2007 International Symposium on Nonlinear Theory and its Applications
NOLTA'07, Vancouver, Canada, September 16-19, 2007

Modeling Knowledge Sharing Portal Activities using a Multiplicative Random Process with a Birth and Death Mechanism

Kenichi Arai† and Takeshi Yamada† and Yukio Hayashi‡

†NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0237 Japan
‡School of Knowledge Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1292 Japan
Email: ken@cslab.kecl.ntt.co.jp, yamada@cslab.kecl.ntt.co.jp, yhayashi@jaist.ac.jp

Abstract—We propose a new model of the evolution of the posting activities in knowledge sharing portals (KSPs). Typical KSPs include online Bulletin Board Systems, word-of-mouth and Q&A community sites. We have constructed a model based on extensive analysis using three different KSPs. First, we show that the number of posted messages obeys Gibrat's law and can be modeled as a multiplicative random process. Next, we extend the model by adding the birth and death mechanisms of the posting sequence. The proposed model can successfully reproduce the exponential age distributions of posting sequences and the power law distributions of the number of messages.

1. Introduction

In recent years, the Internet has played an important role in our daily lives in terms of acquiring knowledge and sharing it with others. In addition to using search engines to find helpful Web sites that provide necessary information directly, we can also use Q&A community sites and word of mouth sites that provide bulletin board systems (BBSs) in which we can interactively ask questions about specific topics and exchange information. We refer such sites as knowledge sharing portals (KSPs). KSPs enhance the opportunity to obtain information on less common and more specialized topics that would not be available without them. The use of KSPs is now gaining in popularity and thus giving them the potential to become core service functions on the Internet in the near future. Therefore, we believe that it is important to reveal the fundamental mechanism of posting behavior in the KSPs to improve and enhance KSP activities, namely, to increase the number of messages and members and encourage the sharing and exchange of more knowledge and information.

Several related studies have been made on the growth of bipartite graphs. Noh et al. studied the evolution of bipartite graphs focusing their attention on members and the group sizes in KSPs [9] and Ramasco et al. proposed a growth model of bipartite graphs as collaboration networks [10]. These studies did not mention the growth of messages and their models could not express the increases for a unit period but only relative growth rates of nodes and links. What seems to be locking is the growth model of messages, authors and so on from viewpoint of evolution dynamics.

In this paper, we investigate posting behavior characteristics by analyzing the time series of the number of messages posted by an author on a board during a fixed period of time. In particular, we focus on the birth and death mechanism of the time series. As a result, we found that the time series obeys Gibrat's law and that the birth and death rates are almost constant. Based on these results, we propose an evolution model based on a random multiplicative processes (RMP) with the birth and death mechanism of the processes. Finally, we show that our model can reproduce certain statistics of the posting behavior in the KSP by numerical experiments and approximation analysis.

2. KSP and collected data

KSPs provide forums, or places for interactive communication, either on the Internet or on an intranet in an organization, and these forums have several meeting rooms for discussing specific subjects, which we call boards. Participants can read and post messages to discuss certain issues, ask and answer questions, and exchange information. A participant who has posted messages is called an author.

We have collected data from three real KSPs. Each set of data consists of a list of authors, a list of boards and a collection of messages that include information about their authors, their targeted boards and the time they were posted. The first is from "Fujisawa citizens' BBS" [1]from June 1, 1999 to September 24, 2005, which is a well-known active municipal BBS where, among other things, citizens discuss regional issues, and enjoy a chat. It has 879 authors, 73 boards and 52,881 messages. The second is from "Oshiete! goo" [2], one of the largest Q & A community sites in Japan. The collected data includes 400,690 authors and 8,902,882 messages, consisting of questions and responses, from July 29, 1999 to July 20, 2006. We treat "Oshiete! goo" as one board in our analysis since the active period of a question and its all responses is short and there are few responses to one question: the rate at which a question and its all re-

