様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成　24 年　6 月　11 日現在

機関番号：８２７２３
研究種目：基盤研究(C)
研究期間：2009～2011
課題番号：21510172
研究課題名（和文）　ウェブ上フォーラムのコミュニティサイズ推定

研究課題名（英文）　Community size estimation of Internet Forums

研究代表者
久保　正男（KUBO MASAO）
防衛大学校・電気情報学群・准教授
研究者番号：30292048

研究成果の概要（和文）：
インターネット上の安全なコミュニケーションをサポートする技術の一つとして，電子掲示板のコミュニティのサイズを誰でも概算できる手法を開発した．ここでのコミュニティとは将来的に投稿をおこなう人のことで，サーバーの管理者でも容易ではない．本研究では，既知である投稿者数から集団の普遍的な特性を利用することによって未投稿者数を推定する．サーバーの負荷や視聴率を用いた検証から，提案手法は単純に投稿者数をコミュニティとするよりも自然な結果が得られることがわかった．したがって，この方法を利用者が用いればコミュニティの規模や自身の発言がもたらす影響を事前に認識しやすくなると考えられる．

研究成果の概要（英文）：
The technique that can estimate size of the community of Internet forums as a technology to support safe communication in the Internet is proposed. The community here is defined as a group of person contributing it in the future. The proposed estimates the number of contributors who have not posted yet. This is not easy for even the administrator of the server. Using the invariant emergent property of the group the number of the remaining member is able to be estimated by the number of the members who have posted so far. I showed that the proposed method can provide a more reasonable community than a community consists of only posted members by the experiments using server load data of the internet forums and TV viewing rates.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,500,000	0	1,500,000
2010 年度	1,000,000	0	1,000,000
2011 年度	1,000,000	0	1,000,000
総　計	3,500,000	0	3,500,000

研究分野：複合新領域
科研費の分科・細目：社会・安全システム科学・社会システム工学・安全システム
キーワード：社会システム，データマイニング，ソーシャルネットワークサービス，消費者行動，マーケティング

１．研究開始当初の背景

対面では起こりにくい人間模様がネット上では起こりやすいことの原因の一つとして，コミュニケーションの様相を決めるために必要な情報が知覚困難になる為と考えた．その中でもコミュニティのサイズは重要な

項目であろう．電子掲示板の先には大勢の聴衆がいることが明確に把握できれば，犯罪に係る行為や不用意な行動は慎まれると考えた．

２．研究の目的

電子掲示板におけるコミュニティの規模を推定するには他人の振る舞いを知る必要があり，通常管理者でなければ困難である．しかし，過去の研究から電子掲示板の投稿には一人当たりの投稿回数とその該当者数の間には多くの掲示板に共通の普遍的な性質があることがわかっていた．そこで，この分布が生成される原理とそれに基づく人数推定方法の確立および検証をおこなうことを目的とした．

３．研究の方法

ここで着目した特性とはベキ分布という自然界でよく現れる特性で，単語の頻度に関する Zipf の法則が有名である．ここでははじめに発言したものほどより多く発言しているという定性的な特性に着目して，バラバシらが提案する BA モデルを生成原理とする推定手法の構築を試みた．

まず．エージェントシミュレーションをおこなった．各エージェントは掲示板の投稿者に相当し，既存の発言に意見を書き込む場合と新たな話題を投稿する．これをエージェント数を一定のもとでシミュレーションをおこなって，現実の掲示板と比較した．その結果，２者での言い争いのような現象がおこらなければ，ベキ状の分布になることも多いと考えられた．

そこで推定モデルを構築した．BA モデルは全連結にある極めて小規模なグラフにノードが逐次的に追加される成長を扱うモデルである．新規ノードと既存のグラフとのリンクは各ノードの次数に比例する．この場合，ノードの次数とその頻度との間にはベキ分布がなりたつことが知られている．ここでは有限のノード集団から逐次的に一人ずつ掲示板に書き込むと仮定して，待ち行列理論をもちいて総ノード数の推定方法を提案した．

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

このモデルは過去の発言の重要度 w と総人数(コミュニティサイズ)N からなる．この二つのパラメータを実データから推定する方法を開発した．

４．研究成果

(1)推定方法の構築

掲示板開設から十分に時間が経過した場合と各投稿時刻が既知である場合についてコミュニティサイズを推定する方法を提案した．前者ではこのモデルが十分に時間が経過した後では投稿回数比(P(k)/P(k+1))が一定になることに着目し，w を最小二乗法などで推定する．

