2008－MPS－71 (14)
2008／9／19

利用者の発言頻度の規則性に基づく電子フォーラムのコミュニティサイズ推定法

久保 正男 †, 成瀬 継太郎 ††, 佐藤 浩 †, 円田 孝志 † 松原 隆 †

† 防衛大学校 †† 会津大学

従来、電子フォーラムや掲示板への投稿は個々の利用者の自発的な行動の結果であり、そこには大局的な特徴はないと思われていた．これに対し、利用者ごとの投稿数とその頻度の分布は対数正規分布に従う規則性があることを明らかにする．この規則性は実際の掲示板には現れていない投稿数0件の利用者、すなわち閲覧だけを行う利用者数の推定を可能にする．これにより、従来ではアクセスログなど電子フォーラムの管理者しか知り得なかった閲覧者を含む利用者数を、誰もが観測できる投稿数の分布から推定することができる．本論文では、この新しい原理に基づくコミュニティサイズ推定法を提案し、実際の既知のコミュニティサイズと比較することにより、その有効性を検証している．

Population Estimation of Anonymous Online Forum Community by Posted Article Distribution

Masao Kubo † Keitaro Naruse †† Hiroshi Sato † Takashi Enda † Takashi Matsubara †

† National Defense Academy

†† University of Aizu

Traditionally, it seems that there are no typical patterns or regularities on an electronic forum and the bulletin board because it is a result of voluntary behavior of an individual user. However, recent studies on them including us show distribution of the number of posted articles per user on them follows as a lognormal distribution and a power law distribution. We think that this regularity of the emergent behaviour is very important for understanding social science and human communication, not only computer science because it is possible to estimate the number of lurkers of each bulletin board and forum thread only using by usually observable data. In this paper, we proposed a set of estimation methods based on preferential attachment which is caused by reciprocity as a human trait. The effectiveness of our proposed estimation methods are verified by comparing it with indices to be strongly related to the population of its corresponding human behaviour.

1 はじめに

ここでのコミュニティとは，(1) 特定の掲示板に投稿を行う可能性がある人からなる集団で，(2) 実際に記事を掲示板に書き込んだ人と閲覧だけしているの人から構成される．また，投稿された記事はコミュニティのすべての人に瞬時に伝わり，投稿はこの集団に属する人からのみ行われると仮定する．またコミュニティは常に変動しているはずだが，ここでは問題を簡単にするために一日単位でメンバの変化が起こるものとする．

さて、コミュニティサイズを算出するには投稿した人と閲覧だけをしている人の数を求める必要がある．投稿した人の数は記事につけられているユーザ ID から誰でも算出することができる．一方，閲覧だけしている人の数を厳密に求めることは、他の web ページの閲覧者数や訪問者数を決めることと同様に決して容易なことではない [4) 5)]．一般に，掲示板の管理者がアクセス情報などを開示しない限り，ユーザは他の閲覧者の情報を知ることはできない．一方，掲示板の管理者であれば，サーバへのリクエストを観測することによって可能なように思える．しかし厳密には (1) 巡回ソフトウェアによる多くの機械的なアクセスの除去，(2) アドレスの打ち間違いなどによる誤アクセスの判定，(3) ユニークアドレスの抽出などが必要で，これには莫大なコストを要する．そのため，より簡便な方法 (例えば一定時間内のアクセス数) で代用するのが一般的である．さらに，(4) キャッシュやプロキシーサーバの利用者は web サイトにアクセスしなくてもページを閲覧できるので，管理者でも正確には知り得ない．我々の知るところこれらを克服した枚挙的閲覧者数算出手法は提案されていない．

ところで、人間の自発的な活動の結果できた人工物の中には思いもよらない規則性が見られることが報告されている．例えば英語の単語の使用頻度の順位がべ

キ分布に従うことが知られている[10]．またインターネットの web サイトのリンク接続数がベキ分布に従うことが明らかにされている．もしこのような規則性が電子掲示板にみられるのであれば、この性質から掲示板のコミュニティの属性を推定することも可能であろう．先行研究[9]では、話題や参加者が異なるにもかかわらず多くの掲示板では投稿回数の頻度に同様の規則性が乗算過程によって生まれることが報告されている．本論では、返事がくると一層投稿しやすくなるという返報性[7]を仮定し、利用者間の相互作用のモデル化を行う．そして、そのモデルから投稿回数ごとの人数の分布関数を導出し、コミュニティサイズの概算を実現する．

