag{2.7}$$

Q_0 represents a criterion to prejudge whether the trend of a hot online topic will be growing or withering.

At the same time, we establish set H which satisfies $(R, D) \in H = \{(R, D) | R > 0, D > 0, R + D \leqslant \frac{A}{d}\}$, and $\overline{Q_0} = Q_0 = A\beta/(\gamma - B)d$.

From the expression of $\overline{Q_0}$, we infer that $\overline{Q_0} \leqslant 1$ or $\overline{Q_0} > 1$.

(1) When $\overline{Q_0} < 1$, for treating of the stability of $F_0(R_0, 0)$ in system (2.2), we transform the equilibrium $F_0(R_0, 0)$ to the origin by means of a coordinate transformation $x = R - R_0$ so that system (2.2) can be adapted to:

$$\begin{cases} \frac{dx}{dt} = -\beta(x + R_0)I - dx \\ \frac{dD}{dt} = (B - \gamma)D + \beta(x + R_0)D \end{cases} \text{or} \begin{cases} \frac{dx}{dt} = -\beta(x + \frac{A}{d})D - dx \\ \frac{dD}{dt} = (B - \gamma)D + \beta(x + \frac{A}{d})D \end{cases} \tag{2.8}$$

While $(x, D) \in \overline{H} = \{(x, D) | x > -R_0, D > 0, x + D \leqslant \frac{A}{d} - R_0\}$

We utilize $V_1 = x^2/2 + R_0 D$ as the Lyapunov function.

Differentiating V_1 along the curve of the solution of system (2.8) results in

$$\left.\frac{dV_1}{dt}\right|_{(2.8)} = -\beta\left(x^2 + \frac{A}{d}x\right)D - dx^2 + (B - \gamma)D\frac{A}{d} + \beta\left(x + \frac{A}{d}\right)D\frac{A}{d} = -\beta x^2 D - dx^2 + \left[(B - \gamma) + \frac{\beta A}{d}\right]D\frac{A}{d}.$$

Because of $\overline{Q_0} < 1, Q_0 < 1$ and $[(B - \gamma) + \beta A/d]DA/d < 0$, dV_1/dt is negative in $\overline{H}$ all but $x = D = 0$, and thus F_0 is globally asymptotically stable in field H.

(2) When $\overline{Q_0} = Q_0 = 1, dV_1/dt$ always maintains negative in field H. From the first equation of system (2.8), we can easily come to a conclusion: If $x = 0$ is the solution of this system (2.8), D must be zero. Therefore, if there is no non-trivial solution curve in the set of which each element satisfies the condition $dV_1/dt = 0$, and then F_0 is still globally asymptotically stable in field H. Thus, this implies that no one is writing or commenting upon article posts with regard to hot online topic.

(3) When $\overline{Q_0} > 1$, there is a positive equilibrium $F^+(R^+, D^+)$ in addition to $F_0(R_0, 0)$ for system (2.2). We now testify the stability of $F^+(R^+, D^+)$, $(R^+ = (\gamma - B)/\beta, D^+ = A/(\gamma - B) - d/\beta)$.

Next we will prove the global asymptotic stability of F^+ and the instability of F_0 in field H.

Letting $x = R - R^+$, system (2.2) can be rewritten as

$$\begin{cases} \frac{dx}{dt} = -dx - \beta xD - (\gamma - B)(D - D^+) \\ \frac{dD}{dt} = \beta xD \end{cases} \tag{2.9}$$

Note that $D^+ = \frac{A}{\gamma - B}\left(1 - \frac{1}{Q_0}\right) > 0$.

Then $V_2 = x^2/2 + (\gamma - B)(D - D^+ - D^+ \ln D/D^+)/\beta$ is used as the Lyapunov function, we differentiate V_2 along the curve of system (2.9), which leads up to

$$\left.\frac{dV_2}{dt}\right|_{(2.9)} = -dx^2 - \beta x^2 D - (\gamma - B)(D - D^+)x + (\gamma - B)(D - D^+)x = -dx^2 - \beta x^2 D.$$

It is suggested that $x = 0$ is necessary and sufficient for $dV_2/dt|_{(2.9)} = 0$; moreover, seeing that $x = 0$ is the solution of system (2.9), D must be equal to D^+. The set of solutions of system (2.9) meeting $dV_2/dt|_{(2.9)} = 0$ does not include non-trivial solution. Thus, $F^+(R^+, D^+)$ keeps up global asymptotic stability in field H, and obviously F_0 is unstable.

In the light of previous discussion, $\overline{Q_0}$ will be a threshold of the trend of hot online topic;

(1) When $\overline{Q_0} = Q_0 > 1$, positive equilibrium $F^+(R^+, D+)$ retains asymptotic stability. This implies that the hot topic will persist for a long time.
(2) When $\overline{Q_0} = Q_0 < 1$, non-trivial equilibrium $F_0 = (A/d, 0)$ holds asymptotic stability and the hot topic will eventually die out. So that Q_0 is the watershed for the propagation trend of a hot topic.

5.2. Time-varying hot topic propagation model (THTPM)

In this subsection, we will introduce the time-varying state equations-system (2.10) (THTPM) and explore the stability of the equilibrium of THTPM as following:

$$\begin{cases} \frac{dR}{dt} = A(t) - \beta(t)RD - d(t)R \\ \frac{dD}{dt} = B(t)D + \beta(t)RD - \gamma D \end{cases} \tag{2.10}$$

Where R, D are state variables, and $A(t)$, $\beta(t)$, $d(t)$ and $B(t)$ are time-varying, continuous and bounded parameters, $\beta(t) > 0$, $d(t) > 0$, $B(t) \geqslant 0$, $A(t) > 0$, $t \geqslant 0$, $\gamma - B(t) > 0$, and $R(t) > g > 0 (\exists g > 0,\ \forall t \geqslant 0)$. Taking advantage of the former analysis of the constant dynamic propagation model (CDPM) for hot topic, we get $F_0 = (R_0, 0) = (A(t)/d(t), 0)$, and it is easy to achieve its stability. But it is difficult to obtain the stability of positive equilibrium $F^+(R^+, D^+)$ of system (2.10); we cannot do it just as the previous way in subsection 5.1. By way of analyzing the second equation of system (2.10), we find out the threshold as:

$$Q^* = \frac{\left(\beta\left(\sup\left(\max_t\left(\frac{\gamma - B(t)}{\beta(t)}, \frac{A(t)}{d(t)}\right)\right)\right)\right)^*}{(\gamma - B)_*}$$

Where $(\gamma - B)_* = \lim_{t\to\infty} \inf\left(\frac{1}{t}\int_0^t (\gamma - B)ds\right)$.

$$\left(\beta\left(\sup\left(\max_t\left(\frac{\gamma - B(t)}{\beta(t)}, \frac{A(t)}{d(t)}\right)\right)\right)\right)^* = \lim_{t\to\infty} \sup\left(\frac{1}{t}\int_0^t \beta\left(\sup\left(\max_t\left(\frac{\gamma - B(t)}{\beta(t)}, \frac{A(t)}{d(t)}\right)\right)\right)ds\right).$$

Theorem 5.1. *When $Q^* < 1$, the hot online topic on a given topics category T will die out, in other words for every solution of system (2.10), $\lim_{t\to\infty} D(t) = 0$.*

Proof. See Appendix A. □

Next we specify the uniformly weak persistence of the hot topic. We introduce the concept of uniformly weak persistence, which implies that there exists some $\varepsilon > 0$ such that $D^\infty = \lim_{t\to\infty} D(t) > \varepsilon$ for all solutions of system (2.10) with $D(r) > 0$ for some $r \geqslant 0$. In other words, if the hot online topic maintains uniformly weak persistent, there will always exist active users on blog network or BBS site to write or comment upon article posts with regard to hot topic. The size of $D(t)$ may not be large, but discussers will continue to follow the topic discussion.

