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[††]

掲示板に代表される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実証的解析を行い，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，一定期間の投稿記事数の時間推移に対する投稿数増加率が Gibrat 則を満たすことから，記事数の時間発展を乗算確率過程としてとらえることができることを示した．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，投稿系列に対応する乗算確率過程の生成消滅のためのメカニズムを導入した新たな知識共有サイト投稿行動モデルを提案した．この提案行動モデルにより投稿系列の投稿継続期間が指数分布となることや投稿数分布が定常的なべき分布となることを計算機シミュレーションおよび解析結果から示した．提案モデルは単純な乗算確率過程に比べ現実の投稿行動に近いモデルであり，よく実データの性質を再現できるモデルとなっている．

Time Evolution of Knowledge Sharing Portal Activities as Multiplicative Random Process

Kenichi Arai,[†] Takeshi Yamada[†] and Yukio Hayashi[††]

We propose a new evolution model of the article posting activities in the Knowledge Sharing Portal (KSP), in which one can post messages, exchange opinions, and ask and answer questions. Typical examples of KSP include online Bulletin Board System (BBS), intra-company information exchange service, word-of-mouth and Q&A community sites. We have constructed a model based on extensive analysis using real data of three different KSPs. First, we show that the number of articles posted in a fixed time interval obeys Gibrat's law, and can be modeled as Multiplicative Random Process (MRP). Next, we extend the model by introducing the birth and death mechanisms of posting sequences. The proposed model can successfully reproduce exponential distributions observed for the age of posting sequences and Pareto distributions for the number of postings. Compared to the simple MRP model, the proposed model is a more practical one that can explain the real posting behaviors.

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．Web サイトから必要な情報を検索し入手する方法に加えて，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欲する情報の提供を求めたりするなど，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．この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る．知識共有サイトは今後ますます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．この課題に取り組む手法の 1 つとして，コンテンツを解析してトピック推移や参加者の役割などの調査を行うことは有効であろう．しかしその一方で，特にインターネット上のサイトなどデータ量が大量である場合には，コンテンツの解析を行うことは困難な場合も多い．むしろデータ間の関係性だけを用い全体構造

† 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories, NTT Corporation
†† 北陸先端科学技術大学院大学
Japan Advanced Institute of Science and Technology

をネットワークととらえ大局的な解析を行うことの方が有効な場合も多い[1),2),4),14)]．知識共有サイトにおいても，参加者や掲示板をノードとし，記事を参加者と掲示板を結び付けるリンク，記事数をリンクの重みと見なすと，知識共有サイトは重み付き 2 部グラフとしてとらえることができる．つまり，参加者の参入や掲示板の新設，記事の投稿などの知識共有サイトの拡大を 2 部グラフのネットワークの成長と見なし，ネットワーク解析の手法により数理的モデルを構築することは有効であろう．実際，複雑ネットワークの視点から優先接続（preferential attachment）による成長モデルやノードの次数分布などについての研究は行われている．たとえば，掲示板などの知識共有サイトの成長を複雑ネットワークの視点から扱ったものとして，掲示板としての 2 部グラフについての成長モデルの提案，解析をしたもの[15)]，掲示板の詳細な解析と簡単なモデルの構築をしたもの[8)]，また，2 部グラフの成長モデルを構築したもの[16)] などいくつかの研究がある．これらの研究では，優先接続を基本的な考え方とし，その解釈の妥当性を検証したり成長モデルにより生成されたネットワークと実データの統計的な性質を比較したりしているものが多い．2 部グラフの通常のネットワーク成長モデルでは，各ノードの次数の増加はネットワーク全体との相対的な関係で決まり，特定のノードに着目した成長ダイナミックスを構築するのは難しい．さらに，ノードの増加にともない時間軸をうまくスケールすることが必要であるなどの課題がある．このため，記事数などのダイナミックスや記事数の時系列の生成消滅を記述するモデルとしてネットワーク成長は必ずしも扱いやすいものではないと考える．

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．

本論文の構成は以下のとおりである．知識共有サイトのモデル化のための準備と収集したデータについて 2 章で述べる．収集したデータの投稿数の時系列推移や生成消滅についての解析結果を 3 章で示し，モデル化およびシミュレーションの結果を 4 章で述べる．最後に，5 章でまとめを述べる．

2. 知識共有サイトのデータ構成

本論文で扱う「知識共有サイト」とは，インターネットもしくはイントラネット上のサービス（Web ページ）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サブシステムから構成される．ここでは，これらのサブシステムを掲示板と呼ぶことにする．参加者は設定された話題に沿った内容の文章などの記事を掲示板にアップロードする．このことを，「記事の投稿」と呼ぶ．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．なお，ここでは，記事の参照関係などによる話題に細分化や参加者の属性などによる分類については考慮しないことにする．つまり，知識共有サイトモデルの構成要素は，参加者，掲示板，記事の 3 つであり，投稿された各記事に含まれる情報から基礎データとしてこれらを収集した．データの具体的構成としては記事固有の番号である記事番号，記事が投稿された日時，記事の投稿先の掲示板に割り当てられている ID，投稿者 ID である．ただし，今回収集した知識共有サイトでは記事の投稿には事前に投稿者の登録が必要であり，このとき割りふられる登録者固有の ID を投稿者 ID として投稿者の識別のために用いた．また，実際のデータの例として一部を示したのが **表 1** である．

我々は，参加者，掲示板，記事のデータを次の 3 つの知識共有サイトから収集した．1 つめは，地方自治体が運営する市民参加型の議論や会話の場として高い活動を続けている「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

表 1 収集データの一例
Table 1 Example of collected data.

