The Possibility of an Epidemic Meme Analogy for Web Community Population Analysis

Masao Kubo[1], Keitaro Naruse[2], Hiroshi Sato[1], Takashi Matubara[1]

[1] National Defense Academy of Japan, Dep. of Computer Science, Hashirimizu 1, Yokosuka, Kanagawa,239-8686, Japan
{masaok,hsato,matubara}@nda.ac.jp
[2] Univ. of Aizu, Dep. of Computer Software, Aizu-Wakamatsu, Fukushima-ken, 965-8580, Japan
naruse@u-aizu.ac.jp

Abstract. The aim of this paper is to discuss the possibility of understanding human social interaction in web communities by analogy with a disease propagation model from epidemiology. When an article is submitted by an individual to a social web service, it is potentially influenced by other participants. The submission sometimes starts a long and argumentative chain of articles, but often does not. This complex behavior makes management of server resources difficult and a more theoretical methodology is required. This paper tries to express these complex human dynamics by analogy with infection by a virus. In this first report, by fitting an epidemiological model to Bulletin Board System (BBS) logs in terms of a numerical triple, we show that the analogy is reasonable and beneficial because the analogy can estimate the community size despite the submitter's information alone being observable.

Keywords: Kermack–McKendrick models, SIR, BBS, SNS, Web Mining.

1 Introduction

Service industry is rapidly growing. Abe of Fujitsu [1] says, ''As a result, various services that were previously processed, handled, and separated as merely mass services have now become possible to provide as individually targeted and personalized services''. The shift of this quality of service requires more scientific approach to service system. For example, in these service systems, it is difficult to manage the system resource based on the intuition and experience because a huge number of different services have to be involved.

We believe that resource allocation for Social Networking Service (SNS), public Bulletin Board Systems (BBSs), and other social applications is a service that should be researched, as stated above. For a server manager who must allocate resources adequately to provide comfortable services, estimation of the number of accesses in the near future is an important problem. In this paper, we are interested in estimation

of size of access in near future when a time series data of their access log is observable because we think it makes its job easier.

This is currently performed as a regression analysis of time series data for the number of accesses. However, this method is inadequate when a web service offered belongs to Web 2.0 because customers interact with each other. For example, even a small community produces many accesses, so that the interaction among members is complex. Therefore, the population size of the community and the characteristics of its members are also important in making precise predictions.

This observation suggests using the model for an epidemic. An appropriate epidemiological model describes the dynamics in terms of three groups: Susceptible, Infected, and Recovered (SIR). The numbers in the SIR groups are described by differential equations having two constants, β and γ, which represent infection and recovery speed, respectively (details are described in Section 3). Because there are no actual contagious diseases when connected to the Internet, this approach may seem unnatural. However, consider supposing that a *meme* spreads among people who meet via web services? This is a popular idea, as we describe in the next section, but as far as we know, there is little research that uses actual logged access data. If the behavior of web service communities is described by the epidemiological model, we think this can be a powerful tool for server management, which usually depends on experience, because the population of the community is $(S + I + R)$ and the population of the three subgroups can be estimated precisely.

This paper reports on some of the initial work. First, numerical verification must be performed. It will be a weak argument if there are big gaps between the time series of the population of the epidemiological model and real server access data, even if the epidemiological model explains human behavior around web services well from a semantic viewpoint. Therefore, in this paper, by showing some curve fitting results for the actual log data of a TV-related BBS, we conclude that this approach is plausible.

This paper is organized as follows. In the next section, related work about population predictions for web services is summarized. The epidemiological model used here is explained in Section 3. In Section 4, an analogy between human behavior in a web community and the epidemiological model is proposed. Then, the parameters of the model are estimated by curve fitting in Matlab. At the end of Section 4, we discuss our next step: agent-based simulation.

2 Background

If you are managing a web site, you may want to know the number of accesses to your site in the future. Initially, the number of accesses per unit time is calculated by the methods in [2] and [6] for example, because real log data is noisy. Next, by using this smoothed time series data, the number of future server accesses is estimated.

The simplest estimation method is regression analysis. A type of distribution, such as an exponential distribution and Gumbel distribution, is assumed a priori. This is simple, but it does not consider why people are accessing the service. For further analysis, a model is required that can explain why and how people are accessing.

Gruhl et al. [4] listed two candidate models for such human behavior, innovation propagation dynamics and disease propagation. They adopted the innovation propagation model because of their objective of time-order reconstruction of SNS sites. Therefore, the latter model was not examined.