Analyzing system (2.10), we regard

$$Q_* = \frac{\left(\beta\left(\inf\left(\min_t\left(\frac{\gamma - B(t)}{\beta(t)}, \frac{A(t)}{d(t)}\right)\right)\right)\right)_*}{(\gamma - B)^*} > 1 \Rightarrow \left(\beta\left(\inf\left(\min_t\left(\frac{\gamma - B(t)}{\beta(t)}, \frac{A(t)}{d(t)}\right)\right)\right)\right)_* > (\gamma - B)^*$$

as the threshold.

Where $(\gamma - B)^* = \lim_{t\to\infty} \sup\left(\frac{1}{t}\int_0^t (\gamma - B)ds\right)$,

$$\left(\beta\left(\inf\left(\min_t\left(\frac{\gamma - B(t)}{\beta(t)}, \frac{A(t)}{d(t)}\right)\right)\right)\right)_* = \lim_{t\to\infty} \inf\left(\frac{1}{t}\int_0^t \beta\left(\inf\left(\min_t\left(\frac{\gamma - B(t)}{\beta(t)}, \frac{A(t)}{d(t)}\right)\right)\right)ds\right).$$

We know that $\exists r > 0$ for $\forall t \geqslant r > 0$, $R \geqslant \inf(\min_t((\gamma - B(t))/\beta(t), A(t)/d(t)))$.

Theorem 5.2. *When $Q_* > 1$, the hot online topic will be uniformly weakly persistent.*

Proof. See Appendix B. □

6. Experiments

For consideration of both theoretical motivation and practical application, we excogitate Method (I) to validate the model mainly for theoretical motivation and verify the validity of our theorems, we exemplify how to establish the threshold Q^* or Q_* and how to prejudge the trend of hot online topic by the threshold. Meantime, we design method (II) to overcome the shortcoming of Method (I) to be practically applied for predicting the trend of single-peak hot topic as well as multi-peak one. Method (II) has internal relationship with Method (I) actually; in each cycle of calculation of method (II) we would employ some steps of Method (I), there are subtle difference between Method (I) which is charge in predicting the trend of hot topic for a relative long time and method (II) which is charge in predicting the population of discusser in the following two days. In Method (II), we bring forward the first universal hot topic trend prediction theory based on moving time windows and it can be applied to distinct hot topic effectively and our results shows the proposed prediction method is valid for both single-peak and multi-peak hot topic.

6.1. Method (I): propagation prediction method of single-peak topic

We set up the following experiment for hot topics which have only one peak. (See Table 1)

Step 1: First to extract raw data from the LQQM BBS, the BMY BBS and the Sina blog during the period from 1 Nov07 to 15 Nov09 about hot online topics.

Step 2: Then we separate users' group on blog network into related, discussion and exited groups according to definitions in Section 3. For example, for the hot topic "Roh Moo-hyun's death" on the Sina blog, we look upon those who wrote or commented upon article posts with regard to other topics in the category (that is, politics) where the topic "Roh Moo-hyun's death" are as members of the related group; those who wrote or commented upon article posts in the Sina blog with regard to the topic "Roh Moo-hyun's death" are members of the discussion group. We consider those who wrote or commented upon article posts with respect to hot online topic at time $t-1$ but quit the discussion group at time t as members of the exited group.

Step 3: Select some data extracted from the BMY BBS, the LQQM BBS and the Sina blog as sampling sets to be fitted for parameters of the model.

Step 4: By treating the propagation of hot online topic as an epidemic propagation, we procure the coefficient of overflowing rate γ. Because $1/\gamma$ is the average life span of discusser based on the result of epidemiology [25]; this feature is helpful to estimate γ.

Step 5: Use data-based modeling [19] to filter the raw data. Then we should calculate β of system (2.10). However, it is very difficult to directly obtain the coefficient of the spreading rate β because it is almost impossible to exactly capture concrete information on the contact rate about the hot online topic. It is not too difficult to get the spreading rate $\beta(t)R(t)$ which is the input rate from the related group to the discussion group. Using the least-squares method, we estimate the spreading rate $\beta(t)R(t)$ of a hot online topic in system (2.10).

Step 6: By the least squares estimation method, we estimate the input rate of the related group $A(t)$, the output rate of the related group $d(t)$ and the input rate of the discussion group $B(t)$ for the hot online topic in system (2.10).

Step 7: via employing our model, we predict the size of the discussion group. And the threshold of hot online topic can also be obtained. Then comparison diagram between real data and prediction shows the validity of method (I).

(1) Based on least-squares method, we capture the probability distribution of γ and the mean value of the probability distribution in step 4 of method (I), the mean value is the estimated value of γ. The results indicate that the probability distributions follow exponential distribution with parameter $1/\gamma$ for all hot topics in Fig. 6.1.

Table 1
Describes the source of the raw data.

Hot topic	Resource	Time
Huawei recruitment	LQQM BBS	15 Nov07 to 25 Dec07
2008 Olympics	LQQM BBS	1 Jul08 to 19 Aug08
Tomato garden	LQQM BBS	20 Aug08 to 24 Sep08
Melamine-contaminated milk powder incident	LQQM BBS	12 Aug08 to 14 Sep08
Shenzhou spacecraft	LQQM BBS	8 Sep08 to 6 October08
Melamine-contaminated milk powder incident	BMY BBS	12 Sep08 to 14 Oct08
08 Olympic	BMY BBS	1 Jul08 to 19 Aug08
Policeman beating somebody to death	BMY BBS	1 Sep08 to 2 Oct08
Huawei recruitment	BMY BBS	15 Nov07 to 25 Dec07
Obama's inauguration	Sina blog	1 Jan09 to 15 Nov09
Roh Moo-hyun's death	Sina blog	23 May09 to 4 Jul09
Drag racing and vehicular manslaughter in Hangzhou	Sina blog	28 Jul09 to 6 Sep09
Influenza A [H1N1]	Sina blog	17 Apr08 to 16 May09

