2007－MPS－63 (6)
2007／3／3

知識共有サイトにおける投稿数の乗算確率過程的成長モデル

新井 賢一 †, 山田 武士 †, 林 幸雄 ‡
日本電信電話株式会社 †, 北陸先端科学技術大学院大学 ‡

概要: 掲示版などの知識共有サイトにおいて，複数の実データを用いた投稿行動の解析を行い，それに基づき投稿記事数の数理的成長モデルを提案した．まず，一連の投稿行動である投稿系列において，投稿記事数の増加率がGibrat 則を満たすことから，記事数の推移を乗算確率過程として捉えられることを示した．次に，投稿系列の生成消滅が頻繁に生じるという知識共有サイトの特徴を考慮し，乗算確率過程の生成消滅のための機構を導入した新たな知識共有サイトモデルを提案した．この提案モデルは従来の乗算確率過程に較べより現実の投稿行動に即したものであり，投稿系列の投稿継続期間や投稿数分布など実データの性質をよく再現できることを示した．

Time Evolution of Knowledge Sharing Portal activities as Multiplicative Random Process

Kenichi Arai †, TAkeshi Yamada †, Yukio Hayashi ‡
NTT Corporation †, Japan Advanced Institute of Science and Technology ‡

Abstract: We propose a new evolution model of the article posting activities in the Knowledge Sharing Portal (KSP). Typical examples of KSP include online Bulletin Board System, word-of-mouth and Q&A community sites. Our model has been constructed based on extensive analysis using three different KSPs. First, we show that the number of posting obeys Gibrat's law, and can be modeled as Multiplicative Random Process. Next, we added the birth and death mechanisms of posting sequences to the model. The proposed model can successfully reproduce distributions for age of posting sequences and the number of postings.

1 はじめに

近年，情報や知識の獲得や流通に関して，インターネットは大きな役割を担っている．最近では，Q&A コミュニティサイト，口コミサイトなどの掲示板サイト（Bulletin Board System, BBS）などを利用した情報や知識の獲得や流通も活発に行われている．このようなサイトでは，直接不特定多数に疑問を投げ掛けたり情報の提供を求めるなど，対話的なコミュニケーションを通じて情報の獲得などを行うのが特徴である．このようなサイトを「知識共有サイト（Knowledge Sharing Portal, KSP）」と呼ぶことにする．知識共有サイトを用いれば，これまで入手し難かった特定かつ専門的な話題に関する情報が比較的容易に手に入るなどこれまでになかったサービスを享受することができ，今後益々重要になり発展するだろうと考えられる．

知識共有サイトにおいて知や情報の流通を効率化・活性化させ，アクティビティ（投稿数，会員数など）を維持，拡大させるためにも，投稿行動に関する基本的な知見やそのメカニズムを探ることは重要な課題である．本論文では，一定期間にある参加者がある掲示板に投稿する記事数の時系列やそれら時系列自体の生成や消滅の特徴について解析を行った．その結果，増加率が Gibrat 則を満たすことや時系列の生成消滅が頻繁かつ一定率で生じることを見い出した．これらの結果に基づき，生成消滅する乗算確率過程として知識共有サイトの投稿行動に関する数理的モデルの構築提案を行った．さらに，シミュレーションや解析により，知識共有サイトにおける実際の投稿行動を再現できることを示した．

2 知識共有サイトのデータ構成

「知識共有サイト」とは，インターネットなどの上のサービスであり，参加者による議論，情報交換などのコミュニケーションの場として用いられているものである．通常これらのサイトは，特定のタイトルやテーマなどの話題が設定された複数の掲示板から構成される．参加者は話題に沿った内容の文章などの記事を掲示板に投稿する．参加者はシステムにより一覧表示された記事を基にして，記事を投稿することができる．つまり，知識共有サイトモデルの構成要素としては，参加者，掲示板，およびどの参加者がどの掲示板に何時投稿したかという情報を含む記事である．