We also think the latter disease propagation model from epidemiology is as reasonable a candidate as the dynamics of knowledge propagation. If people are interested in an event, such as the press release of a new book, they will continue to search and make notes in their web pages, SNS, BBS, etc. Therefore, this will increase the chance that someone will notice the event. This seems similar to infection by a disease. Then, with the proliferation of such pages and other descriptive material, effective pages that offer infected people convincing arguments will emerge. People who meet such definitive pages will stop their search and discussion of the event. We think that the above process mimics the infection/recovery dynamics of disease propagation models. Therefore it is reasonable to ask if this propagation model is appropriate for explaining human behavior in the web community.

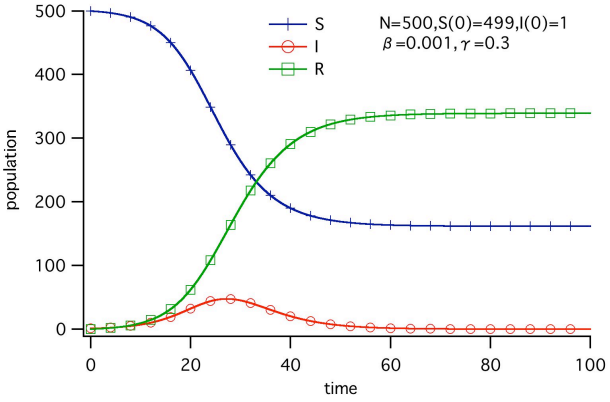

Fig. 1. An example of an SIR model. N = 500, β = 0.001, γ = 0.3, γ/β = 300.

As we noted in Section 1, this idea is not new. In particular, as in the survey in [4], there is much research from the network analysis point of view. For example, the network of migration is important for the accurate prediction of an outbreak of disease when people go to a city centre to work and then go back home. In the case of sexually transmitted diseases, the network of sexual activity is very important for suppression of the outbreak, because the number of sexual activities of a person follows a power law. However, as far as we know, there is no work that applies an epidemiological model based on network theory to actual communication data in the Internet community, except for computer virus analysis.

3 The Disease Propagation Model in Epidemiology

3.1 The Kermack–McKendrick model

There are many disease propagation models in epidemiology because of the different propagation conditions for diseases. However, the Kermack–McKendrick model (1927) and the Reed–Frost model (1928) [5] are much simpler and more general than others.

The Kermack–McKendrick SIR model gives the differential equations for a deterministic general epidemic [7]. Let

$$S + I + R = N, \tag{1}$$

where S, I, and R are the number of Susceptible, Infected, and Recovered people, and N is constant. S(t), I(t), and R(t) are represented as follows:

$$dS/dt = -\beta S \cdot I, \tag{2}$$

$$dI/dt = \beta S \cdot I - \gamma I, \tag{3}$$

$$dR/dt = \gamma I, \tag{4}$$

where β is the infection probability and γ is the recovery probability. Clearly, there is no direct transition from S to R. From the epidemiology point of view, β is the number of people who are infected by a patient and it is necessary for $1/\gamma$ unit time on average until a disease is cured [10]. Usually, in the case of person-to-person disease infection, these parameters are estimated statistically. When β and γ are available, the important information is given as follows. All the members of S are not always infected (see Fig. 1). The condition that an epidemic ends is given by $dI/dt = 0$. The solution is

$$I = 0 \text{ or } S = \gamma/\beta. \tag{5}$$

In addition, as $R(0) = 0$, the number of people who were not contagious, $S(\infty)$, is satisfied as follows:

$$S(\infty) = S(0)\exp(-(N - S(\infty))/(\gamma/\beta)). \tag{6}$$

3.2 The network and epidemic model

Research using epidemiological models based on networks usually assumes a degree distribution. In our case, the propagation path for each member of the community is usually unknown and different. Therefore, for simplicity, we assume a fully mixed model [9].

4 The Analogy and Experiments

4.1 Data

Compared with 10 years ago, it is now easy to collect data on the behavior of human groups because of the development of web-crawling agent technology and social networking services. There are various communities with different cultures and subjects of interest. We think that it is important to choose the largest communities possible, to maximize the generality of this discussion.

Therefore, we looked to BBSs (aka Internet forums) such as Slashdot and Google Groups, for data to analyze. As is widely known, BBSs on the Internet are social network services that offer the chance of communication and discussion only. When an individual submits an article to a BBS, it is influenced by other users of the BBS.

We chose the biggest Japanese open anonymous BBS, "2 channel" (http://www.2ch.net/). This BBS includes more than 600 categories, and each category contains from 100 to 500 threads. It processes more than 100 million page views per day. Anyone, without special privileges, can access the same data that we acquired, and this site is frequently used by other web researchers in Japan. We would expect that this is the most widely accessed BBS under present conditions.