[1]http://www.city.fujisawa.kanagawa.jp/˜denshi/
[2]http://oshiete.goo.ne.jp/

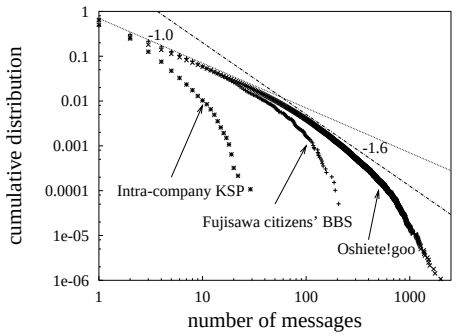

Figure 1: Distributions of number of messages for a month.

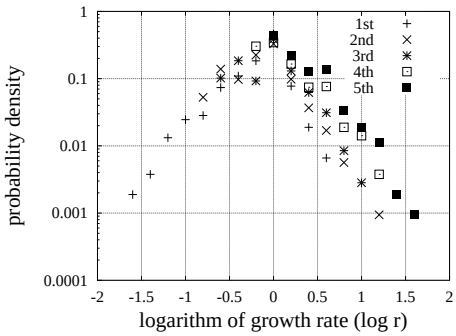

Figure 2: Dependence of growth rates on number of posted messages in previous months for Fujisawa citizens' BBS.

sponses were posted in the same month is 94.3 % and the average number of responses to a question is 3.3. The third set of data is from "Intra-company KSP" from January 9, 2004 to October 18, 2006, which is a certain company's intranet BBS where employees discuss issues of interest and exchange information about their business. There are 242 authors, 751 boards and 11,894 messages in the data.

3. Dynamics of posting sequence

We focus on the number of messages in a posting sequence (PS) to investigate the activity of a KSP in this section. Here, a PS is defined as a series of posting activities by a fixed author on a fixed board.

3.1. Number of messages and its growth rate

Let $x_{ij}(t)$ indicate the number of messages that the i-th author posted on the j-th board during the t-th month. In Figure 1, we show the distributions of $x_{ij}(t)$ for the three KSPs on a double logarithmic chart. We can see straight parts in the curves, where the distributions obey the power law. The exponents of the cumulative distribution of "Oshiete! goo" are about -1.0 and -1.6 in the left and middle regions, respectively. It follows that the distributions fundamentally obey the power law but they may also obey double Pareto distributions or log-normal distributions. Note that the rapid decreases in the right sides of the curves are due to the cutoff effects of observations.

Figure 2 shows the distributions of the growth rate $r_{ij}(t) = x_{ij}(t)/x_{ij}(t-1)$. The data are divided into five groups according to the $x_{ij}(t-1)$ values, i.e. the largest $x_{ij}(t-1)$ group, the second largest $x_{ij}(t-1)$ group and so on. The growth rate distributions for all data groups almost coincide except for the small $r_{ij}(t)$ region where some groups have no distributions. The fact that the growth rate follows the same distribution, independent of the denominator $x_{ij}(t-1)$, is known as Gibrat's law [2].

3.2. Birth and death of PS

We consider the birth and death of a PS: the starting and quitting of posting by a certain author on a certain board.

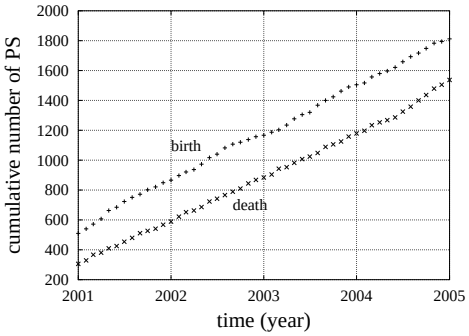

Figure 3: Time evolution of cumulative numbers of birth and death PSs for Fujisawa citizens' BBS.

Note that the birth of a PS includes the point at which an existing author starts posting messages on a board for the first time as well as the appearance of new authors and the addition of new boards. The death of a PS means that no messages are posted in the PS after that moment.