記事 ID	投稿時間	掲示板 ID	投稿者 ID
20764	2001/4/19 18:20	107	860
12824	2001/4/19 19:17	74	794
20765	2001/4/19 20:44	107	1258
23683	2001/4/19 21:05	109	12
24299	2001/4/19 21:19	122	12
24074	2001/4/20 00:32	115	531
10997	2001/4/20 00:37	48	383
18434	2001/4/20 00:44	92	1056
24075	2001/4/20 00:51	115	1451
23684	2001/4/20 06:50	109	531

☆ http://www.city.fujisawa.kanagawa.jp/˜denshi/

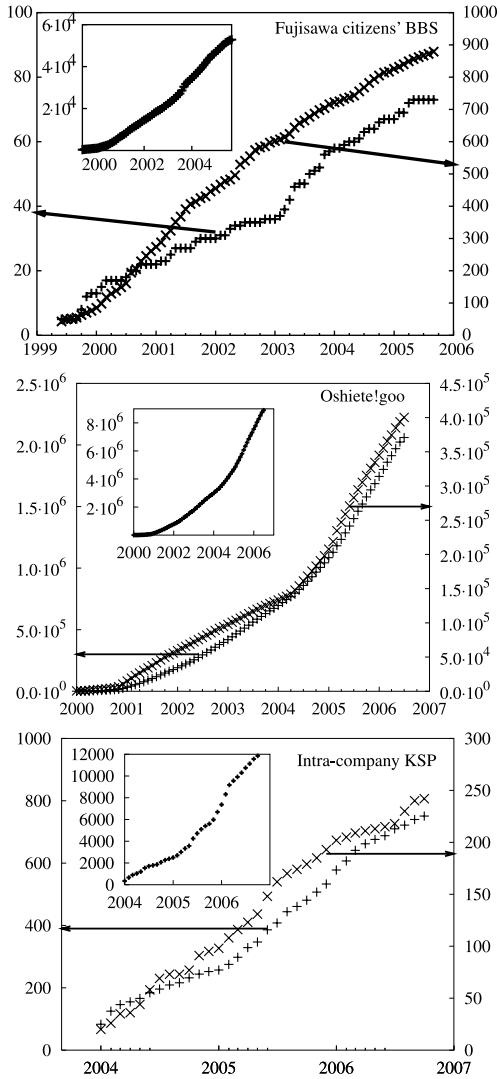

図 1 「藤沢市市民電子会議室」（上図），「教えて!goo」（中図）および 「社内情報共有サイト」（下図）における掲示板数（左 y 軸），参加者数（右 y 軸），記事数（inset）の推移．ただし，「教えて!goo」は掲示板数ではなく質問数である．

Fig. 1 Time evolutions of the cumulative number of bulletin boards (left y-axis), authors (right y-axis) and articles (inset) in "Fujisawa citizen's BBS"(upper), "Oshiete!goo"(middle) and "Intra-company KSP"(lower). Note that, in "Oshiete!goo", left y-axis indicates the number of questions instead of bulletin boards.

板数 73 であり，記事総数は 52,881 である．ただし，これらの数字は収集データから得られたものであり，公式な登録者数，掲示板数，記事数とは異なる．2 つ目は，日本最大級の Q&A コミュニティサイトである「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 日までであり，参加者は 400,690 人，記事総数は 8,902,882 である．「教えて!goo」などの Q&A コミュニティサイトでは，参加者が質問と回答を行う．一連の質問と回答を 1 つの掲示板と見ることもできるが，質問者が満足できる回答が寄せられると質問に対する回答が締めきられるため質問あたりで見ると期間も回答数も限定されてしまう．実際，質問のあった当月だけに回答が限られるものは全体の 94.3% であり，複数月にわたって回答が続くことは少ない．これらの状況から，「教えて!goo」全体を 1 つの掲示板と見なすことにした．3 つ目として，某社内で様々な話題について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 日までのデータについても解析を行った．参加者が限定されているので，他に比べて小規模であり，参加者は 242 人，掲示板数 751 であり，記事総数は 11,849 である．

また，これらの知識共有サイトでの参加者数，掲示板数，記事数の増加の時系列を**図 1** に示す．いずれのデータについても，おおむね時間とともに一定の割合で増加しているといえる．ただし，「教えて!goo」については 2004 年春頃を境にグラフの傾きが変わっている．特に参加者において増加率の変化はシャープであり，それに影響され記事数の増加割合も大きくなったのではないかと考えられる．変化点の前後の期間で参加者増加率の要因が変わったと考えられるが，前後の期間を区分的に見れば，やはり一定の割合で増加していることがいえる．

3. 投稿数のダイナミックス

知識共有サイトの活性度として，各々の参加者が各々の掲示板へ一定期間内に投稿する記事数に着目し，その時間推移について調べた．ここでは一定期間として 1 カ月間を用いた．i 番目の参加者 A_i が j 番目の掲示板 B_j に t 月に投稿した記事数を $x_{ij}(t)$ と書くことにする．

3.1 投稿数分布とその相関

本論文で最も基本とする量は 1 カ月あたりの投稿数 $x_{ij}(t)$ である．まず，投稿数 $x_{ij}(t)$ の分布を調べた．**図 2** は 4 期間（2003 年 07 月から 2003 年 12 月まで，2004 年 01 月から 2004 年 06 月まで，2004 年 07 月から 2004 年 12 月まで，2005 年 01 月から 2005 年 06 月まで）の累積月間投稿数の分布の重ね書きであるが，4 つの分布はほぼ一致しており，投稿数分布は時間に関して不変であると考えてよいことが分かる．

このように，投稿数に関する統計的性質が時間に関

☆ http://oshiete.goo.ne.jp/

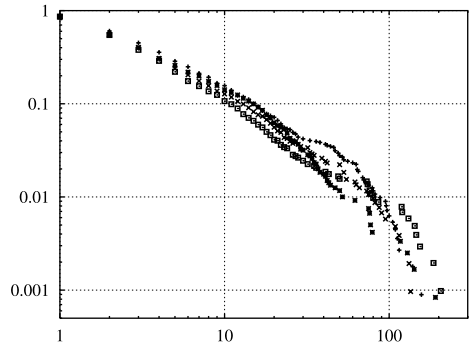

図 2 「藤沢市市民電子会議室」における 2003 年 07 月から 2003 年 12 月まで（+），2004 年 01 月から 2004 年 06 月まで（×），2004 年 07 月から 2004 年 12 月まで（∗），2005 年 01 月から 2005 年 06 月まで（⊡）の月間投稿数の累積分布

Fig. 2 Cumulative probability distributions of several numbers of semiannually posted articles in “Fujisawa citizen’s BBS”.