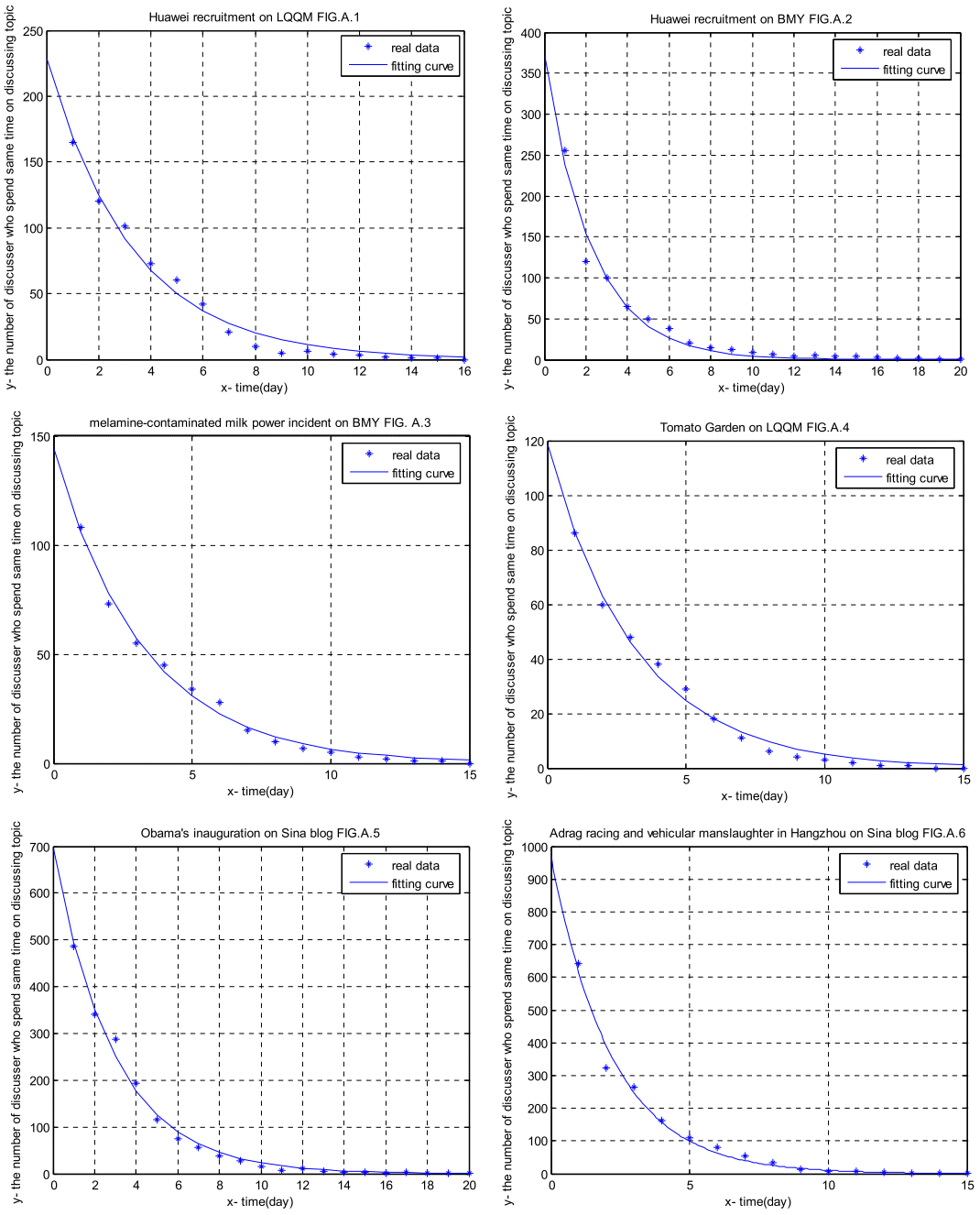

Fig. 6.1. Estimation on overflowing rate: y denotes the actual total number of users spending the same time on writing or commenting in blog network or on BBS site upon article posts with respect to hot online topics "Huawei recruitment" on the BMY BBS and the LQQM BBS, "Tomato Garden" on the LQQM BBS, "melamine-contaminated milk powder incident" on the BMY BBS, "Obama's inauguration" and "Drag racing vehicular manslaughter in Hangzhou" on the Sina blog respectively. And x represents the fitting time. The real data are consistent with the estimated exponential distribution.

For example, the fitting distribution is $f(x) = 1560^{*}\exp(-1.056^{*}x)$ for the topic "Huawei recruitment" on the LQQM BBS, $f(x) = 118.3*\exp(-0.3143*x)$ for the topic "melamine-contaminated milk powder incident" on the LQQM BBS, $f(x) = 1267*\exp(-1.218*x)$ for the topic "Huawei recruitment" on the BMY BBS, and $f(x) = 144.2*\exp(-0.3092*x)$ for the topic "Tomato Garden" on the BMY BBS. $1/\gamma$ is the mean value of the fitting distribution.

(2) Fig. 6.2 shows the comparison on the spreading rate $\beta(t)R(t)$ between the real data and fitting distribution for all hot topics in Sina blog or on BMY BBS and LQQM BBS. When the data are de-noised, the data reflecting the regularity are clear. The fitting result is more precise. By the least square method, we find out the gauss distribution to be the fitting distribution based on least-residual in the step 5 of method (I).

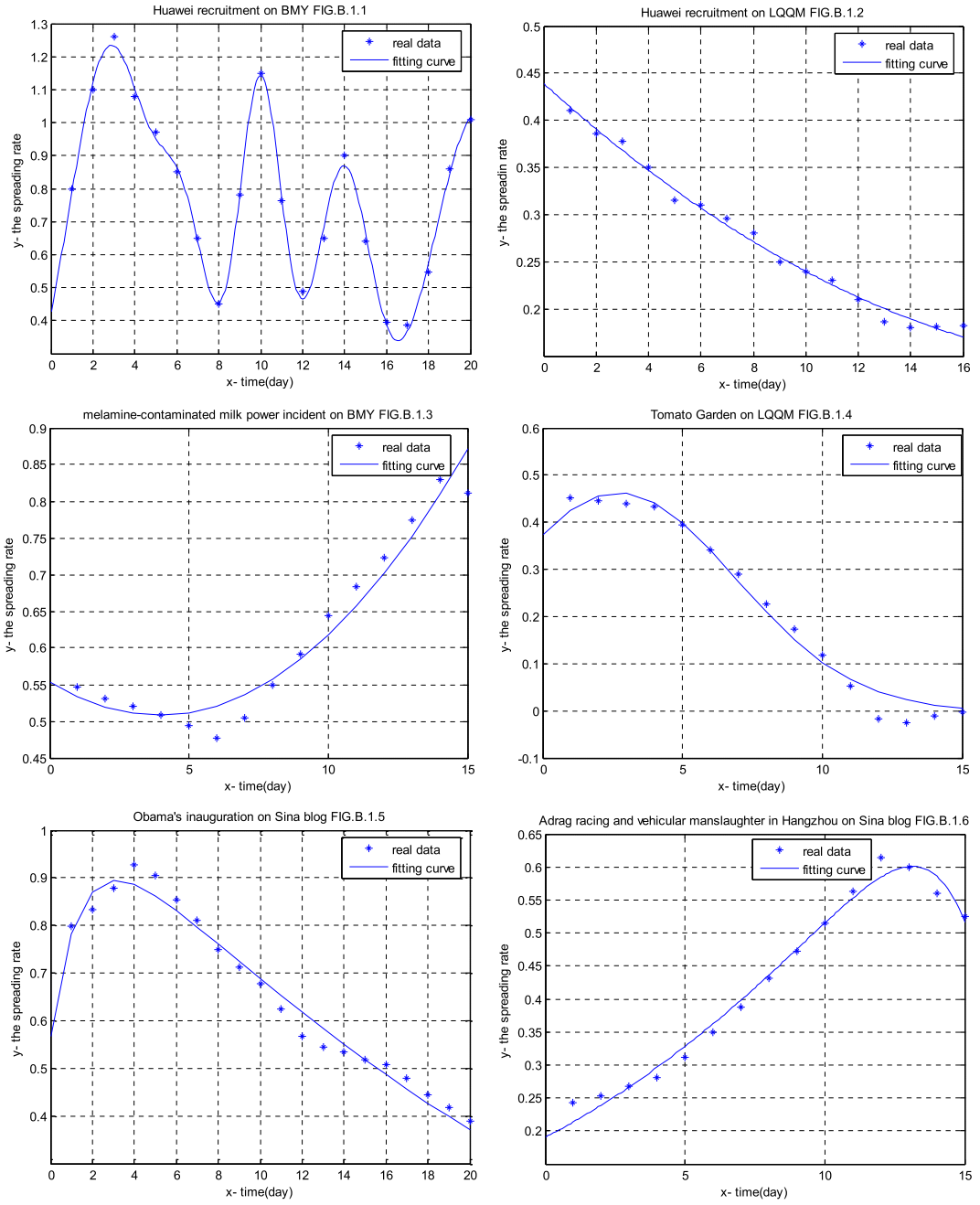

Fig. 6.2. Estimation on spreading rate: we compare real data with the fitting distribution about the spreading rates for hot online topics "Huawei recruitment" on the BMY BBS and the LQQM BBS, "melamine-contaminated milk powder incident" on the BMY BBS, "Tomato Garden" on the LQQM BBS, and "Obama inauguration" and "Drag racing and vehicular manslaughter in Hangzhou" on the Sina blog. The *x*-axis denotes time, and the *y*-axis denotes the spreading rate. Star (*) represents de-noised raw data.