For time series analysis using the SIR model, the start time is important. However, this BBS is available 365 days per year, and anyone can post an opinion at any time. In fact, it is difficult to specify when a discussion actually starts. Gruhl [3][4] identified two behavior types for SNSs by their cause: *spike* or *chat*. Spike refers to bursty behavior by events outside the community. By contrast, chat is a burst based on a conversation within the community. For our purposes, we would like to pick a spike at the time when a known event happens. Therefore, we adopted a TV program and its related BBS where participants talk to each other even outside the broadcast time. However, we know beforehand when the maximum external stimulus will happen, from newspaper information. In addition, we can suppose that there will be no spike following the broadcast.

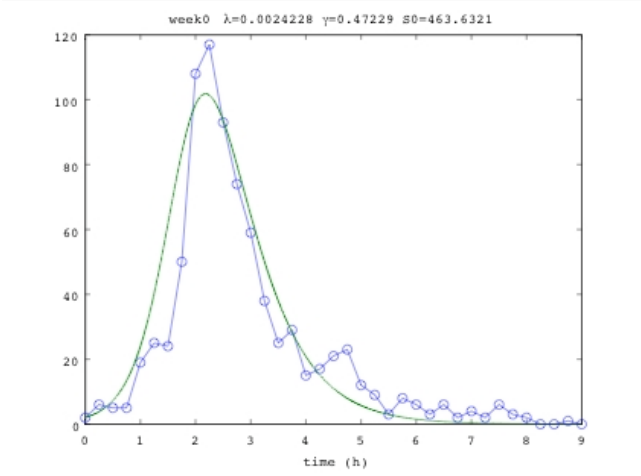

Fig. 2. Posting data for the BBS (Jan. 10 to Jan. 11, 2007) and the fitting result for the SIR model.

4.2 The proposed interpretation of BBS data by the SIR model

Here, we assume that "Susceptible" means a person who is interested in the TV program. A person who has so strong an opinion as to post to the BBS corresponds to "Infected". A "Recovered" person leaves the BBS, being no longer interested in the topic. Therefore, we aim to minimize the RMS difference between the log data of the number of posters and the "Infected" group, as calculated using (3).

4.3 Fitting and results

We counted the number of posters every 15 minutes from 9 pm, Jan. 10, 2007 to 6 am, Jan. 11, 2007. It was 10 pm when the TV program broadcast started, and it finished at 11 pm. This TV program was so famous and general as to be watched by over 18% of households in Japan. The zigzag line of Fig. 2 indicates the logged data. The x-axis indicates the time sequence and each tick is an hour. In this figure, the TV program starts at x = 1 and ends at x = 2. Note that there is a big burst with a small drop around x = 1.5. We think this is reasonable because posters will also want to watch the TV program!

The smoother line of Fig. 2 represents the fitting result of the SIR model using Matlab. It seems that this fits well. The resulting estimate for the triple (S(0), β, γ) is (463.6321, 0.0024228, 0.47229).

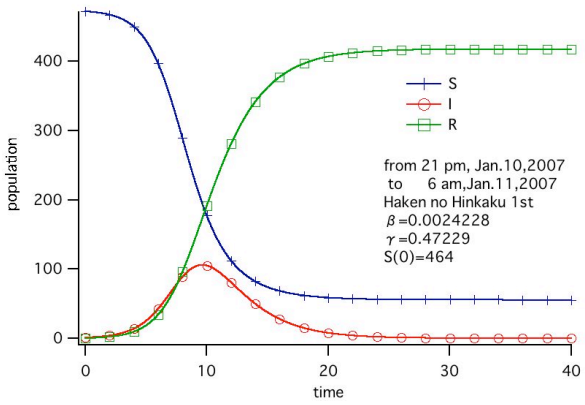

Fig. 3. The resulting behavior of SIR (Jan. 10 to Jan. 11, 2007)

Fig. 3 shows the progress of S, I, and R. The I in Fig. 3 is the same as the solid line of Fig. 2. If our assumption of Section 4.2 is valid, this offers the following insights about the community: (1) about 464 persons came to this BBS, (2) about 400 people left, and (3) 60 people still enjoy the community. This third point suggests that the broadcast and the surrounding discussion in this BBS enlarged its community by about 60 people.

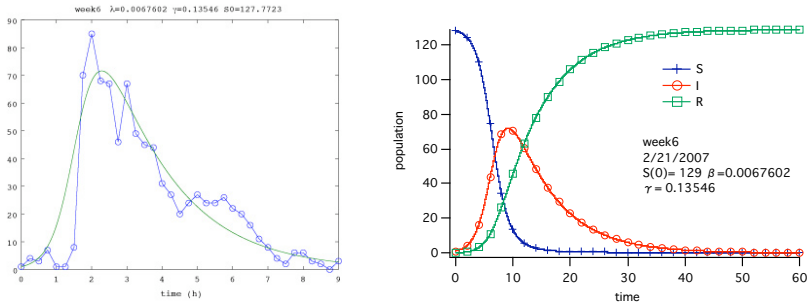

Fig. 4. The fitting results for another broadcast (Feb. 21 and Feb. 22, 2007).