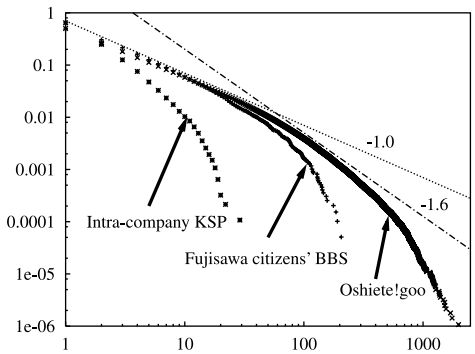

図 3 「藤沢市市民電子会議室」,「教えて!goo」および「社内情報共有サイト」における月間投稿数の累積分布

Fig. 3 Cumulative probability distributions of number of monthly posted articles in “Fujisawa citizen’s BBS”, “Oshiete!goo”and “Intra-company KSP”.

してほぼ不変と考えられるので，収集したすべての月のデータを用いて解析した（**図 3**）．すべての月のデータを用いることにより，プロットに用いることのできるデータ数は多くなり，分布の関数形はより鮮明になっている．いずれの曲線も両対数グラフで直線的であり，ほぼべき的に分布しているといえる．「教えて!goo」の投稿数累積分布では，べきの指数が投稿数により変わっており，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．投稿数 1,000 付近☆の急激な落ち込みは有限効果と考えられる．「藤沢市市民電子会議室」では，

☆ 1 カ月あたりの投稿数が「藤沢市市民電子会議室」で 100 通を超え，「教えて!goo」で 1,000 通を超える場合があるが，比較的短い記事の投稿や各種データのアップロードなども多くあり，同一投稿者の記事数と見ても可能な範囲であると考える．

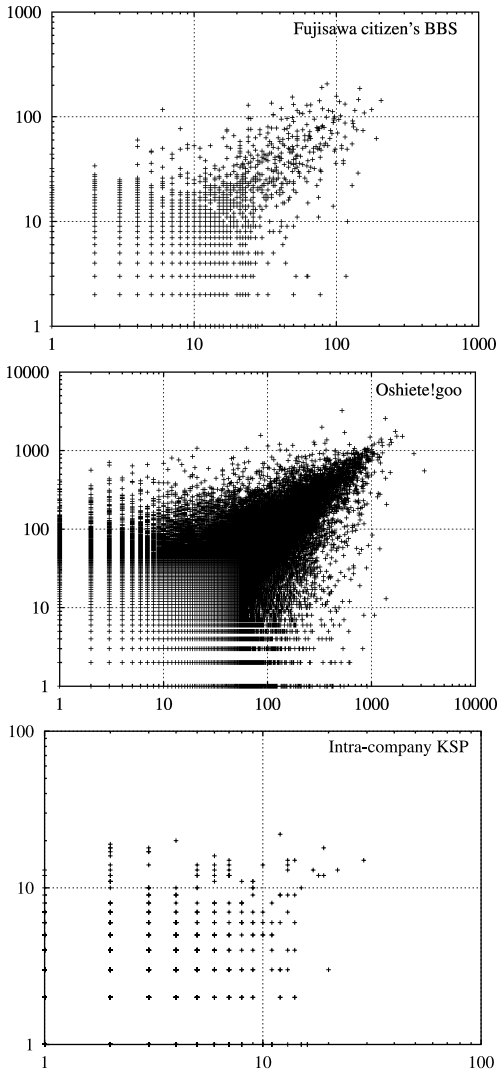

図 4 「藤沢市市民電子会議室」（上図），「教えて!goo」（中図）および「社内情報共有サイト」（下図）における前月翌当月の投稿数の相関

Fig. 4 Scatter plot of the number of posted articles in current and the previous months for “Fujisawa citizen’s BBS”(upper), “Oshiete!goo”(middle) and “Intra-company KSP”(lower).

投稿数の比較的少ない領域では「教えて!goo」と同様に −1.0 くらいであり，その後有限効果と思われる落ち込みがある．「社内情報共有サイト」はさらに急な曲線となっており，投稿数の少ない領域で指数は −2 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，両対数グラフで上に凸な曲線となる傾向がある．その理由として，観測の有限効果との区別は難しいがある閾値でべき指数が切り替わる 2 重パレート分布や対数正規分布である可能性も考えられる．

図 4 は前月と当月の投稿数の相関図，つまり，横軸

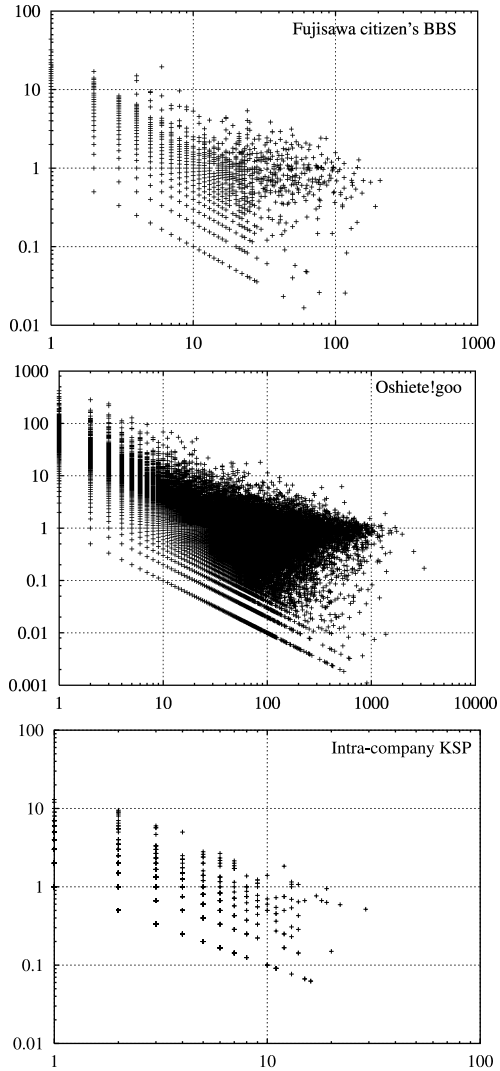

図 5 「藤沢市市民電子会議室」（上図），「教えて!goo」（中図）および「社内情報共有サイト」（下図）における前月の投稿数と当月の投稿増加率の相関

Fig. 5 Scatter plot of $x(t-1)$ and $r(t)$ for "Fujisawa citizen's BBS" (upper), "Oshiete!goo" (middle) and "Intra-company KSP" (lower).