(3) Fig. 6.3 illustrates the comparison on the input rate of the discussion group between the real data and fitting distribution for all hot topics in Sina blog or on BMY BBS and LQQM BBS. When the noises in raw data are filtered, the internal regularity of de-noised data is disclosed. By the least square method, we find out satisfied distribution based on the least-residual in the step 6 of method (I).

(4) Fig. 6.4 shows the comparison on the input rate of the related group between the real data and fitting distribution for hot topic-"Huawei recruitment" on LQQM BBS. And Fig. 6.5 illustrates the comparison on the output rate of the related group for hot topic-"Huawei recruitment" on LQQM BBS. When the noises in raw data are filtered, the internal regularity of de-noised data is disclosed. By the least square method, we obtain the fitting distribution based on the least-residual in the step 6 of method (I).

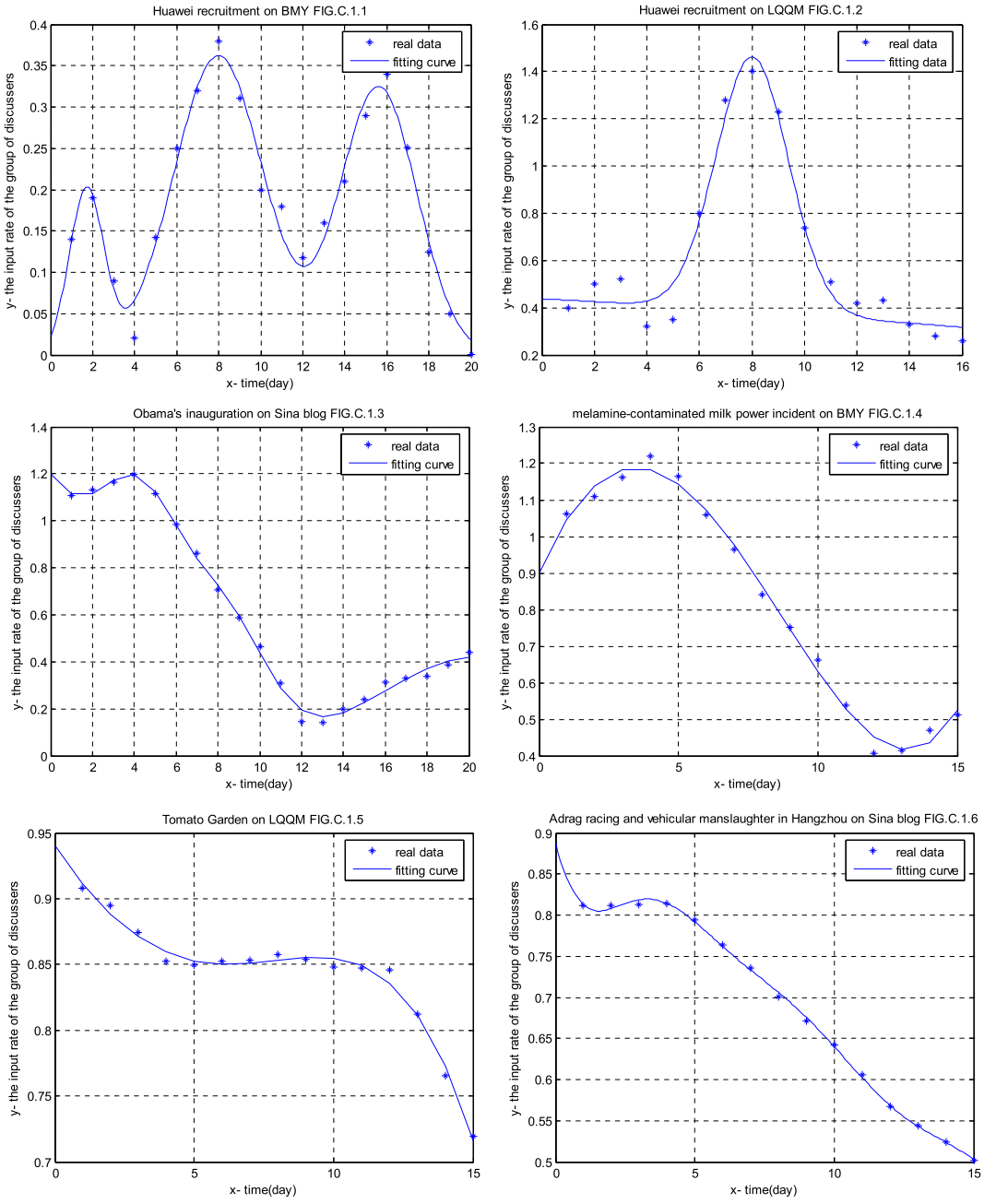

Fig. 6.3. Estimation on input rate of discussion group: we compare real data for the input rate of the discussion group with the fitting distribution for hot online topics "Huawei recruitment" on the BMY BBS and the LQQM BBS, "melamine-contaminated milk powder incident" on the BMY BBS, "Tomato Garden" on the LQQM BBS, and "Obama inauguration" and "Drag racing and vehicular manslaughter in Hangzhou" on the Sina blog. The *x*-axis denotes time, and the *y*-axis denotes the input rate of the group of discussers. A star (*) represents the de-noised raw data.

(5) We check the comparison with real data and prediction on the size of discusser in blog network or BBS sites in Fig. 6.6 for all hot topics based on the step 7 of method (I). From the following figures, we know when the size of the discusser is larger; prediction result is more precise, because some results about hot topics show our method's efficiency. In Fig. F.6, there are two peaks, actually, we choose sample data for 20 days, and so if we choose sample data for fewer 18 days, the accuracy of the result is not too high. We may not find out the second peak.

Prediction results reveal that method (I) based on our model is valid and reliable for single-peak hot topic. Especially, when the size of discusser of target hot topic is really larger, our prediction's accuracy of method (I) is better, and the results also show that our model can describe the user's state transition of hot online topic efficiently.

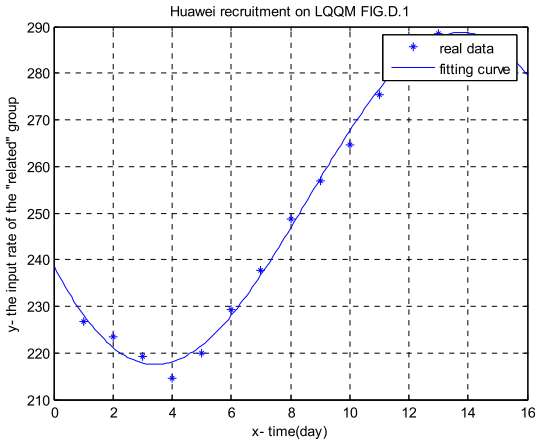

Fig. 6.4. Estimation on input rate of related group: we compare real data with the fitting distribution for the input rate of the related group.

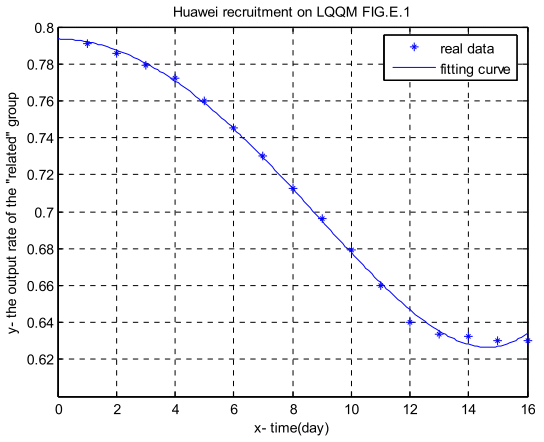

Fig. 6.5. Estimation on output rate of related group: we compare real data with the fitting distribution for the output rate of the related group for the hot topic "Huawei recruitment" on the LQQM BBS. The *x*-axis denotes fitting time. A star (*) indicates de-noised raw data.