Figure 3 shows that the cumulative numbers of the births and deaths of PSs grow linearly and it follows that the numbers of births and deaths in a month are almost constant. In addition, the birth and death curves are parallel, which means that the number of living, or active, PSs remains roughly constant. Another important point is that a considerable number of PSs die and are born in a month; about 20 PSs die and about 20 are born each month while the number of active PSs in "Fujisawa citizens' BBS" is 200 or 300.

Figure 4 shows the age distributions of PSs. Here, the age of a PS is the period, or the number of months, from its birth to the present. The age distributions decay in an exponential form and the faster decreases at the lefts end of the curves are due to cutoff effects.

4. Modeling posting behavior

The following features of the KSPs were described in the previous section. (1) The number of messages posted in a month obeys the power law distribution. (2) The ratio

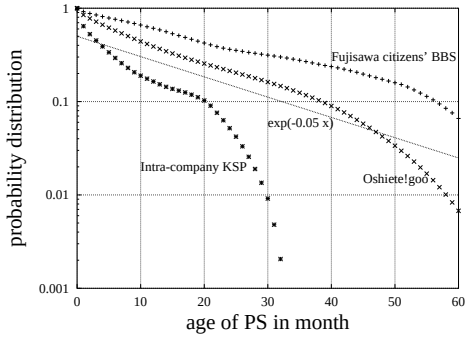

Figure 4: Age distributions in month.

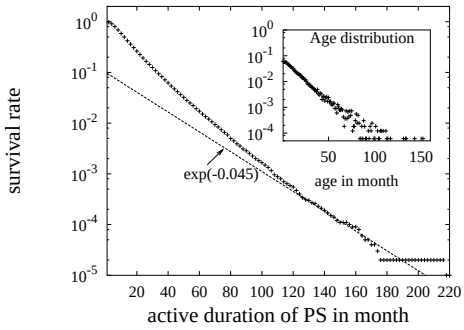

Figure 5: Evolution of number of active PSs and Age distribution in months (inset).

of the numbers of messages between consecutive months obeys Gibrat's law. (3) The numbers of births and deaths of PSs are almost constant. (4) The PS age distribution has an exponential form. In this section, we develop the model of the evolution of posting behaviors based on (1) and (2) and then we will show that the model can reproduce (1) and (4).

4.1. Random multiplicative process

It follows from Gibrat's law that the number of messages $x_{ij}(t)$ obeys a RMP with independent and identically-distributed random variables $r_{ij}(t)$, as follows:

$$x_{ij}(t+1) = r_{ij}(t)x_{ij}(t). \quad (1)$$

Suppose that $\ln r(t)$ has a mean value μ and a variance σ^2, then for large t the distribution of $\ln x(t)$ approaches a normal distribution with mean $t\mu$ and variance $t\sigma^2$ according to the central limit theorem. Therefore, $x(t)$ does not have a time-invariant distribution and it is not an appropriate model of the observed posting behavior.

In fact, the RMP is used in econophysics with modification. For example, it is known that an RMP with reflecting barriers or additive noise reproduces power law distributions [4, 6, 8]. However, as regards the model of the posting behavior, certain issues remain in that the interpretation of the barrier and the noise is unclear and their parameter values are difficult to determine. On the other hand, Reed, and Huberman and Adamic, showed that the exponential growth of the number of processes led to a power law distribution [3, 7] but the number of PSs in the KSPs does not increase exponentially. In the next subsection, we propose an RMP model that includes the birth and death mechanism of PSs and show that it can reproduce the distribution of the posting behavior in the KSPs.

4.2. Birth and death mechanism

In 3.2, we saw that the numbers of births and deaths are almost constant. Thus, as regards the birth of new PSs in our model, a constant number of new processes are added every month. In addition, when $x(t) < \theta$ (≤ 1), we treat the PS as dead. Then, the model can be described as follows.