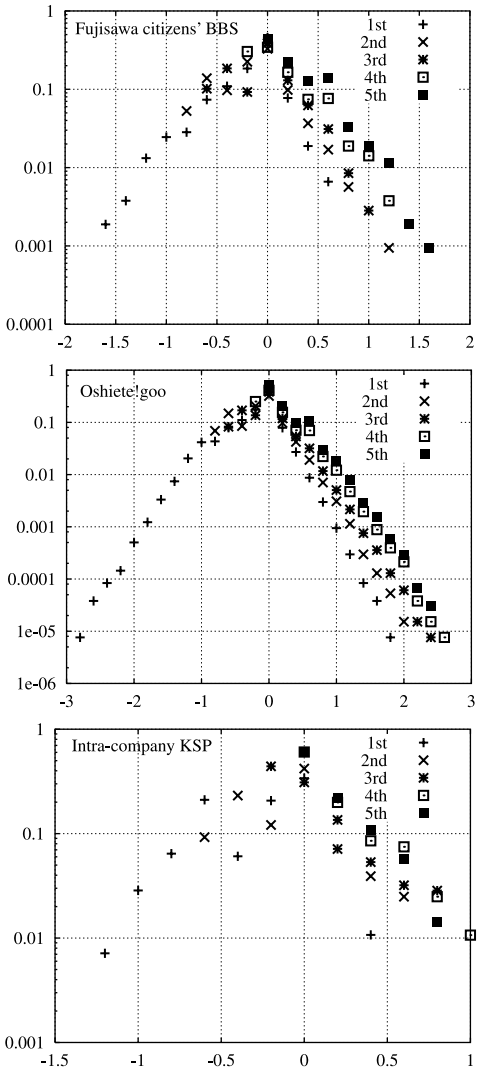

図 6 投稿増加率の前月投稿数への依存性．「藤沢市市民電子会議室」（上図），「教えて!goo」（中図）および「社内情報共有サイト」（下図）

Fig. 6 Probability distribution of growth rate depending on a number of articles. "Fujisawa citizen's BBS" (Upper), "Oshiete!goo" (middle), "Intra-company KSP" (lower).

を前月の投稿数，縦軸を当月の投稿数としてプロットしたものである．連続する 2 カ月の投稿数がともに 1 以上でないと，有効なデータが得られないので，データ数が限られ，有効なデータ数は，「藤沢市市民電子会議室」が 5,301，「教えて!goo」が 656,870，「社内情報共有サイト」が 1,401 であった．グラフの対角線に対してほぼ対称に分布していると見られ，詳細釣り合いが成り立っていると考えることができる．Fujiwara らによると，次節で述べる Gibirat 則と詳細釣り合いから，べき分布が導き出せることを指摘しており，我々のデータ解析結果とおおむね一致する[6]．

3.2 投稿数増加率と Gibrat 則

投稿数の月ごとの増減についてその比，つまり増加率

$$r_{ij}(t) = x_{ij}(t)/x_{ij}(t-1)$$

に着目する．**図 5** は，横軸を前月の投稿数，縦軸を当月の投稿増加率をプロットしたものである．投稿数は自然数であり 1 未満が存在しないため，図の左下の部分 $r_{ij}(t)x_{ij}(t-1) < 1$ のデータが欠損しているように見える．

投稿増加率 $r_{ij}(t)$ の分布の前月の投稿数に対する依

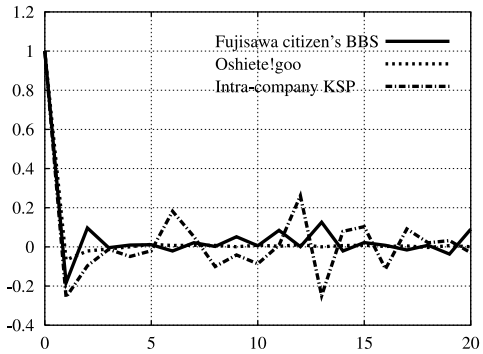

図 7 投稿増加率の自己相関関数
Fig. 7 Auto correlation function of growth rate.

存性を見るため，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 6** に示した．投稿数が最多のグループは “+”，以降順番に “×”，“∗”，“⊡” で最小のグループが “■” のマークでプロットしてある．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．この部分を除いては，「藤沢市市民電子会議室」および 「教えて!goo」の増加率分布はほぼ一致していることが分かる．「社内情報共有サイト」においては，データ数が少なく，ばらつきが大きい部分が多いが，おおむね一致している傾向があることが分かる．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いえる．これは，Gibrat 則と呼ばれるものであり，主に企業規模の成長率について詳しく調べられている[5)~7),21),22)]．

次に，投稿増加率の時間相関を見るために，自己相関関数

$$R(\tau) = \frac{E((r(t)-\mu)(r(t+\tau)-\mu))}{\sigma^2}$$

を**図 7** に示す．$\tau = 0$ で，時間相関 R は，1 であるが，$\tau \geq 1$ 以降はデータ数によって揺らぎはあるが急速に時間相関が減衰しており，ほぼ 0 と見なしてよいことが分かる．

3.3 投稿系列の生成消滅と継続月齢分布

これまで，ある参加者がある掲示板へ 1 カ月間に投稿する記事数に着目し，その分布や時間推移について解析してきた．ここでは，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，投稿系列の生成や消滅に着目する．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投稿系列がた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，モデルを考えるうえで大変重要である．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いままで投稿したことのない既存の掲示板に記事を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見なす．