3つの知識共有サイトからデータを収集した．一つめは，高い活動を続けている市民参加型の議論や会話の場「藤沢市市民電子会議室」の 1999 年 6 月 1 日から 2005 年 9 月 24 日までのデータである．収集したデータの中に現れる参加者は 879 人，掲示板数 73 であり，記事総数は 52881 である．二つ目は，日本最大級の Q&A コミュニティサイト「教えて!goo」を用いた．収集したデータは，1999 年 7 月 29 日から 2006 年 7 月 20 日まであり，参加者は 400690 人，記事総数は 8902882 である．「教えて!goo」においては参加者が質問と回答を行う．一連の質問と回答を一つの掲示板と見ることもできるが，質問当月内に回答が終了するものが全体の 5.7% であり，質問を掲示板とすると期間も回答数も限定されてしまう．ここでは，「教

[0]http://www.city.fujisawa.kanagawa.jp/~denshi/
[0]http://oshiete.goo.ne.jp/

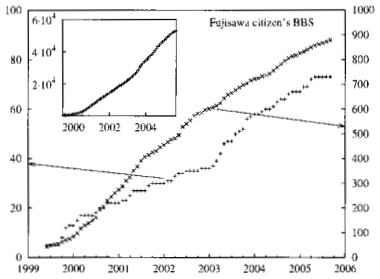

図 1: 「藤沢市市民電子会議室」の掲示板数（+），参加者数（×），記事数（inset）の推移.

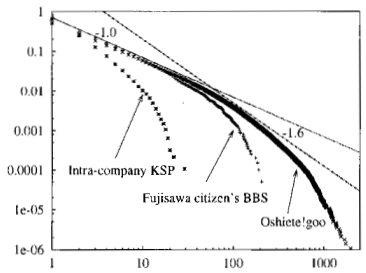

図 2: 月間投稿数の累積分布.

えて!goo」全体を一つの掲示板とみなすことにした．三つ目として，某社内で議論や情報共有をする「社内情報共有サイト」の 2004 年 1 月 9 日から 2006 年 10 月 18 日までのデータである．参加者は 242 人，掲示板数 751 であり，記事総数は 11849 である．

「藤沢市市民電子会議室」の参加者数，掲示板数，記事数は概ね時間と共に一定の割合で増加している（図 1）．「教えて!goo」や「社内情報共有サイト」についても同様の増加傾向がみられる．

3 投稿数のダイナミックス

ある参加者による掲示板への一連の記事投稿行為をこの参加者と掲示板の「投稿系列」と呼び，知識共有サイトの活性度として，各々の投稿系列の 1ヶ月の投稿記事数に着目し，その分布，時間推移についてしらべた．

3.1 投稿数分布と投稿数増加率分布

i 番目の参加者が j 番目の掲示板に t 月に投稿した記事数を $x_{ij}(t)$ と書くことにする．まず，投稿数 $x_{ij}(t)$ の分布を調べた（図 2）．いずれの曲線も両対数グラフで直線的でほぼべき的に分布している部分と，有限効果と思われる投稿数の大きな所の急激な落ち込みがみられる．「教えて!goo」の投稿数累積分布では，投稿数の比較的少ない領域では -1.0 くらいであり，比較的多いところでは -1.6 程度である．これらの分布は基本的にはべき分布と考えられるが，2 重パレート分布や対数正規分布の可能性も考えられる．

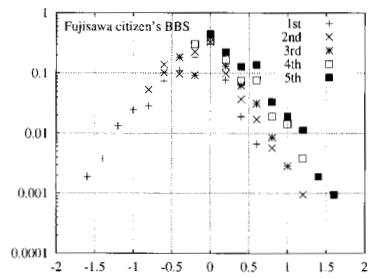

図 3: 投稿増加率の前月投稿数への依存性.

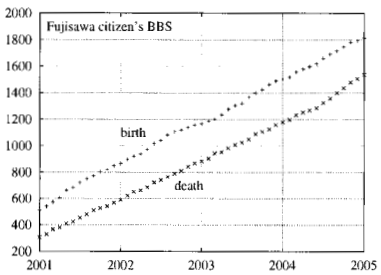

図 4: 投稿系列の生成消滅累積数の推移.