We take the topic "Huawei recruitment" on the LQQM BBS for example to specify how to predict the trend of hot online topic discussion by threshold.

The threshold is derived from either $Q^* = (\beta(\sup\ (\max_t((\gamma - B(t))/\beta(t), A(t)/d(t)))))^*/(\gamma - B)_*$ or $Q_* = (\beta(\inf\ (\min_t(\gamma - B(t)/\beta(t), A(t)/d(t)))))_*/(\gamma - B)^*$.

Factually, $(\beta(\sup\ (\max_t((\gamma - B(t))/\beta(t), A(t)/d(t)))))^*$ almost approximates $(\sup\ (\beta\ (t)R(t)))^*$ or $(\beta(\inf\ (\min_t((\gamma - B(t))/\beta(t), A(t)/d(t)))))^*$ approximates $(\inf\ (\beta\ (t)R(t)))^*$.

When $Q^* > 1$, the topic of discussion will die out as $t \to \infty$;

When $Q_* > 1$, the topic uniformly and weakly persists as $t \to \infty$.

For example, for the topic "Huawei recruitment" on the BMY BBS,

$$Q^* < 1$$

For the topic "Huawei recruitment" on the LQQM BBS,

$$Q^* < 1$$

The topic "Huawei recruitment" dies out on both the BMY BBS and the LQQM BBS over a relatively long period of time.

6.2. Method II: propagation prediction method of single-peak and multi-peak topic

In real network circumstances, some topics are involved in multi-peak process instead of only single peak one. Method (I) is not available for multi-peak topic. Therefore, we design method (II) based on moving time windows to predict the trend of hot online topic discussion. We test a lot of hot topics that we find out that the hot topic discussion will be stable when

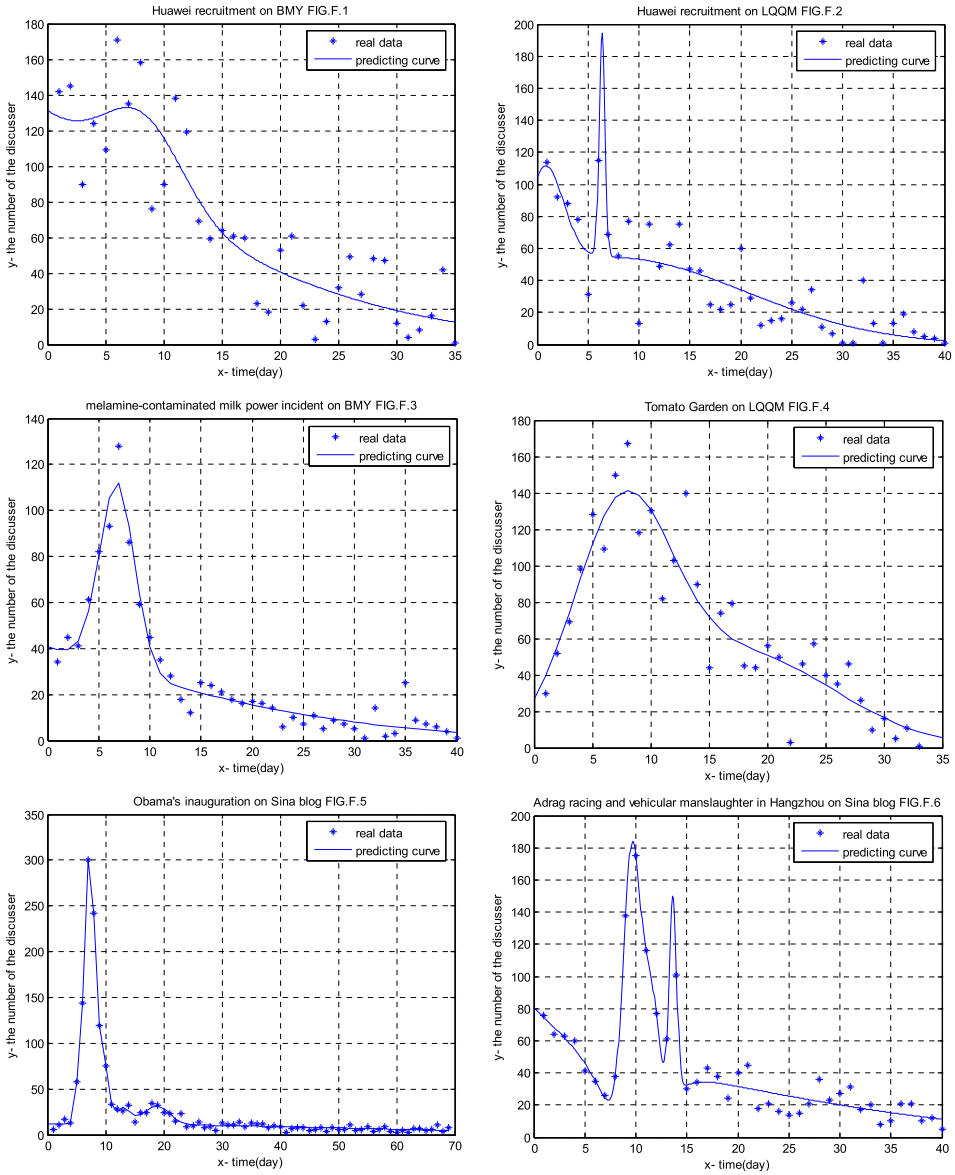

Fig. 6.6. Prediction on the number of discussers: we compare real data on the number of discussers on blog network or BBS site $D(t)$ with the prediction for the hot online topics "Huawei recruitment" on the BMY BBS and the LQQM BBS, "melamine-contaminated milk powder incident" on the BMY BBS, "Tomato Garden" on the LQQM BBS, and "Obama inauguration" and "Drag racing and vehicular manslaughter in Hangzhou" on the Sina blog. The x-axis denotes time in days, and the y-axis denotes the number of discussers.

almost all the average life span of discussers are ten days as to hot topics. So we choose moving time windows of ten days time section. We quote an instance-the hot topic "08 Olympics" on the BMY BBS to describe the procedure of method (II).

Hypothesis 1. defining a_i as the number of discussers in the hot topic "08 Olympics" on the ith day, where $a_i(i \geqslant 1)$ is determined by real data. Meantime, we use $b_j(j \geqslant 11)$ to indicate the prediction size of discussers on blog network or BBS site. a_0 is the initial value.

Step 1: First to extract raw data $a_i(i = 2, \ldots, 11)$ from 2 Jul08 to 11 Jul08 on the BMY BBS.
Step 2: Then we separate users on BMY BBS into related, discussion and exited group according to the definitions of the three subgroups discussed above.
Step 3: Select some data extracted from the BMY BBS as sampling sets to be fitted for parameters of the model.

Fig. 6.7. Prediction on the spreading rate of discussion group: we compare real data with the prediction for the spreading rate of the group of discusser.

Step 4: We estimate γ in terms of step 4 of Method (I) and the spreading rate $\beta(t)R(t)$ of a hot online topic in system (2.10) on the basis of step 5 of Method (I) regardless the de-noising processing. Then we predict the value γ and $\beta(t)R(t)$ of the following two days by the fitting distribution of γ and $\beta(t)R(t)$ of the previous ten days.

Step 5: According to the least squares method, we estimate the input rate of the discussion group $B(t)$ in system (2.10) just like the way in step 6 of method (I) for the hot online topic. Then we predict the value of $B(t)$ of the following two days by the fitting distribution of $B(t)$ of the previous ten days similarly.