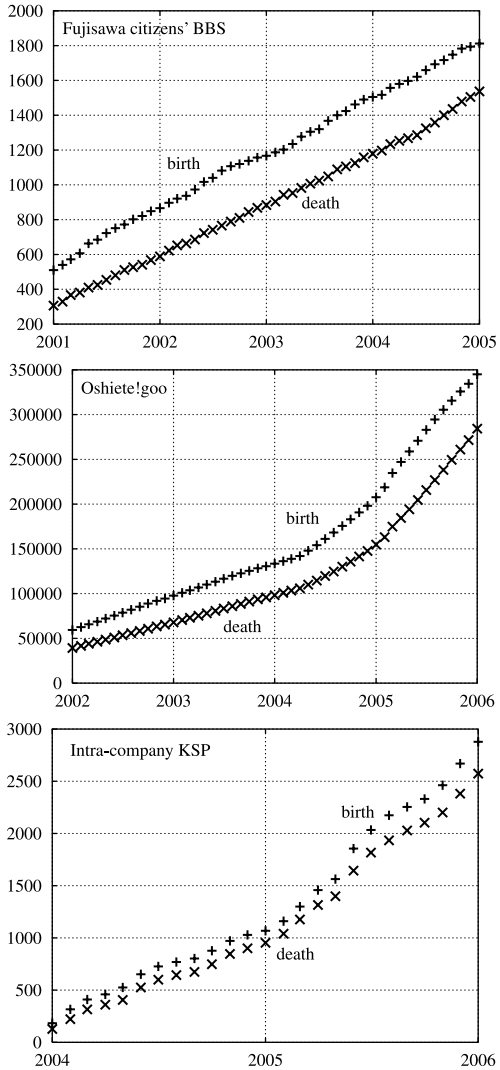

図 8 投稿系列の生成・消滅と有効な投稿系列の推移．「藤沢市市民電子会議室」（上図），「教えて!goo」（中図）および 「社内情報共有サイト」（下図）
Fig. 8 Time evolution of birth and death numbers of posting sequences. “Fujisawa citizen's BBS”(Upper), “Oshiete!goo”(middle), “Intra-company KSP”(lower).

投稿系列の生成消滅の頻度を見るために，新しく生成された投稿系列数と消滅した投稿系列数の累積数を**図 8** に示す．いずれのグラフもほぼ直線的な増加をしており，毎月ほぼ同数の投稿系列が生成され，毎月ほぼ同数の投稿系列が消滅している．ただし，「教え

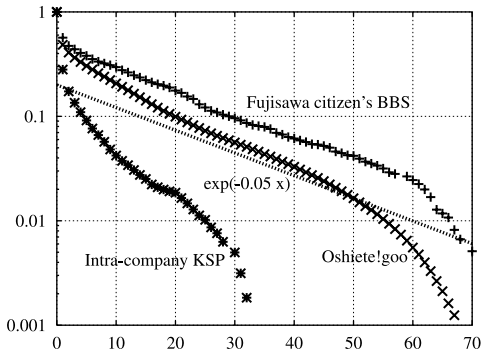

図 9 投稿系列寿命の累積分布．「藤沢市市民電子会議室」(+)，「教えて!goo」(x) および 「社内情報共有サイト」(*)

Fig. 9 Life time distribution of posting sequences. "Fujisawa citizen's BBS"(+), "Oshiete!goo"(x) and "Intra-company KSP"(*).

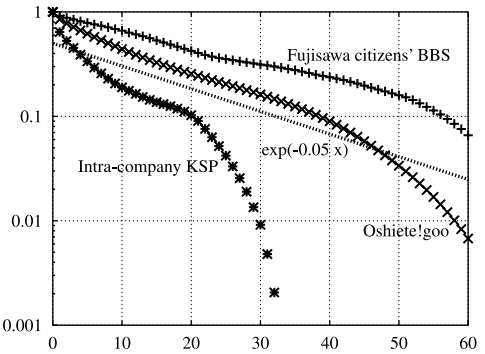

図 10 月齢分布．「藤沢市市民電子会議室」(+)，「教えて!goo」(x) および 「社内情報共有サイト」(*)

Fig. 10 Age distribution of posting sequences. "Fujisawa citizen's BBS"(+), "Oshiete!goo"(x) and "Intra-company KSP"(*).

て!goo」においては，2004 年の中期以降で直線の傾きが急になっており，1 カ月あたりの生成数，消滅数の変化を意味している．図 1 でも見たように，同時期に参加者増加率の急激な変化も見られその影響を受けたもの見られるが，期間を前後に分けると生成消滅率はやはりほぼ一定である．また，生成と消滅を表す曲線は平行，つまり投稿系列の生成数消滅数はほぼ同数であり，投稿系列の入れ替わりはあるが投稿系列数は平衡状態にある．「教えて!goo」ではアクティブな投稿系列が 50,000 程度であるときにこの 5 分の 1 の約 10,000 の投稿系列が毎月生成消滅しており，「藤沢市市民電子会議室」においてもアクティブな投稿系列が 200 から 300 程度に対して 10 分の 1 程度が毎月生成消滅をしている．

このように，投稿系列には生成と消滅があり，最初に投稿により生成され，最後の投稿により消滅をする．投稿系列は生成後消滅するまでをアクティブであるということにする．投稿系列がアクティブな状態にある期間の長さを投稿系列の寿命と呼ぶ．投稿系列が t 月以上アクティブな状態にある確率，すなわち累積寿命分布を**図 9** に示す．この寿命累積確率分布によると，投稿系列は生成直後の段階において消滅するものが多く，その後累積寿命分布は比較的小さな指数の指数関数により減衰していく．ただし，各曲線において寿命の長い部分での急速な減衰は，データの観測期間の有限効果によるものだと考えられる．

このようにそれぞれの投稿系列が生成された時刻は区々であり，ある時刻においてアクティブである投稿系列もそれぞれ別の生成時刻を持つ．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 10** に示す．月齢分布も時間に対して安定な分布をしており，グラフの曲線をなめらかにするため数カ月の期間の各月ごとの月齢分布の平均値を示している．具体的には，「藤沢市市民電子会議室」は 2004 年 6 月から 2005 年 6 月まで，「教えて!goo」は 2004 年 6 月から 2006 年 6 月まで，「社内情報共有サイト」は 2005 年 10 月から 2006 年 10 までである．この図に見られるように月齢分布においても初期の急激な減衰とその後の小さな指数の指数関数的な減衰が見られ，寿命累積確率分布に振舞いが似ている．

4. 投稿行動モデル

これまでの実データの解析によると，知識共有サイトはおおむね次のような性質を持つことが分かった．

投稿記事数 月間の投稿記事数はべき的に分布する．

投稿増加率 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．

投稿系列の生成消滅 投稿系列の生成消滅は毎月ほぼ一定数である．

寿命累積分布および月齢分布 これらの分布は指数関数的な振舞いをする．

この章では，投稿増加率の Gibrat 則と投稿系列の生成消滅率の両者の解析結果から数理的投稿行動モデルを構築し，この数理モデルにより寿命および月齢分布の指数的な振舞いと投稿記事数のべき分布を導出できることを示す．