次に，投稿増加率 $r_{ij}(t) = x_{ij}(t)/x_{ij}(t-1)$ の分布の前月の投稿数に対する依存性をみるため，データ全体を前月の投稿数 $x_{ij}(t-1)$ に応じて 5 分割して，それぞれの投稿増加率の分布を図 3の図に示した．前月の投稿数 $x_{ij}(t-1)$ が少ないグループでは r_{ij} が小さい領域の分布が存在しないが，この部分を除いて，増加率分布はほぼ一致していることがわかる．以上のことから，投稿増加率の分布は，前月の投稿数には依存せず，ほぼ同じ分布であると言える．これは，Gibrat 則とよばれるものである [2, 8].

3.2 投稿系列の生成消滅と継続月齢分布

通常投稿系列はある程度の期間継続されるが，参加者の興味の変化や掲示板のトピック推移などにより，投稿系列の生成消滅が頻繁に生じる．投稿系列の生成とは，新規参入者や掲示板の新設および既存の参加者が未投稿の掲示板に投稿した場合も含まれる．また，ある時期以降投稿がない場合に，投稿系列が消滅したとみなす．

投稿系列の生成と消滅の累積数曲線を図 4に示す．投稿系列の生成数や消滅の累積数はほぼ直線的に増加しており，1 月当りほぼ一定である．また，生成と消滅を表す曲線は平行であり，投稿系列数はほぼ一定数に保たれている．アクティブな投稿系列が 200 から 300 程度であるに対して毎月 20 程度が生成し消滅をしている．投稿系列が絶えず入れ替わることは知識共有サイトの大きな特徴であり，モデルを考える上で大変重要である．

ある時刻におけるアクティブである投稿系列の月齢とは，その投稿系列が生成されてからその時点までの経過期間の長さ（月数）のことである．投稿系列の月齢分布を図 5に示す．月齢分布には初期の急激な減衰とその後

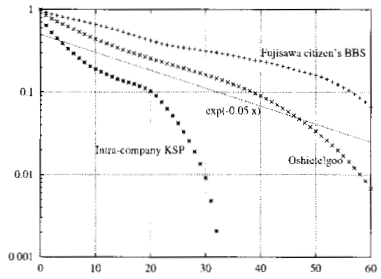

図 5: 月齢分布.

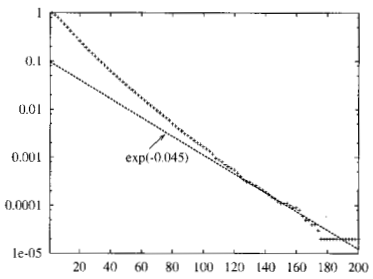

図 6: アクティブ投稿系列数の推移（シミュレーション）

の小さな指数の指数関数的な減衰がみられる.

4 投稿行動モデル

これまでの実データの解析によると，知識共有サイトは概ね次のような性質をもつことが分かった.（1）投稿記事数はべき分布する.（2）投稿増加率は前月の投稿数に依存せずかつ時間相関がない. Gibrat 則が成立している.（3）投稿系列の生成消滅の毎月ほぼ一定数生じる.（4）月齢分布は指数関数的な振舞いをする. この章では,（2） Gibrat 則と（3）投稿系列の生成消滅の両者の結果から数理的投稿行動モデルを構築し,（4）月齢分布と（1）投稿記事数分布を導出する.

4.1 乗算確率過程

Gibrat 則から，$r_{ij}(t)$ をある独立同一の確率分布に従う確率変数とみなせば，月間投稿数 $x_{ij}(t)$ は乗算確率過程,

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

にしたがって時間発展することになる. $\ln r(t)$ の平均を μ，分散を σ^2 とすると，十分大きな t では，$\ln x(t)$ は平均 $t\mu$，分散 $t\sigma^2$ の正規分布に漸近する. 乗算確率過程は時間に不変な投稿数分布をもたず，前章のデータ解析結果と符合せず，純粋な乗算確率過程である式 (1) は投稿モデルとしては適切とはいえないだろう.