Initialization: N_0 pairs of $\{i, j\}$ are prepared and the variable $x_{ij}(0)$ is set as the initial value x_0.

Every month

existing PS: $x_{ij}(t-1)$ is mapped into $x_{ij}(t)$ using Eq. (1).

death of PS: $\{i, j\}$ are eliminated if $x_{ij}(t) < \theta$.

birth of PS: New n pairs of $\{i, j\}$ are added and $x_{ij}(t)$ is initialized as x_0.

Numerical simulations were performed to investigate the above KSP model. Random variables $\ln r(t)$ were chosen from the normal distribution with $\mu = -0.3$ and $\sigma^2 = 1.0$ ($\mu = -0.0785$ and $\sigma^2 = 0.969$ in "Oshiete! goo"). The simulations, using parameter values of $N_0 = 100000$, $x_0 = 10.0$ and $n = 0$, shows that the fraction of active PSs after t months, namely survival rate, decays exponentially in a large t region (Figure 5). Note that the survival rate as a function of time can be interpreted as the cumulative distribution of lifetime. Our simulations also show that ages of PS are distributed in the exponential form as shown in the inset of Figure 5 and it agrees well with that of real KSPs, where the age distribution obeys the exponential function except the cutoff effect in the tails of the curves. shown in Figure 4.

The distribution $p_t(x)$ of $x(t)$ at time t is exactly the log-normal distribution as follows.

$$p_t(x) = \frac{1}{\sqrt{2\pi t}\sigma x} \exp\left\{-\frac{(\ln x - t\mu - \ln x_0)^2}{2t\sigma^2}\right\}. \quad (2)$$

Since the death condition is $x < \theta$, the probability $f(t)$ that a t months-old PS is still active is approximated [1] by

$$f(t) = \int_\theta^\infty p_t(x)dx \sim \frac{\sigma}{\sqrt{2\pi t}(-\mu)} e^{-\frac{\mu^2}{2\sigma^2}t} \sim e^{-t\mu^2/2\sigma^2}. \quad (3)$$

Therefore, the lifetime distribution shows the exponential decreases for large t with the exponent $-\mu^2/2\sigma^2$, or -0.045 for the parameter values in the simulation. This is supported by the simulation shown in Figure 5, where the life time distribution decreases exponentially with the exponent -0.045 for large t. The expected number of active PSs is $nf(\tau)$ since the number of PSs generated τ months ago is

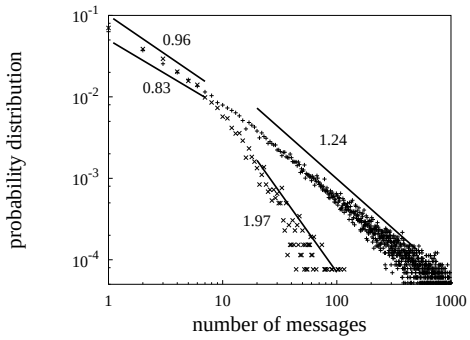

Figure 6: Distributions of number of messages (simulation) $\gamma_1 = 0.96$ and $\gamma_2 = 1.24$ for $\mu = -0.1$ and $\sigma = 1.0$. $\gamma_1 = 0.83$ and $\gamma_2 = 1.97$ for $\mu = -0.4$ and $\sigma = 0.7$.

n. It follows that the age distribution $g(\tau)$ is proportional to $f(\tau)$. Therefore, the age distribution also obeys an exponential function from Eq. (3), which is confirmed by the simulation as shown in the inset of Figure 5.

Next, we performed a simulation with $N_0 = 1000$ and $n = 1000$. Figure 6 shows the probability density distributions of the number of messages for two sets of parameter values. The distributions in this figure and Figure 1, which shows the distribution of the real KSPs, behave similarly: they obey similar power law distributions.