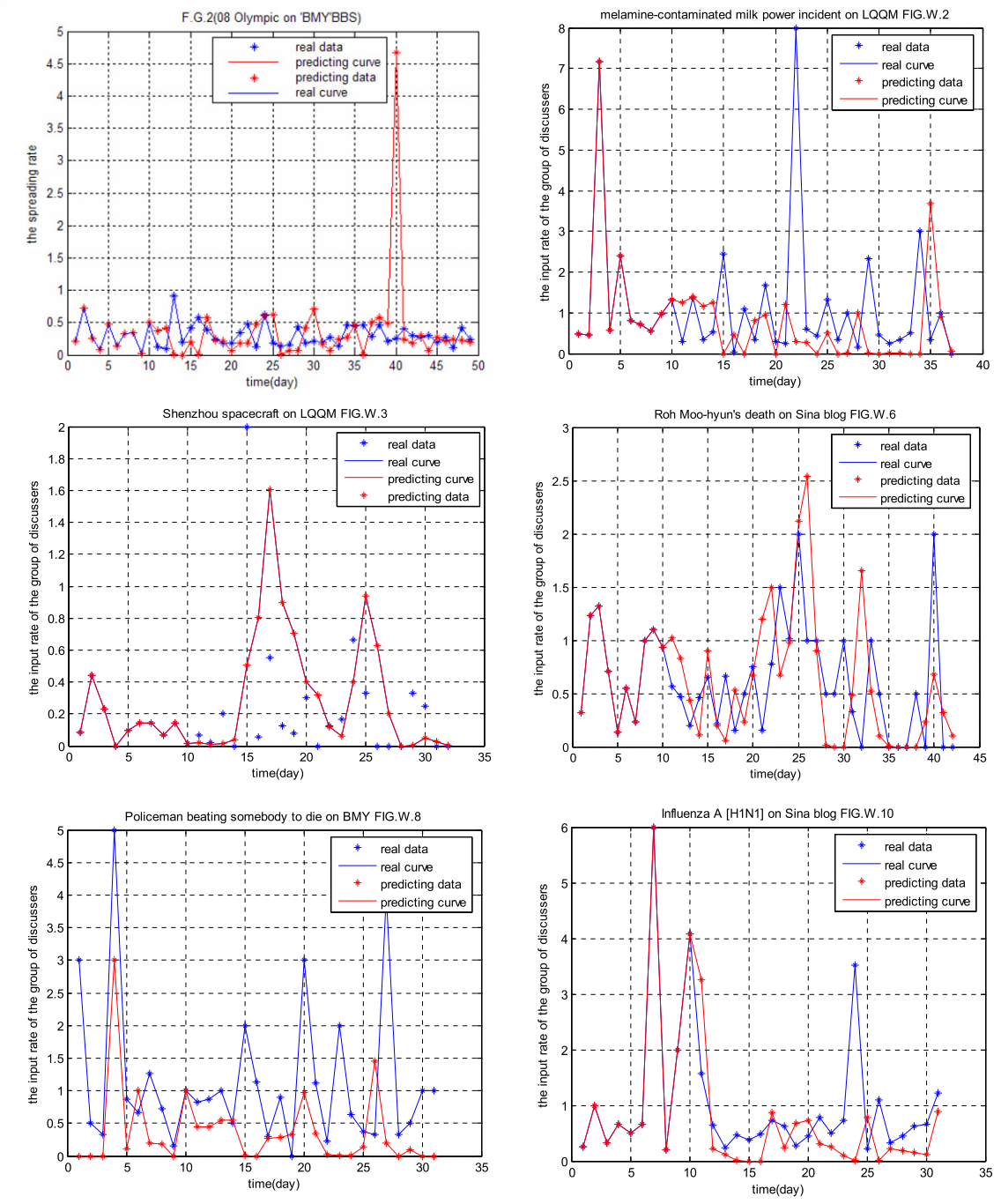

Fig. 6.8. Prediction on the input rate of discussion group: we compare real data with the prediction for the input rate of the group of discussers for the hot topic "08 Olympic" and "Policeman beating somebody to death" on the BMY BBS, "Shenzhou spacecraft" and "melamine-contaminated milk power incident" on the LQQM BBS, and "Roh Moo-Hyun's death" and "Influenza A[H1N1]" in the Sina blog.

Step 6: After finishing estimation of $B(t)$, γ and $\beta(t)R(t)$, we obtain their fitting distributions and employ our model to predict the size of discussers b_{12} on 12 Jul08 and b_{13} on 13 Jul08.

Step 7: By repeating procedure from step 1 to step 6 iteratively, then we can obtain b_{14} and b_{15}. Without loss of generality, by extracting $a_{2i}, a_{2i+1}, \ldots, a_{2i+10}$ $(i > 1)$ from the 2ith day to the $(2i + 10)$th day and taking a_{2i-1} as the new, then we could predict the size of discussers b_{2i+11} and b_{2i+12} of the following two days in sequence.

Step 8: Finally, we verify whether the real data are consistent with the prediction or not.

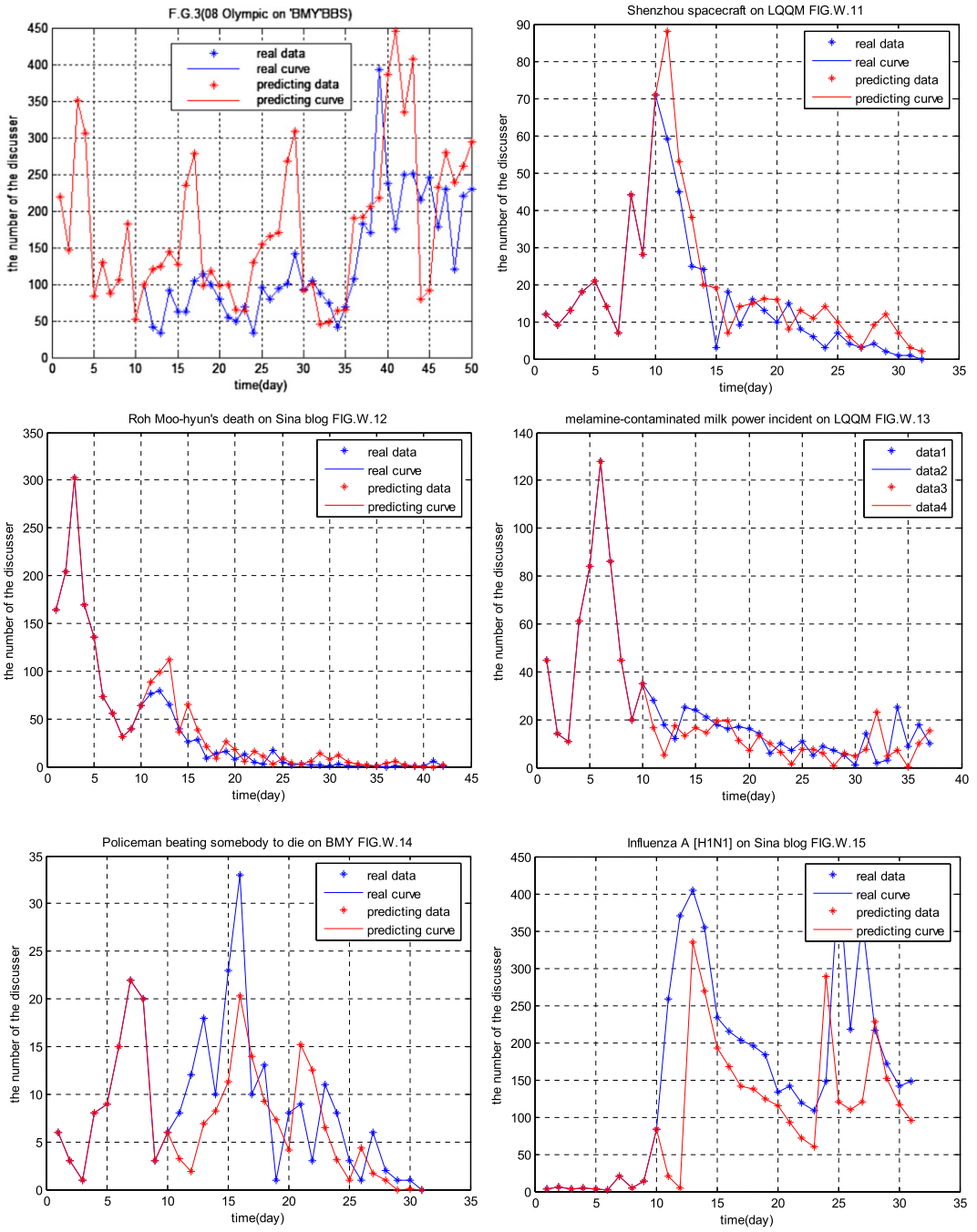

Fig. 6.9. Prediction on the number of discussers: we show comparison of real data with the prediction based on method (II) with respect to the number of discusser for the hot topics "Policeman beating somebody to death" on the BMY BBS, "Shenzhou spacecraft" and "melamine contaminated milk powder incident" on the LQQM BBS, and "Roh Moo-Hyun's death" on Sina blog and "Influenza A[H1N1]" on the Sina blog.