この概算方法の特筆するべき点として、(1) 書き込み内容を考慮しない、(2) ユーザ ID や書き込み時間などの誰もがアクセスできる情報のみを利用している点にある．これによって，従来困難であったが，ユーザでもコミュニティサイズの概算が可能となる．

2　掲示板における投稿数分布の解析

2.1　対象掲示板

議論の一貫性を保つために，対象を大量のデータを含む単一の掲示板に絞るものとし，誰でも容易に投稿できるという観点からここでは 2 ちゃんねる[1]という巨大匿名掲示板群を対象とする．

2.2　各スレッドに共通の性質

データの取得は，2005 年 6 月から 8 月の期間に行った (総スレッド数 327,282)．この中から 1 日あたりの記事数が十分多い（400 件以上）スレッドを対象に解析を行った (対象スレッド数 585)．

Fig.1 は，三つのスレッドにおける単位時間あたりの投稿記事数の推移を表している．ここでは，横軸は時間（秒）を線形のスケールで，縦軸は投稿数を対数スケールで表している．赤線 (+) はあるテレビ番組に関するスレッドであり，番組放送開始と同時に記事が殺到し，スレッドの最大記事数に短時間に到達した様子を示している．一方，青線 (○) は株式の市況に関するスレットであり，市場が開いている間は頻繁に書き込みがあるが，市場が閉じるとほとんど投稿は見られない．最後に，緑線 (□) は時事ニュースに関するスレッドであり，定常的に一定量の記事が投稿されていることを示している．他のスレッドについても投稿数の推移を観察したが，各スレッドのテーマやそこに集まる人々により大きく変化し，各スレッドに共通の性質は見出せない．

次に，スレッドにおける各投稿者の記事数の分布を観察する．Fig.2 は，同三つのスレッドにおいて，各投稿者がそのスレッドに何件の記事を書き込んだかを示している．単位時間当たりの投稿記事数の推移は異

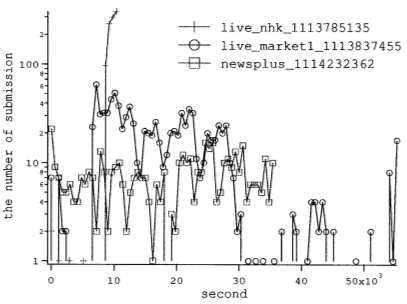

Fig. 1　単位時間当たりの投稿記事数の推移.

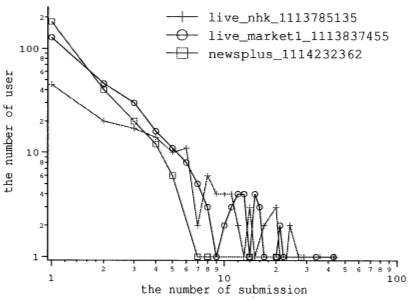

Fig. 2　各投稿者当たりの記事数の分布.

なる三つのスレッドであるが，両対数スケールにおける投稿者の分布は，ほぼ同じように直線的に減少している．この傾向は，他の大部分のスレッドでも同様の分布が観察された．

そこで，これらの実際の分布がどのような統計モデル（分布）により良く表現されるかを AIC（赤池の情報量基準）により検討した．

最も小さい AIC になった分布のスレッド数を Table.1 に示す．ほとんどの掲示板 (98.6%) で対数正規分布が最もよく投稿回数の人数分布を表していることがわかった．対数正規分布は両対数スケール上で直線状になる性質を持つ．以上から多くの掲示板において共通する性質がみられることがわかった．

3　返報性に基づく投稿モデルの提案

掲示板のコミュニティは潜在的に投稿者になる人から構成されているものと仮定したが，その目的が情

Table 1　AIC による最良モデル

	Normal	Exponential	Lognormal
Best Model	0.0 % (0 thread)	1.37 % (8 threads)	98.63 % (576 threads)

報収集あるいは議論など何らかのコミュニケーションであれば，自分の投稿した記事への返信などの反応があった場合，記事を投稿する頻度が高くなるであろう．実際，Joyce らは初めて投稿したものが再度投稿する確率は返事をもらうか否かで 12%向上することを報告している [7]．一方，この性質が成り立つ時には投稿回数が多いものほど更に投稿する機会が増えると考えられる．これは前節で明らかにした「両対数スケール上で直線状」になるものの一つであるベキ分布を発生する一つのダイナミクスであるプリファレンシャルアタッチメント (以下 PA) と類似している [6]．そこで，このダイナミクスを用いて，投稿ダイナミクスの定式化を試みる