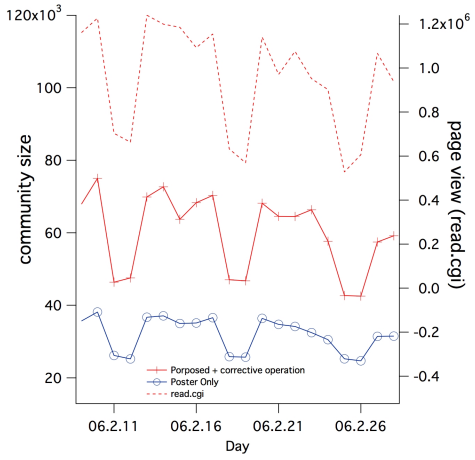
$$\lim_{t\to\infty} \frac{P(K)}{P(K+1)} = \left(\frac{1+(1+k)w}{1+kw}\right)^{\frac{w-1}{w}}$$

その後，１回だけ投稿したメンバ数が分かれば未知数である０回投稿者の人数を推定できる．

後者では，投稿回数に準じた状態遷移モデルから，総数 N を推定する方法を提案した．これによって人数推定に必要な最小限のデータは連続する 3 投稿回数のメンバ数だけでよく，実測する際のデータの欠けやばらつきに対応しやすくなった．

また，推定した結果 w の値に応じてヒューリスティックルールを提案した．

図１　サーバーの負荷との比較



(2) ３つのケースで検証

提案した手法を実データに適用して，その検証をおこない妥当性，有効性を確認した．３件のケースを扱った．そのうちの二つは一般に公開されているデータを用いている．

① サーバー負荷による検証

図１に示すのはある電子掲示板(2 ちゃんねる)が運営されているサーバーの約３週間のアクセス量の変化(pv.kakiko.com を参照した)と含まれる計 33,069 の掲示板のコミュニティの人口の総和を比較したものである．点線がサーバーへのアクセス量，○で投稿者総数，＋が提案手法によって推定した人口である．サーバーへのアクセス量とのピアソンの積率相関係数は 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．電子掲示板にアクセスするのはすでに投稿した人だけでなく，未だ投稿していない人も多く含まれるはずである．したがって，コミュニティに未投稿者数を含めた提案手法のコミュニティサイズの方がサーバーの負荷とより強い相関があることは自然な結果と考えられる．一方のその差は非常に微小であり，提案手法の有用性を明確にするには至っていない．またサーバーの負荷は各ユーザの振る舞いによって左右されるため，厳密な検証がむずかしいことが指摘できる．

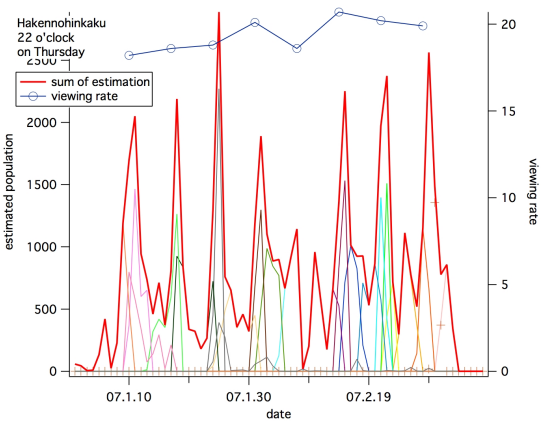


図２推定人口とテレビ番組視聴率の変化

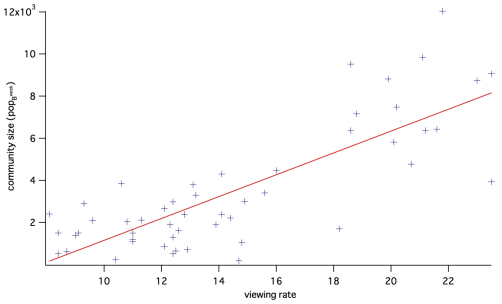


図３　テレビ番組視聴率と相関

	Only Poster Data	Proposed
correlation coeffi-cient	0.7565	0.7987

表１　提案手法の効果の視聴率による評価

② 視聴率をもちいた評価
提案手法を厳密に評価するために，人口に相当するデータでの検証が望ましい．そこで，視聴率を用いた評価をおこなった．視聴率は世帯数を単位とするもので，比較して人口に近しいデータである．図２に示すのは提案手法によるテレビ番組関連の掲示板の推定人口の変化と公開されている視聴率との比較である．関連する掲示板は複数あるので，提案手法により，日ごとの各コミュニティのサイズを算出してその和を太線で示している．右軸にそのテレビ番組の週間視聴率を示す．次に放送前の一週間の人口の総和と視聴率との関係を調べた．７つのテレビ番組について調査した結果，51組の視聴率とコミュニティサイズの関係についてのサンプルが得られ，これを図３に示す．横軸が視聴率で縦軸が推定したコミュニティの人口である．テレビや電子掲示板を閲覧する人になんらかの偏りがなければ，これらの間には線形な関係が生じるはずである．そこで，提案手法によって求めたコミュニティサイズと視聴率とのピアソンの積率相関係数は 0.7987 であった．投稿者数と視聴率との相関係数 0.7565 を上回った．テレビを視聴する人の中には電子掲示板を閲覧するけれども投稿はしていない人が多く含まれると考えるのが自然であり，提案する手法でコミュニティを推定すればより我々の感覚と近い値が得られることがわかった．またこの調査の副産物的な結果として，投稿者数や提案手法を用いたコミュニティサイズから比較的正しく視聴率が推定できることがわかった．