(1) Fig. 6.8 shows the comparison on the input rate of the discussion group $B(t)$ between the real data and prediction for all hot topics on Sina blog or BMY BBS and LQQM BBS, and Fig. 6.7 shows similar result for the spreading rate $\beta(t)R(t)$. The prediction result of the input rate of the discussion group and the spreading rate may not be satisfied respectively, but they are mutual offset for predicting the number of discusser for target hot topic T_D.

(2) We demonstrate the comparison with real data and prediction on the size of discusser in Fig. 6.9 for all hot topics. The prediction results reveal that method (II) is capable of anti-noise.

7. Conclusions and future work

The propagation of hot online topic is correlative with the collective behavior of different user's groups on blog network or BBS site. Thus by constructing a dynamic propagation model which is time-varying state equations of different user's group just like epidemic modeling, we can approximate to actual hot-topic propagation process on blog network or BBS site. This time-varying dynamic model (THTPM) is first proposed to describe the propagation of hot online topic. Furthermore, Theorems 5.1 and 5.2 in this paper contribute to theoretical research on social opinions, particularly online social opinions and are instructive for describing and predicting trends in topic discussion in real online environments. For predicting the trend of hot online topic discussion, we design method (I) to predict the trend of single-peak hot topic for theoretical motivation based on the threshold of Theorem 5.1 or Theorem 5.2 and method (II) to predict the trend of both single-peak and multi- peak hot topic for applicable motivation. Although method (II) only can predict the number of discussers- users who write or comment upon article posts about hot online topic in the following two days, the method works well for prediction precision and it can automatically make iterative prediction by previous history data. And method (II) adjusts to hot topics without considering their type as well as the number of hot-topic peak, until now we have not seen results comparable to or better than those obtained by method (II).

Moreover, by method (II) we can predict trend of hot online topic across different forums without requiring any empirical parameters. During the development of a single-peak hot online topic, by method (I) we can predict the trend in the hot online topic discussion well.

Experimental results indicate that the dynamic model indeed can describe the state transition process of related, discussion and exited group on different online social media, and our methods do not depend on any empirical parameters, and can be successfully applied to different kinds of hot online topic.

There are two specific directions for our future work. On the one hand, we will investigate the effectiveness of our method in the online forum sites other than "blog network" and "BBS site". On the other hand, we plan to search a way to model the propagation of hot online topic with considering both collective behavior and individual behavior of leader in the discussion group in the future work.

Appendix A

Theorem 5.1

Proof. Note that

$$Q^* = \frac{\left(\beta\left(\sup\left(\max_t\left(\frac{\gamma-B(t)}{\beta(t)},\frac{A(t)}{d(t)}\right)\right)\right)\right)^*}{(\gamma-B)_*} < 1 \Rightarrow \left(\beta\left(\sup\left(\max_t\left(\frac{\gamma-B(t)}{\beta(t)},\frac{A(t)}{d(t)}\right)\right)\right)\right)^* < (\gamma-B)_*,$$

while $0 < g \leqslant R(t) \leqslant \sup(\max_t(R^+(t), R_0(t))) \leqslant \sup(\max_t(\gamma - B(t)/\beta(t), A(t)/d(t)))$..

Hence, $\exists r > 0$,

$$\exists: \frac{1}{t}\int_0^t \beta\left(\sup\left(\max_t\left(\frac{\gamma-B(t)}{\beta(t)},\frac{A(t)}{d(t)}\right)\right)\right)ds + \frac{1}{t}\int_0^t (B-\gamma)ds \leqslant -\varepsilon, \forall t \geqslant r$$

$$\frac{1}{t}\int_0^t \beta S ds + \frac{1}{t}\int_0^t (B-\gamma)ds \leqslant \frac{1}{t}\int_0^t \beta\left(\sup\left(\max_t\left(\frac{\gamma-B(t)}{\beta(t)},\frac{A(t)}{d(t)}\right)\right)\right)ds + \frac{1}{t}\int_0^t (B-\gamma)ds$$

$$\frac{1}{t}\int_0^t \beta S ds + \frac{1}{t}\int_0^t (B-\gamma)ds \leqslant -\varepsilon, \forall t \geqslant r$$

From the second equation of system (2.10),

$$\frac{1}{t}\ln\frac{D(t)}{D(0)} = \frac{1}{t}\int_0^t \beta S ds + \frac{1}{t}\int_0^t (B-\gamma)ds, \forall t \geqslant r > 0 \tag{2.11}$$

Therefore,

$$\frac{1}{t}\ln\frac{D(t)}{D(0)} \leqslant -\varepsilon, \forall t \geqslant r$$

and

$$D(t) \leqslant D(0)e^{-\varepsilon t}, \forall t \geqslant r$$

So, if $t \to \infty$, then $\lim_{t\to\infty} D(t) = 0$. □

Appendix B

Theorem 5.2

Proof. By

$$Q_* = \frac{\left(\beta\left(\inf\left(\min_t\left(\frac{\gamma-B(t)}{\beta(t)},\frac{A(t)}{d(t)}\right)\right)\right)\right)_*}{(\gamma-B)^*} > 1 \Rightarrow \left(\beta\left(\inf\left(\min_t\left(\frac{\gamma-B(t)}{\beta(t)},\frac{A(t)}{d(t)}\right)\right)\right)\right)_* > (\gamma-B)^*,$$

where $(\gamma-B)^* = \lim_{t\to\infty}\sup\left(\frac{1}{t}\int_0^t(\gamma-B)ds\right)$.

$$\left(\beta\left(\inf\left(\min_t\left(\frac{\gamma-B(t)}{\beta(t)},\frac{A(t)}{d(t)}\right)\right)\right)\right)^* = \lim_{t\to\infty}\inf\left(\frac{1}{t}\int_0^t\beta\left(\inf\left(\min_t\left(\frac{\gamma-B(t)}{\beta(t)},\frac{A(t)}{d(t)}\right)\right)\right)ds\right).$$

We know that $\exists r>0$ for $\forall t\geqslant r>0$ and $R\geqslant\inf\left(\min_t\left(\frac{\gamma-B(t)}{\beta(t)},\frac{A(t)}{d(t)}\right)\right)$.

Suppose that for every $\varepsilon>0$, there are some solutions such that $D^\infty=\lim_{t\to\infty}D(t)<\varepsilon$.