Fig. 4 shows the results for another week (from 9 pm, Feb. 21, 2007). This curve also fits well to the week's data. The estimated γ is smaller than that for the earlier week, with posts continuing longer than for the Jan. 10 to 11 period.

4.4 Discussion

In this paper, we validated the disease propagation model as a model for explaining human behavior in a community. By using curve-fitting techniques, we report how reasonable this idea is.

We adopted the Kermack–McKendrick version of the disease propagation model and applied it to the posting data of a BBS. The characterizing triple (S(0), β, γ) for this model was estimated by RMS minimization and hill climbing in Matlab. As shown in Figs. 2, 3, and 4, the SIR model fitted well and gave new and insightful information.

This approach has the following attractive aspects. Firstly, this model can estimate the total community size, namely (S + I + R), which regression analysis via a statistical distribution cannot achieve. Secondly, it is easy to understand the behavior intuitively, with the propagating speed being β, and the durability of conversation seeming to be γ. We hope it will ease the management of server resources, with (5) and (6) describing the population dynamics of each community.

Note that, in Section 4.2, we proposed understanding the logged data of BBSs in terms of an epidemiological propagation analogy. That is, we assumed that people join the BBS so as to post. In actual BBS communities, there are many "lurkers", and [8] estimates the total community size including lurkers. However, we think the analogy remains reasonable because the driving force of a BBS is undoubtedly the group of people who want to post.

The unique point of this analogy is that it deals with migration among communities. S and R represent the population outside the observing BBS community. We think these estimates justify evaluating the nonlinear differential equations because this information is expensive even if you can obtain access to it.

The proof that the assumptions above are correct is very important and an urgent task for us. We think that Agent-Based Simulation (ABS) is a powerful tool for SSME [11]. As is well known, one definition of "Engineering" is that it is a methodology for obtaining desirable results. However, in a service industry, it is sometimes difficult to test a new method. In such cases, the agent simulation approach is one that is both practicable and meaningful. It is possible for this bottom-up simulation methodology to use a set of programs that behave like participants using the service. As tastes vary, we could use a variety of agents, with carefully chosen parameters. In our case, verification of posting article behavior of clients is required because this component is out of focus of any disease propagation model.

5 Conclusion

In this paper, we have proposed a new approach to understanding the behavior of the Internet community by analogy with a disease propagation model from epidemiology. The SIR of the Kermack–McKendrick model was applied to data comprising the number of posts per 15 minutes to a BBS. The characterizing (S(0), β, γ) of this model was estimated by RMS minimization and hill climbing in Matlab. This new interpretation fits well, and we can say that the analogy is a promising approach that gives new and insightful information, namely:

(1) A response from a big event for community is represented by infection speed, recovery speed, and initial community size.

(2) This framework offers information about community migration. In particular, the total community size is highly valuable information for server managers of web service sites, who have to allocate resources.

References

1. Abe T.:What is Service Science Dec-05, FRI Research Report, No.246 (2005)
2. Fujiki, T., Nanno, T., Suzuki, Y., Okumura, M.: Identification of Bursts in a Document Stream. In: First International Workshop on Knowledge Discovery in Data Streams (in conjunction with ECML/PKDD 2004) (2004).
3. Gruhl, D., Guha, R., Kumar, R., Novak, J., Tomkins, A.: The Predictive Power of Online Chatter. In: Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, pp. 78–87, ACM Press, New York (2005).
4. Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A.: Information Diffusion through Blogspace. In: Proceedings of the 13th International Conference on the World Wide Web, pp. 491–501, ACM Press, New York (2004).
5. Jacquez, J. A.: A Note on Chain-Binomial Models of Epidemic Spread: What is Wrong with the Reed–Frost Formulation? Mathematical Biosciences 87(1), pp. 73–82 (1987).
6. Kleinberg, J.: Bursty and Hierarchical Structure in Streams, In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2002).
7. Krebs, C. J.: Ecology, Benjamin-Cummings Publishing Company (2002).
8. Naruse K., Kubo M.: Lognormal Distribution of BBS Articles and its Social and Generative Mechanism, Web Intelligence, pp. 103–112 (2006).
9. Newman, M.E.: Spread of Epidemic Disease on Networks, Physical Review E 66, 016128 (2002).
10. Nishiura, H., Inaba, H.: Prediction of Infectious Disease Outbreaks with Particular Emphasis on the Statistical Issues Using a Transmission Model, In: Proceedings of the Institute of Statistical Mathematics Vol. 54, No. 2, pp. 461–480 (2006), (in Japanese).
11. Rahmandad, H., Sterman, J.: Heterogeneity and Network Structure in the Dynamics of Contagion: Comparing Agent-Based and Differential Equation Models, MIT Sloan School of Management, Cambridge MA 02142 (2004).