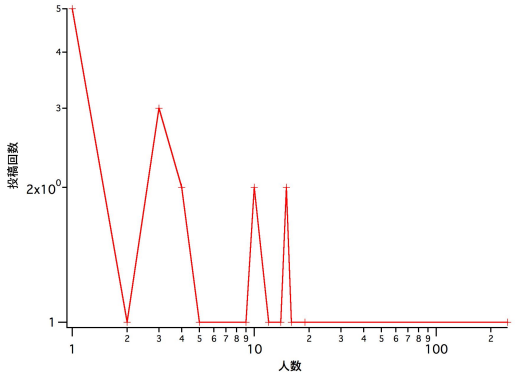


図４　学内掲示板への投稿件数

③サーバー履歴による評価
上記の二つの方法は一般に公開されているデータを用いた検証で，将来第三者による提案手法の検証が期待できる点で有意義である．しかし，提案手法がコミュニティのサイズである以上，人数という単位での検証ができることが好ましい．そこで，より正確に検証を行うため，学内の電子掲示板のログを調査した．学内公開されたある一つの電子掲示板へのアクセスと書き込み数について 2009 年，2010 年のログを調査した．その結果，この掲示板には 2 年間で計 22 人から 396 件の書き込みがあった．一方この期間の内，このウェブページには 2010 年 4 月から 11 月までで 5527 件のアクセスがプ

ロキシサーバーをのぞく 139 台のコンピュータからあった．書き込み件数が検証対象としては十分でなかったが，提案するベキ型モデルの特徴である両対数グラフ上での直線形状をなんとか確認できた．しかし，アクセス履歴より IP アドレスもしくはコンピュータ名が分かるものの，利用者との対応に時間を要し，聴衆数を正確に計測することができなかった．そのため，学内での電子掲示板のログを調査では，提案するモデルの正しさを検証するにいたらなかった．

以上のことから，より正確に検証するためには，より多くの投稿と厳密なユーザ単位でのサーバーへのアクセス履歴の取得が必要なことが改めて浮き彫りになった．

(3)研究結果のまとめ
電子掲示板におけるコミュニティの規模を概算できる手法を開発し，実証実験をおこなった．その結果，実際に適用できることがわかった．これによってユーザは掲示板の管理者の情報開示に頼ることなく自身でコミュニティの規模を見積もることができる．またこの方法の存在が広まることによって改竄なく管理者が利用者情報を開示することが促進されると考えられ，その性能以上に社会的なインパクトの大きい研究と結論できる．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 4 件）
①Masao Kubo, Keitaro Natuse, Hiroshi Sato, community size estimation of internet forum by posted article distribution, innovation in intelligent machine, vol.2,pp225-239,springer-verlag,2012,査読あり

② Masao Kubo, Hiroshi Sato, and Takashi Matsubara, Word Familiarity Distributions to Understand Heaps ‘ Law of Vocabulary Growth of the Internet Forums,LNAI 6883, pp. 627- 636, Springer,2011,査読あり

③Masao Kubo, Hiroshi Sato ,Sub-linearity in Vocabulary Growth in Internet Forums, Proceedings of Joint 5th International Conference on Soft Computing and Intelligent Systems and 11th International Symposium on Advanced Intelligent System,pp570-575,2010,査読あり

④Masao Kubo, Keitaro Naruse, Hiroshi Sato, Takashi Matsubara, Population Estimation of Internet Forum Community by Posted Article Distribution, Lecture Notes in Computer Science,vol.6279, pp298-307,Springer,2010,査読あり

〔学会発表〕（計 2 件）
①Masao Kubo, Keitaro Natuse, Hiroshi Sato, Behavior of Posting Activities of 2ch Users ,Proceedings of 13th Asia Pacific Symposium on Intelligent and Evolutionary Systems,pp1-5,2009 年 12 月 5 日発表，九州大学(福岡県春日市)

②久保正男, 成瀬継太郎，松原隆, 佐藤浩 , 待ち行列理論を用いた発言頻度の規則性からの電子フォーラムのコミュニティサイズ推定,the 19th Intelligent Systems Symposium (Fan 2009) and The 1st international Workshop on Aware Computing (IWAC 2009) ,pp338-341, 2009 年 9 月 18 日発表，会津大学(福島県会津若松市)

６．研究組織
(1)研究代表者
久保正男（KUBO MASAO）
防衛大学校・電気情報学群・准教授
研究者番号：30292048

(2)研究分担者
なし
(3)連携研究者
なし