3.1 提案モデル

バラバシらの方法 [6] とここで扱う掲示板では次のような違いがある．ここではノードをコミュニティの各メンバと捉えるので，(1) 初期ノード数は無視できない，(2) 一日という短い時間間隔なのでこの間に利用者の総数にほとんど変化がないものとする．これを踏まえ，本論では次式の投稿モデルを提案する．

まず、一単位時間に一件の投稿があるものと仮定し、時刻 t での総投稿数を t とする．メンバ i の投稿数を $n_i(t)$ とする．また，コミュニティの総数を N とする．この時、投稿者 i が時刻 t に掲示板に投稿する確率 $\Pi(i)$ を次式で与える

$$\Pi(i) = \frac{wn_i + 1}{\sum_{j=1}^{N}(wn_j + 1)} \tag{1}$$

w は一回の投稿が投稿確率に与える重みであり $w \geq 0$ とする．投稿数が多い投稿者は高い確率で再投稿し，その確率は投稿数 n_i に比例する．また一度も投稿したことがないメンバ (n_i=0) も，低い確率であるが投稿する可能性がある．w=0 の時は各メンバの投稿確率は等しくなり，投稿はランダムに行われることを意味する．

さて，式 1 をバラバシらの方法 [6] 同様に展開すると次ように投稿件数分布 $P(k)$ が得られる．

$$P(k) = \frac{\partial}{\partial k}(1 - \int_0^{(t+\frac{N}{w})\frac{1+\frac{1}{w}}{k+\frac{1}{w}} - \frac{N}{w}} P_{first}(t)dt) \tag{2}$$

$$= \frac{\Gamma(\frac{(1+w)(N+tw)}{w(1+kw)} - \frac{1}{w})}{(1+kw)\Gamma(\frac{(1+w)(N+tw)}{w(1+kw)})} \tag{3}$$

となる．ここで $P_{first}(t_i)$ はユーザー i が初めて掲示板に投稿する時刻が t_i である確率である．

4 コミュニティサイズの同定

4.1 計測時間が既知な時の閲覧者数の推定原理

実測した投稿件数から式 3 の分布関数 $P(k)$ の各パラメータを求めることができれば，コミュニティサイズが推定できるはずである．

いま時刻 t まで、0 回投稿した人の数を $Post_0(t)$, 同様に x 回投稿した状態にいる人の数を $Post_x(t)$ とする．またその期待値を表すときには “ $\bar{}$ ” を用いる．いま、時刻 $t+1$ に 0 回投稿した人の数 $Post_0(t+1)$ は式 1 より確率 $\frac{Post_0(t)}{tw+N}$ で 1 減る．したがって，期待値は次のように遷移する．

$$\overline{Post_0}(t+1) = \overline{Post_0}(t) - \frac{\overline{Post_0}(t)}{tw+N} \tag{4}$$

この性質を利用すると，時刻 t において x 回投稿した人と $(x+1)$ 回投稿した人の比率 $Q_{x/x+1}(t)$ は

$$Q_{x/x+1}(t) = \frac{\overline{Post_x}(t)}{\overline{Post_{x+1}}(t)} = \frac{(x+1)(-1+N+(-(x+1)+t)w)}{(-x+t)(1+xw)} \tag{5}$$

である．したがって、時刻 t において，例えば隣り合う 3 投稿数の人数が与えられたならば，式 5 を連立することによって N および w を求めることができる．

以上から、もしコミュニティが掲示板が開かれている間ほとんど変化せず、w が一定と仮定できる時，コミュニティのサイズを推定する方法を提案した．

5 検証

各掲示板の真のコミュニティサイズを得る事は困難であるが，各掲示板の利用者の総計はサーバのアクセス量と密接な関係があると思われる．加えて，投稿せず閲覧だけしている人のアクセスがサーバに与える影響は無視できるほど小さいとは考えにくい．したがって，もし，コミュニティサイズが正しく推定できれば，投稿者のみをコミュニティとする場合に比べ，より緊密な相関係数が得られることが予想される．

5.1 使用したデータ

深夜時間帯以外の書き込みが多い掲示板が数多く運営されているサーバとして ex13.2ch.net を選んだ．次に 2006 年の 2 月から 2006 年 3 月まで，このサーバの掲示板データを収集し，計 33,069 掲示板，計 3,128,155 件の書き込みデータを得た．次に 2 月の 2 週目から 4 週目までの各日について各掲示板のコミュニティサイズを推定し，これの合計とこのサーバのアクセス量を比較した．