経済物理などで乗算確率過程を用いる場合は，乗算確率過程を現象に合わせて変更したモデルが使われることが多い. 例えば，Souma は，企業倒産の仕組みとしてリセットイベントを導入し，企業サイズ分布がべき則にしたがうことを示している [8]. しかし，投稿系列のモデルとしてはリセットイベントに相当する現象はなく知識共有サイトのモデルとしては妥当でない. また，境界条件（反射壁）や雑音を加えた場合も安定なべき則が得られることが知られている [4, 6, 9]. しかし，投稿系列モデルとしては雑音や反射壁の意味付けや実データからの推定が難しいという課題が残る. 一方，Reed らや Huberman らは，モデルの構成要素数が指数的に増大するとき，要素のサイズ分布がべき則にしたがうことを示した [3, 7] が，知識共有サイトにおいて構成要素が指数的に増大することはない. ここでは前説でみた投稿系列の生成消滅を乗算確率過程の追加削除のプロセスとしてモデルに導入し，知識共有サイトの振舞いを再現する自然なモデルを提案する.

4.2 投稿系列の生成消滅モデル

投稿系列の生成消滅数が毎月ほぼ一定数であることを 3.2 でみた. まず，投稿系列生成として，新たな投稿系列に対応する乗算確率過程を毎月一定数追加する. また，$x(t) < 1$ の場合は投稿がなかったとみなし，さらに，消滅条件 $x(t) < \theta$ (< 1) を満した投稿系列は消滅したとみなす. 知識共有サイトの成長モデルは次のようにまとめられる.

初期化 $\{i,j\}$ の組を N_0 個用意し，乗算確率過程の変数 $x_{ij}(0)$ を x_0 に初期化する.
毎月 既存投稿系列 式 (1) に従い $x_{ij}(t)$ へ推移.
消滅投稿系列 消滅条件を満した $\{i,j\}$ を削除.
生成投稿系列 n 組の新規 $\{i,j\}$ を用意し，乗算確率過程の変数 $x_{ij}(t)$ を x_0 に初期化する.

知識共有サイトの投稿系列生成消滅モデルの振舞いを調べるためシミュレーションを行った. $\ln r(t)$ は 平均 $\mu = -0.3$, 分散 $\sigma^2 2 = 1.0$（「教えて!goo」は $\mu = -0.0785$, $\sigma^2 = 0.969$）の正規乱数とした. このとき，時刻 t での $x(t)$ 分布 p_t も次の正規対数分布になる.

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

初期条件 $N_0 = 100000$, $x_0 = 10.0$ のとしての，アクティブな投稿系列の割合の時間推移を図 6に示す. t が大きなところでは指数関数的な減少をしている. このシミュレーションの結果の分布は累積寿命分布に相当し，実データの累積寿命分布（図 6）によく一致する.

投稿系列生成から t 月後にアクティブである割合 $f(t)$ は，消滅条件から $x > \theta$ である確率なので,

$$f(t) = \int_\theta^\infty p_t(x)dx = \frac{1}{2} erfc\left(\frac{\ln\theta - y_0 - t\mu}{\sqrt{2t}\sigma}\right) \qquad (3)$$

となる. ただし，$y_0 = \ln x_0$ であり，$erfc(x)$ は相補誤差関数を表す. 大きな t に対して，近似式 $erfc(t) \simeq e^{-t^2}/\sqrt{\pi}t$ が成立するので [1],

$$f(t) \sim \frac{\sigma}{\sqrt{2\pi t}(-\mu)} e^{-\frac{\mu^2}{2\sigma^2}t} \sim e^{-t\mu^2/2\sigma^2} \qquad (4)$$

と近似できる. つまり，累積寿命分布は時間に対して指数 $-\mu^2/2\sigma^2$ の指数関数で減衰する. 実験パラメータで指数は -0.045 となりシミュレーション結果と合う.

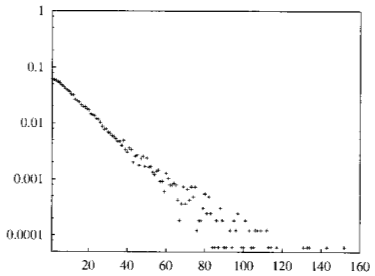

図 7: 月齢分布（シミュレーション）.