4.1 乗算確率過程

月間投稿増加率 $r_{ij}(t)$ に関する 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見なせば，月間投稿数 $x_{ij}(t)$ は乗算確率過程，

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

に従って時間発展することになる．式 (1) の両辺の対数をとると，

$$\ln x(t+1) = \ln r(t) + \ln x(t)$$
$$= \ln r(t) + \ldots + \ln r(0) + \ln x(0) \quad (2)$$

となる．ただし，表記の簡単のため下付文字 ij を省略した．$\ln r(t)$ の平均を μ，分散を σ^2 とすると，十分大きな t では，初期状態 $x(0)$ の影響は無視でき，かつ中心極限定理より $\ln r(t)$ の分布に関係なく $\ln x(t)$ は平均 $t\mu$，分散 $t\sigma^2$ の正規分布に漸近する．

μ が負の場合，$\ln x(t)$ は時間とともに負の無限大に発散するので，投稿数期待値 $< x >$ は 0 に近づき，掲示板の活動が徐々になくなることを意味し，活動を継続している知識共有サイトのモデルとしては不適切である．μ が正のときは，投稿数期待値が無限大に発散し，この場合も投稿数のモデルとして適切ではない．また，$\mu = 0$ となるよう $r(t)$ の分布を選んだ場合，投稿数の期待値は有限ではあるが投稿数の分散はやはり発散し，時間に不変な投稿数分布を持つモデルにはならず，前章のデータ解析結果と符合しない．これらのことから，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．

このため，経済物理などで乗算確率過程を用いる場合は，上記の課題を避けるため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．たとえば，Souma は，企業倒産の仕組みとしてある確率で系列を初期化するリセットイベントを導入し，企業サイズ分布がべき則に従うことを示している[21]．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，下限反射壁や上限反射壁を設けるなどの境界条件を課したり，あるいは乗算確率過程にさらに雑音を加え変数を値を制限したりする方法に関する研究は多くあり，このように変更された乗算確率過程では変数の分布として安定なべき則が得られることが知られている[10),13),19),20),23]．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従うことを示した[9),17),18)] 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．ここでは図 8 で見たように，知識共有サイトの特徴である頻繁に生じる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，乗算確率過程の拡張として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 3.3 節で見た．前節の知識共有サイトの投稿モデルでは，既存の投稿系列投稿数が乗算確率過程に従って時間推移するだけであった．生成消滅モデルでは，投稿系列生成として，既存投稿系列とは別の新たな投稿系列に対応する新規の乗算確率過程を毎月一定数追加する．また，$x(t) < 1$ となった場合は，t 月は投稿がなかったと見なしたが，さらに，ある閾値 θ に対し，$x(t) < \theta$（< 1）の条件を満した投稿系列は消滅したと見なすことにする．つまり，生成と消滅の間，投稿系列はアクティブであるということになる．つまり，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** 記事数 $x_{ij}(t-1)$ は式 (1) により $x_{ij}(t)$ へ推移する．
- **消滅投稿系列** 消滅条件を満した $\{i, j\}$ の組を削除する．
- **生成投稿系列** 新規の $\{i, j\}$ の組を n 個用意し，乗算確率過程の変数 $x_{ij}(t)$ を x_0 に初期化する．

簡単のために，投稿数の初期投稿数を $x(0) = x_0$ と固定し，$\ln r(t)$ は平均 μ，分散 σ^2 の正規乱数とした．$\ln r(t)$ は独立同一確率分布に従う確率変数なので前月の投稿数に依存せず（図 6 参照），時間相関はなく（図 7 参照），また対称な確率分布なので $\mu = 0$ であれば詳細釣り合いが厳密に成立し（図 4 参照），前章の解析結果と一致するモデルになっている．このとき，時刻 t での $x(t)$ 分布 p_t も正規対数分布

$$p_t(x) = \frac{1}{\sqrt{2\pi t}\sigma x} \exp\left\{-\frac{(\ln x - t\mu - \ln x_0)^2}{2t\sigma^2}\right\} \quad (3)$$

となる．

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，$\mu = -0.3$，$\sigma^2 = 1.0$，$\theta = 0.01$（「教えて!goo」では，$\mu = -0.0785$，$\sigma^2 = 0.969$）として計算機シミュレーションを行った．μ が負であるので，平均投稿数が時間とともに発散することはなく，新規投稿系列の生成がない場合アクティブな投稿系列数は減少する．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減少の様子を**図 11** に示す．t が大きなところでは指数関数的な減

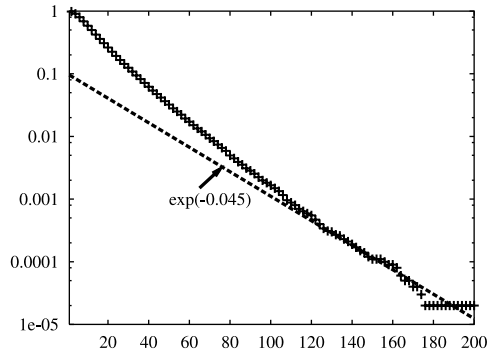

図 11 アクティブな投稿系列の時間推移（シミュレーション）および補助線（$e^{-0.045t}$）
Fig. 11 Lifetime distribution (simulation) and auxiliary line ($e^{-0.045t}$).

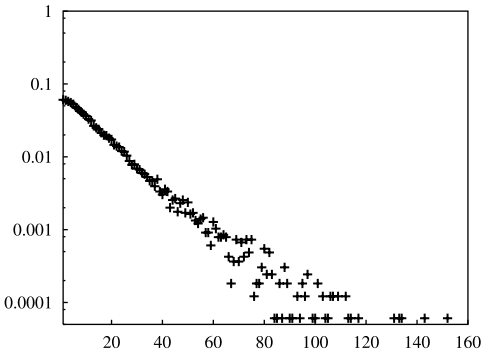

図 12 月齢分布（シミュレーション）
Fig. 12 Age distribution of posting sequences (simulation).