With the continuous approximation of time t, the distribution $h(x)$ of the number of messages is [5]

$$h(x) = \int_0^\infty g(t)p_t(x)dt \sim \begin{cases} C\left(\frac{x}{x_0}\right)^{-\gamma_1} & \text{if } x \le x_0 \\ C\left(\frac{x}{x_0}\right)^{-\gamma_2} & \text{if } x \ge x_0 \end{cases}, \quad (4)$$

$$\gamma_1 = 1 - (1 - \sqrt{2})\frac{\mu}{\sigma^2}, \quad \gamma_2 = 1 - (1 + \sqrt{2})\frac{\mu}{\sigma^2} \quad (5)$$

using the approximation $g(t) \propto e^{-t\mu^2/2\sigma^2}$. $h(x)$ become the double Pareto distribution. We can see that the exponents of the simulations (Figure 6) and the approximate calculations $\{\gamma_1 = 0.96, \gamma_2 = 1.24\}$ and $\{\gamma_1 = 0.83, \gamma_2 = 1.97\}$ from Eq.(5) are well agreed and they have approximately same values of real KSPs. Thus, the proposed model reproduces the double Pareto distribution for the number of messages. Thus, it is quite likely that the distribution of real data also obeys the double Pareto form.

5. Discussion and Summary

We performed an empirical analysis using PSs of three real KSPs, namely a citizens' BBS, a Q&A community site and an intranet BBS, and based on the results we proposed a posting behavior model. The number of messages posted by a author on a board in a month obeys Gibrat's law and it follows that its evolution is formulated by an RMP. Another important characteristic is that the numbers of births and deaths of PSs in a month are almost constant. Thus, our model has a PS elimination-addition mechanism. By numerical simulations and approximation analysis, we have shown that our model successfully reproduces the distributions of the number of messages and of the lifetimes, which agree with the distributions of real KSPs.

We focused on the births and deaths of PSs in this paper. We think that the births and deaths of elements can often be observed in such open systems. In addition, a fluctuation that is proportional to element size is common especially in complex systems. Therefore, our model can apply to phenomena commonly found in open and complex systems.

References

[1] Cody, W. J.: Rational Chebyshev Approximations for the Error Function, *Math. Comp.*, Vol. 23, No. 107, pp. 631–638 (1969).

[2] Gibrat, R.: *Les inegaliteś économiques*, Paris, Recueil Sirey (1931).

[3] Huberman, B. A. and Adamic, L. A.: Evolutionary Dynamics of the World Wide Web (1999). arXiv:cond-mat/9901071.

[4] Levy, M. and Solomon, S.: Power Laws are Logarithmic Boltzmann Laws, *International Journal of Modern Physics C*, Vol. 7, No. 4, pp. 595–601 (1996).

[5] Mitzenmacher, M.: Dynamic Models for File Sizes and Double Pareto Distributions, *Internet Math.*, Vol. 1, No. 3, pp. 305–333 (2003).

[6] Nakao, H.: Asymptotic Power Law of Moments in a Random Multiplicative Process with Weak Additive Noise, *Phys. Rev. E*, Vol. 58, pp. 1591–1600 (1998).

[7] Reed, W. J.: The Pareto Law of Incomes — an Explanation and an Extension, *Physica A*, Vol. 319, pp. 469–486 (2003).

[8] Takayasu, H., Sato, A.-H. and Takayasu, M.: Stable Infinite Variance Fluctuations in Randomly Amplified Langevin Systems, *Phys. Rev. Lett.*, Vol. 79, pp. 966–969 (1997).

[9] Noh, J. D., Jeong, H.-C., Ahn, Y.-Y. and Jeong, H.: Growing network model for community with group structure, *Phys. Rev. E*, Vol. 71, p. 036131 (2005).

[10] Ramasco, J. J., Dorogovtsev, S. N. and Pastor-Satorras, R.: Self-organization of collaboration networks, *Phys. Rev. E*, Vol. 70, p. 036106 (2004).