サーバのアクセス量については pv.kakiko.com で一般公開されているデータを用いた．このサイトには html, read.cgi, dat, text,pic などサーバが扱うファイルの種類ごとのアクセス数が公開されている

5.2 推定結果

Fig.3 の左軸に補正を組み入れた推定方法 (+) および投稿者のみをコミュニティとする場合のコミュニティサイズ (○)，右軸にサーバのアクセス量として read.cgi 値 (点線) をプロットした．これは read.cgi 値が，巡回エージェントによる影響が最も低いと考えられるからである．横軸は日付で 2006 年 2 月 9 日からの 2 月 28 日までを表している．また，相関係数の比較の結果を

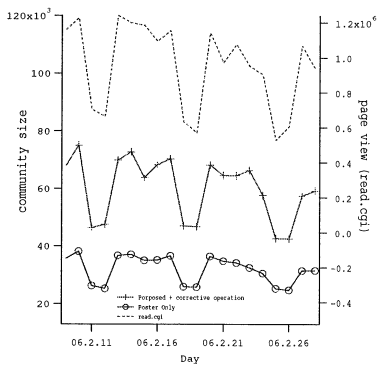

Fig. 3 サーバアクセス量と推定コミュニティサイズ:ex13.2ch.net

Table.2 に示す．提案手法の場合 0.9533 で，投稿者だけをコミュニティとした場合の 0.9433 を上回った．

ここではサーバのアクセス量とサーバに含まれるすべての掲示板のコミュニティサイズの和を比較した．閲覧だけをする人々によるサーバのアクセス量が無視できるほど少ないとは考えにくいので，提案手法が正しく機能すれば，投稿者のみをコミュニティとする場合よりも強い相関が得られるはずである．実験の結果，投稿者数のみをコミュニティとするものよりも，高い相関係数が得られることができた．これによって提案した投稿モデルの正当性を定量的に検証できた．

Table 2 サーバへのアクセス量との相関：ピアソンの相関係数

	Only Poster Data	*Proposed*
total	0.9765	0.9781
read.cgi	0.9433	0.9533
html	0.9461	0.9521
dat	0.9811	0.9820

6 おわりに

本論では，電子フォーラムのコミュニティサイズを管理者でなくとも概算できる手法を，電子フォーラムの投稿ダイナミクスモデルの検証という形で提案し，サーバのアクセス量とサーバが運営するすべての掲示板のコミュニティのサイズ合計との相関を比較した．その結果，投稿者のみをコミュニティとする場合よりも提案手法によるコミュニティサイズを用いた方が高い相関を得ることができた．以上から，本提案手法の有効性，妥当性を定量的に確認した．

参考文献

1) 2 ちゃんねる, http://www.2ch.net/

2) 2 ちゃんねる監修, “2 ちゃんねる公式ガイド 2004”, コアマガジン, 2004.

3) R. Albert and A.-László Barabási, Statistical mechanics of complex networks, Reviews of Modern Physics, vol. 74, no.1 pp. 47–97 (2002).

4) http://www.tlsoft.com/ arbitron/jul95/ arbitron.summary.txt

5) Baldi P., Frasconi P., and Smyth P. 著, 水田 正弘, 南 弘征/小宮 由里子 (訳) :確率モデルによる Web データ解析法 - データマイニング技法から e-コマースまで, 森北出版 (2007).

6) A.-L. Barabási, R. Albert, and H. Jeong, Mean-field theory for scale-free random networks, Physica A, Vol. 272,pp. 173–187, 1999.

7) Joyce, Elisabeth,Kraut, Robert E., Predicting Continued Participation in Newsgroups,Journal of Computer-Mediated Communication, Volume 11,Number 3, pp. 723-747(25)(2006)

8) Massen C.P. and Doye J. P.K. :A self-consistent approach to measure preferential attachment in networks and its application to an inherent structure network,Physica A: Statistical Mechanics and its Applications, Volume 377, Issue 1, 1,pp351-362(2007).

9) Naruse Keitaro and Kubo Masao : Lognormal Distribution of BBS Articles and its Social and Generative Mechanism, Web Intelligence, Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, pp103-112(2006)

10) http://www.nslij-genetics.org/wli/zipf/