少をしている．投稿系列生成の t 月後にアクティブである投稿系列は t 月以上の寿命を持つことを意味するので，図 11 は累積寿命分布に相当し，このシミュレーションの結果は実際の知識共有サイトの累積寿命分布を示した図 9 の振舞いによく一致する．

投稿系列の消滅条件から，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，$x > \theta$ である確率なので，

$$\begin{aligned} f(t) &= \int_{\theta}^{\infty} p_t(x)dx \\ &= \int_{\ln\theta}^{\infty} \frac{1}{\sqrt{2\pi t}\sigma} e^{-\frac{(y-y_0-t\mu)^2}{2t\sigma^2}} dy \\ &= \frac{1}{2} erfc\left(\frac{\ln\theta - y_0 - t\mu}{\sqrt{2t}\sigma}\right) \end{aligned} \quad (4)$$

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，

$$erfc(t) \simeq \frac{1}{\sqrt{\pi}} \frac{e^{-t^2}}{t}$$

という近似式が成り立ので[3]，$f(t)$ の振舞いは次のように近似できる．

$$\begin{aligned} f(t) &\simeq \frac{1}{2} erfc\left(\frac{-\mu}{\sqrt{2}\sigma}\sqrt{t}\right) \\ &\simeq \frac{\sigma}{\sqrt{2\pi t}(-\mu)} e^{-\frac{\mu^2}{2\sigma^2}t} \end{aligned} \quad (5)$$

大きな t では，指数部分が支配的になるので，

$$f(t) \sim e^{-t\mu^2/2\sigma^2} \quad (6)$$

と近似できる．つまり，アクティブな投稿系列や寿命分布は時間に対して指数関数的に減衰し，そのときの指数は $-\mu^2/2\sigma^2$ となる．シミュレーションのパラメータ値での場合，指数は -0.045 となり，図 11 のシミュレーション結果と合う．さらに，実データの寿命累積分布（図 9）と比較すると，観測期間の有限効果が見られる寿命の長い部分を除いて，投稿系列生成直後の急な減少やその後の指数関数な振舞いなど両者はよく一致していることが分かる．

τ 月前に新規生成された投稿系列を $n(\tau)$ 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 $n(\tau)f(\tau)$ となる．つまり，投稿系列の月齢分布 $g(\tau)$ は $n(\tau)f(\tau)$ に比例する．図 8 によると毎月の生成投稿系列数は一定であるので，$g(\tau)$ は $f(\tau)$ に比例する．

$$g(\tau) \propto f(\tau) \quad (7)$$

実データの寿命累積分布（図 9）と月齢分布（図 10）が似た振舞いをしたのはこのような理由からである．シミュレーションにおける月齢分布（**図 12**）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い，アクティブな投稿系列数の時間推移を**図 13** に示す．初期の過渡状態を過ぎると，ほぼ安定した投稿系列数 16,600 前後を保っていることが分かる．つまり，消滅する投稿系列も毎月ほぼ 1,000 程度で安定しているということになる．

単位時間あたり n 投稿系列が生成されるので投稿系列総数 S は，

$$S = \sum_{t=0}^{\infty} nf(t) \quad (8)$$

☆ 相補誤差関数 $erfc(x)$ は次のように定義される．

$$erfc(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-x^2} dx$$

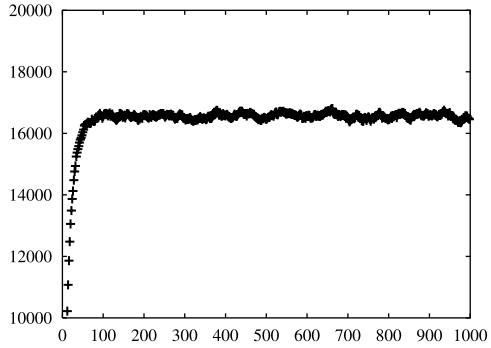

図 13 アクティブな投稿系列数の時間推移（シミュレーション）
Fig. 13 Time evalution of the number of active posting sequences (simulation).

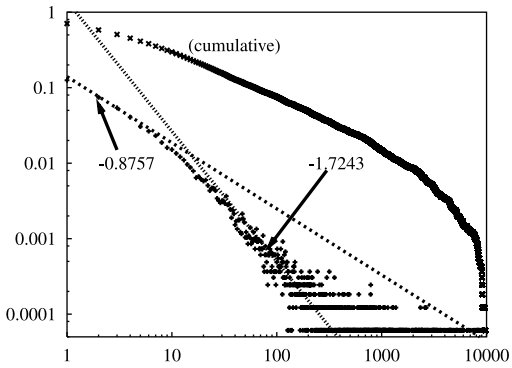

図 14 投稿数の分布と累積分布（シミュレーション）
Fig. 14 Lifetime distribution (simulation).

となる．近似式 (6) を用いれば，$S \sim n/(1-e^{-\mu^2/2\sigma^2})$ となる．パラメータ値から $S \sim 22700$ と計算され，シミュレーション結果の 16,600 からは少しずれるが，この違いは図 13 において t が小さいときのシミュレーション曲線と指数関数の違い分に相当すると考えられる．

上記のシミュレーションの最終月における投稿数 x_{ij} の分布および累積分布を**図 14** 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 3 とよく似たグラフととなっていることが分かる．

投稿数分布 $h(x)$ は t を連続近似すれば，

$$h(x) = \int_0^\infty g(t)p_t(x)dt \qquad (9)$$

となる．投稿系列月齢分布 $g(t)$ は，式 (6)，(7) より指数関数的であるので，$g(t) = \lambda e^{-\lambda t}$ とすれば，$h(x)$ は次のような 2 重パレート分布となる[11),12)]．

$$h(x) = \begin{cases} C\left(\frac{x}{x_0}\right)^{-\gamma_1} & \text{if } x \le x_0 \\ C\left(\frac{x}{x_0}\right)^{-\gamma_2} & \text{if } x \ge x_0 \end{cases} \qquad (10)$$

ただし，

$$C = \frac{\lambda}{x_0\sigma\sqrt{(\mu/\sigma)^2 + 2\lambda}}$$

$$\gamma_1 = 1 - \frac{\mu}{\sigma^2} - \sqrt{\left(\frac{\mu}{\sigma^2}\right)^2 + \frac{2}{\sigma^2}\lambda}$$

$$\gamma_2 = 1 - \frac{\mu}{\sigma^2} + \sqrt{\left(\frac{\mu}{\sigma^2}\right)^2 + \frac{2}{\sigma^2}\lambda}$$