Because $R\geqslant\inf(\min_t((\gamma-B(t))/\beta(t),A(t)/d(t)))$, $0\leqslant D^\infty<\varepsilon$, and the second equation of system (2.10), then

$$\frac{d}{dt}\ln D=\beta(t)R-(\gamma-B(t))\geqslant\beta(t)\left(\inf\left(\min_t\left(\frac{\gamma-B(t)}{\beta(t)},\frac{A(t)}{d(t)}\right)\right)-\varepsilon\right)-(\gamma-B(t))$$

For $t>1$,

$$\frac{1}{t}\frac{d}{dt}\ln\frac{D(t)}{D(0)}\geqslant\frac{1}{t}\int_0^t\left(\inf\left(\min_t\left(\frac{\gamma-B(s)}{\beta(s)},\frac{A(s)}{d(s)}\right)\right)\right)ds-\frac{1}{t}\int_0^t(\gamma-B(s))ds-\frac{1}{t}\int_0^t\beta(s)\varepsilon ds$$

and

$$\frac{1}{t}\frac{d}{dt}\ln\frac{D(t)}{D(0)}\geqslant\left(\beta\left(\inf\left(\min_t\left(\frac{\gamma-B(t)}{\beta(t)},\frac{A(t)}{d(t)}\right)\right)\right)\right)_*-(\gamma-B)^*-\varepsilon\beta^*-\varepsilon$$

Because $Q_*>1$, then $\left(\beta\left(\inf\left(\min_t\left(\frac{\gamma-B(t)}{\beta(t)},\frac{A(t)}{d(t)}\right)\right)\right)\right)_*-(\gamma-B)^*>0$ for sufficiently large time t and $\delta(\varepsilon)>0$, provided that $\varepsilon>0$ is chosen to be small enough. Thus,

$$\frac{1}{t}\frac{d}{dt}\ln\frac{D(t)}{D(0)}\geqslant\delta(\varepsilon),\quad D(t)\geqslant D(0)e^{\delta(\varepsilon)t}.$$

For sufficiently large t and $\lim_{t\to\infty}D(t)=\infty$, we know D is unbounded which is contradicting to the condition that D is bounded. □

References

[1] N.J.T. Bailey, The Mathematical theory of Infectious Diseases and its Applications, Griffin, London, 1975.
[2] L.B. Cao, C.Q. Zhang, Y.C. Zhao, Philip S. Yu, G. Williams, DDDM2007: Domain driven data mining, in: ACM SIGKDD Explorations Newsletter vol. 9 (2), Special issue on visual analytics Workshop Session: KDD 2007 reports: KDD Cup and workshops 84–86, 2007.
[3] M.D. Choudhury, H. Sundaram, A. John, D.D. Seligmann, Multi-scale characterization of social network dynamics in the blogosphere, in: Proceeding of the 17th ACM conference on Information and Knowledge Management, October 26–30, 2008, Napa Valley, California, USA.
[4] K. Denecke, W. Nejdl, How valuable is medical social media data? content analysis of the medical web, Information Sciences 179 (12) (2009) 1870–1880.
[5] O. Diekmann, J. Heesterbeek, Mathematical Epidemiology of Infectious disease, Wiley Series in Mathematical and Computational Biology Chichester, Wiley, 2000.
[6] F. Ginter, H. Suominen, S. Pyysalo, T. Salakoski, Combining hidden Markov models and latent semantic analysis for topic segmentation and labeling: Method and clinical application, International Journal of Medical Informatics 78 (12) (2009). e1–e6, Mining of Clinical and Biomedical Text and Data Special Issue.
[7] D. Gruhl, R. Guha, D. Liben-Nowell, A. Tomkins, Information diffusion through blogspace, in: Proceedings of the 13th International Conference on World Wide Web, 2004, pp. 491–501.
[8] H. Hethcote, The mathematics of infectious diseases, SIAM Review 42 (2000) 599–653.
[9] T. Hironori, L.V. Subramaniam, T. Nasukawa, S. Roy, Getting insights from the voices of customers: Conversation mining at a contact center, Information Sciences 179 (11) (2009) 1584–1591.
[10] V. Hristidis, O. Valdivia, M. Vlachos, et al, Information discovery across multiple streams, Information Sciences 179 (2009) 3268–3285.
[11] W.H. Hsu, A.L. King, M.S.R. Paradesi, T. Pydimarri, T. Weninger, Collaborative and structural recommendation of friends using weblog-based social network analysis, in: Proceedings of Computational Approaches to Analyzing Weblogs – AAAI 2006 Technical Report SS-06-03, Stanford, CA, 2006, pp. 55–60.
[12] S. Kleanthous, V. Dimitrova, Detecting changes over time in a knowledge sharing community, in: Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, September 15–18, 2009, pp.100–107.
[13] R. Kumar, J. Novak, P. Raghavan, A. Tomkins, On the bursty evolution of blog space, in: WWW'03: Proceedings of the 12th International Conference on World Wide Web, 2003, pp. 568–576.
[14] Y.R. Lin, H. Sundaram, Y. Chi, J. Tatemura, B.L. Tseng, Blog community discovery and evolution based on mutual awareness expansion, in: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, 2007, pp. 48–56.
[15] Y.R. Lin, Y. Chi, S.H. Zhu, H. Sundaram, B.L. Tseng, Analyzing communities and their evolutions in dynamic social networks, ACM Transactions on Knowledge Discovery from Data (TKDD) 3 (2) (2009) 1–31.

[16] Q. Mei, X. Ling, M. Wondra, H. Su, C. Zhai, Topic sentiment mixture: modeling facets and opinions in weblogs, in: WWW'07: Proceedings of the 16th International Conference on World Wide Web, 2007, pp. 171–180.
[17] M.E. Newman, J. Forrest, Stephanie, Balthrop, Justin, Email networks and the spread of computer viruses, Physics Review E 66 (3) (2002) 035101. R.
[18] X. Ni, G.R. Xue, Y. Yu, Q. Yang, Exploring in the weblog space by detecting informative and affective articles, in: WWW'07: Proceedings of the 16th International Conference on World Wide Web, 2007, pp. 181–190.
[19] G.M. Ochieng, F.A.O. Otieno, Data-based mechanistic modeling of stochastic rainfall-flow processes by state dependent parameter estimation, Environmental Modeling and Software 24 (2) (2009) 279–284.
[20] T. O'Reilly, What is Web 2.0- design patterns and business models for the next generation of software, Communications and Strategies (1) (2007) 17. First Quarter.
[21] S.G. Ruan, W.D. Wang, Dynamical behavior of an epidemic model with a nonlinear incidence rate, Journal of Differential Equations 188 (1) (2003) 135–163.
[22] Y. Sekiguchi, H. Kawahima, H. Okuda, M. Oku, Topic detection from blog documents using users' interests, in: Proceedings of the Seventh International Conference on Mobile Data Management, 2006.
[23] H. Sundaram, Making sense of meaning: leveraging social processes to understand media semantics, in: Proceedings of the 2009 IEEE International Conference on Multimedia and Expo, 1648–1651, June 28–July 03, 2009, New York, NY, USA.
[24] S.K. Tanbeer, C.F. Ahmed, B.S. Jeong, Y.K. Lee, Sliding window-based frequent pattern mining over data streams, Information Sciences 179 (2009) 3843–3865.
[25] Y.L. Tang, W.G. Li, Global analysis of an epidemic model with a constant removal rate, Mathematical and Computer Modelling 45 (7–8) (2007) 834–843.
[26] X. Wang, C. Zhai, X. Hu, R. Sproat, Mining correlated bursty topic patterns from coordinated text streams. KDD, 2007.
[27] E. Wilde, Deconstructing blogs, Online Information Review 32 (3) (2008) 401–414.
[28] L. Zhao, R.X. Yuan, X.H. Guan, M.Y. Li, Propagation modeling and analysis of incidental topics in blogosphere, Lecture Notes in Computer Science, Online Communities and Social Computing, LNCS 5621 (2009) 401–410.
[29] H.T. Zheng, B.Y. Kang, H.G. Kim, Exploiting noun phrases and semantic relationships for text document clustering, Information Sciences 179 (2009) 2249–2262.
[30] Y.D. Zhou, X.H. Guan, Z.F. Zhang, B.B. Zhang, Predicting the tendency of topic discussion on the online social networks using a dynamic probability model, in: Proceedings of the hypertext 2008 workshop on Collaboration and collective intelligence, 2008.