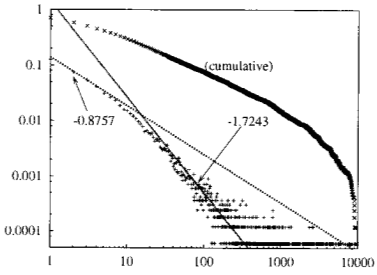

図 8: 投稿数の分布（シミュレーション）.

τ 月前に新規生成された投稿系列を n 個とすれば，この中で現在アクティブな状態にある投稿系列数の期待値は $nf(\tau)$ となる．つまり，投稿系列の月齢分布 $g(\tau)$ は寿命分布 $f(\tau)$ に比例する．シミュレーションにおける月齢分布（図 7）も指数関数的な振舞いをしている．

次に，初期投稿系列数を $N_0 = 1000$，毎月生成される投稿系列数 $n = 1000$ としてシミュレーションを行った．投稿数 x_{ij} の分布および累積分布を図 8 に示す．実際の知識共有サイトで観測された投稿数分布図 2とよく似たグラフととなっていることがわかる．

投稿数分布 $h(x)$ は t を連続近似すれば，

$$h(x) = \int_0^\infty g(t)p_t(x)dt \tag{5}$$

となる．投稿系列月齢分布は $g(t) \propto e^{-t\mu^2/2\sigma^2}$ なので，$h(x)$ は次のような 2 重パレート分布となる [5].

$$h(x) = \begin{cases} C\left(\frac{x}{x_0}\right)^{-\gamma_1} & \text{if } x \le x_0 \\ C\left(\frac{x}{x_0}\right)^{-\gamma_2} & \text{if } x \ge x_0 \end{cases} \tag{6}$$

ただし，$\gamma_1 = 1-(1-\sqrt{2})\frac{\mu}{\sigma^2}$，$\gamma_2 = 1-(1+\sqrt{2})\frac{\mu}{\sigma^2}$ となる．実験パラメータ値では，それぞれ $\gamma_1 = 0.8757$，$\gamma_2 = 1.7243$ となり，この指数の曲線を補助線として x 分布のプロットと一緒に示したが（図 8），分布の傾きとよく一致している．つまり，提案モデルでは投稿数分布は 2 重パレート分布でよく近似され，実データにおいても 2 重パレート分布として近似できる可能性がある．

5 おわりに

市民掲示板や Q&A サイトなどの複数の知識共有サイトについて実証的な解析を行い，これを基に投稿記事数の数理的成長モデルを提案した．ある参加者，掲示板に関する投稿系列の記事数の揺ぎは，記事数に比例しかつ揺ぎの特性は記事数に依存しない，いわゆる Gibrat 則に従う性質のものであり，乗算確率過程として捉えられることがわかった．また，知識共有サイトにおいては投稿系列の生成消滅が頻繁に起るという特徴があることが分かった．このため，提案投稿系列モデルでは，一定の割合で新規投稿系列を追加し，投稿系列の消滅条件を設定した．このモデルにより，投稿系列の寿命分布や投稿数のべき分布がよく再現できることをシミュレーションおよび近似計算から示せた．

今回，投稿系列の生成消滅という現象に焦点を当てモデルを構築提案した．この生成消滅の仕組みのあるシステムとは，閉鎖的ではなく構成要素の新規参入が許される成長する系であると同時に，既存の要素もいつかは消滅してしまうという系である．このような新陳代謝があるシステムは社会システムや自然界には多数存在し，ある意味普遍的な性質であるかもしれない．そうだとすれば，同様の数理モデルが適応できるシステムは多数存在するのではないかと考えている．

このような投稿系列の生成消滅は，参加者の興味の推移や掲示板の活性度，さらには掲示板の話題の特徴や参加者の行動特性にも影響されるであろう．次の重要な研究課題として，参加者や掲示板，それらを結ぶ記事の相互作用，まさにネットワーク解析的な立場から投稿系列の生成消滅の特性を探っていくことは重要だと考える．

参考文献

[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 inegalites economiques* , 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] Souma, W.: Multiplicative stochastic process in Econophysics (in Japanese) , *Proceedings of Ninth Workshop on Informationn-Based Induction Sciences (IBIS2006)*, pp. 192–199 (2006).

[9] 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).