である．また，近似式 (6) より $\lambda = \mu^2/(2\sigma^2)$ とするとべき指数 γ_1，γ_2 は，

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

となる．シミュレーションで用いたパラメータ値では，べきの指数はそれぞれ $\gamma_1 = 0.8757$，$\gamma_2 = 1.7243$ 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 14），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らぎは，記事数に比例しかつ揺ら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ることが分かった．このような揺らぎを持つ時系列は経済現象などでよく見られ，経済物理などでよく用いられる乗算確率過程でモデル化した．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こ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，投稿数がある閾値以下で投稿系列が消滅するとした生成消滅の仕組みを導入した．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．

乗算確率過程でモデル化されるようなシステムでは，揺らぎが大きくブロードな分布を持つ．このような振舞いは予測が困難な経済指標などでよく観測され，これが知識共有サイトの投稿活動でも見られることは興味深い．このような特徴は，揺らぎが環境から加わる雑音ではなく，系自身が揺らぎを内包する人間の活動によく見られる現象なのかもしれない．

本論文の投稿系列モデルでは，投稿数がある閾値以下で系列が消滅するとし，一方で一定の割合で新規投稿系列の生成をいう仕組みを導入した．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．今回，参加者や掲示板の特徴やそれらの間の相互作用などは考慮しなかった．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．システムの構成要素の相互作用が入ることにより，アクティブな投稿系列数のダイナミックスが変わるなど非定常モデルへの拡張という課題についても考えたい．

参 考 文 献

1) Albert, R. and Barabási, A.-L.: Statistical mechanics of complex networks, *Rev. Mod. Phys.*, Vol.74, pp.47–97 (2002).
2) Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. and Hwang, D.-U.: Complex networks: Structure and dynamics, *Phys. Rep.*, Vol.424, pp.175–308 (2006).
3) Cody, W.J.: Rational Chebyshev Approximations for the Error Function, *Math. Comp.*, Vol.23, No.107, pp.631–638 (1969).
4) Dorogovtsev, S.N. and Mendes, J.F.F.: Evolution of networks, *Adv. Phys.*, Vol.51, No.4, pp.1079–1187 (2002).
5) Fujiwara, Y., Aoyama, H., Guilmi, C.D., Souma, W. and Gallegati, M.: Gibrat and Pareto–Zipf revisited with European firms, *Physica A*, Vol.344, pp.112–116 (2004).
6) Fujiwara, Y., Guilmi, C.D., Aoyama, H., Gallegati, M. and Souma, W.: Do Pareto–Zipf and Gibrat laws hold true? An analysis with European firms, *Physica A*, Vol.335, pp.197–216 (2004).
7) Gibrat, R.: *Les inégalité économiques*, Recueil Sirey, Paris (1931).
8) Goh, K.-I., Eom, Y.-H., Jeong, H., Kahng, B. and Kim, D.: Structure and evolution of online social relationships: Heterogeneity in unrestricted discussions, *Phys. Rev. E*, Vol.73, p.066123 (2006).
9) Huberman, B.A. and Adamic, L.A.: Evolutionary Dynamics of the World Wide Web (1999). arXiv:cond-mat/9901071
10) 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).
11) Mitzenmacher, M.: A Brief History of Generative Models for Power Law and Lognormal Distributions, *Internet Math.*, Vol.1, No.2, pp.226–250 (2003).
12) Mitzenmacher, M.: Dynamic Models for File Sizes and Double Pareto Distributions, *Internet Math.*, Vol.1, No.3, pp.305–333 (2003).
13) 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).
14) Newman, M.E.J.: The Structure and Function of Complex Networks, *SIAM Rev.*, Vol.45, No.2, pp.167–256 (2003).
15) 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).
16) Ramasco, J.J., Dorogovtsev, S.N. and Pastor-Satorras, R.: Self-organization of collaboration networks, *Phys. Rev. E*, Vol.70, p.036106 (2004).
17) Reed, W.J.: The Pareto law of incomes—An explanation and an extension, *Physica A*, Vol.319, pp.469–486 (2003).
18) Reed, W.J. and Jorgensen, M.: The Double Pareto-Lognormal Distribution—A New Parametric Model for Size Distributions, *Communications in Statistics: Theory and Methods*, Vol.33, No.8, pp.1733–1753 (2004).
19) Sornette, D.: Multiplicative processes and power laws, *Phys. Rev. E*, Vol.57, pp.4811–4813 (1998).
20) Sornette, D. and Cont, R.: Convergent Multiplicative Processes Repelled from Zero: Power Laws and Truncated Power Laws, *J. Phy. I France*, Vol.7, pp.431–444 (1997).
21) Souma, W.: Multiplicative stochastic process in Econophysics (in Japanese), *Proc. 9th Workshop on Informationn-Based Induction Sciences* (*IBIS2006*), pp.192–199 (2006).
22) Sutton, J.: Gibrat's legacy, *J. Econ. Lit.*,

Vol.35, pp.40–59 (1997).
23) 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).

(平成 19 年 2 月 2 日受付)
(平成 19 年 3 月 23 日再受付)
(平成 19 年 4 月 4 日採録)

新井　賢一

昭和 42 年生．平成 5 年早稲田大学大学院理工学研究科物理学及応用物理学専攻修士課程修了．同年日本電信電話（株）入社．現在 NTT コミュニケーション科学基礎研究所主任研究員．ニューラルネットワークの学習理論，非線形力学理論，複雑ネットワーク理論等の研究に従事．博士（理学）．日本物理学会会員．

山田　武士（正会員）

昭和 39 年生．昭和 63 年 3 月東京大学理学部数学科卒業．同年 NTT 入社．平成 8 年より 1 年間英国コベントリー大学客員研究員．現在，NTT コミュニケーション科学基礎研究所創発環境研究グループリーダ．主としてネットワーク分析，機械学習，組合せ最適化等の研究に従事．博士（情報学）．電子情報通信学会，ACM，IEEE 各会員．

林　　幸雄（正会員）

昭和 37 年生．昭和 62 年豊橋技術科学大学電気電子工学専攻修士課程修了．同年富士ゼロックス（株）入社．平成 3〜5 年（株）ATR 視聴覚機構研究所および人間情報通信研究所に出向．平成 9 年より北陸先端科学技術大学院大学助教授．ニューラルネット，Web サイエンス，複雑ネットワーク科学に関する研究に従事．博士（工学）．電子情報通信学会，日本応用数理学会各会員．