Vol.2010-HCI-138 No.3
2010/5/14

ロングテール時代のための統計分析ツール tailstat の開発

栗 原 一 貴[†1]　土 谷 洋 平[†2]

本論文では母集団分布が既知であり，また中心極限定理が適用できないような小標本の統計分析が現代のインターネット情報社会では重要性も持ってきていることを示し，そのような問題領域における統計手法の再構築と，それを実現する統計分析パッケージ tailstat の開発を行う．また，ソーシャルブックマークサイトにおける具体的な大規模分析事例により tailstat による分析が有効であることを示し，他の事例へ応用するための考察を行う．

Statistical Analysis Tool tailstat for "Long-tailed" Internet Society

Kazutaka Kurihara[†1] and Yohei Tsuchiya[†2]

This article presents a statistical analysis method and introduces the corresponding software package "tailstat," which is believed to be widely applicable to today's internet society. The proposed method facilitates statistical analyses with small sample sets from given populations, which render the central limit theorem inapplicable. A large-scale case study on a social bookmark service demonstrates the effectiveness of the method and provides implications for applying similar analyses to other cases.

1. はじめに

近年インターネットの普及により，大企業や大組織が市民に対し画一的，一方通行的に情報や物資を供給してきた旧来の社会構造に変化が生じてきている．オンライン書店である amazon.com[1] の売り上げは，その約三分の一がこれまで一般の書店が取り扱わなかったような「あまり売れない本」で構成されていることが報告され，世間を驚かせた[10]．これは通常「よく売れる本を店頭に並べることで売り上げを増大させる」という常識を覆した事例である．また， YouTube[14] や Yahoo!オークション[13] などのように，一般市民が主導権を持って提供した情報や物資の総体としてサービスが成立している事例も枚挙のいとまがない．これも従来そのような事業を担ってきた流通・マスメディア業界とは全く異なる仕組みにより運営されており，コストなどの既成概念に画一化されず多様性豊かなコンテンツを育む土壌となっている．これらの潮流は，一般市民による情報発信への参画や小規模で多様なコンテンツの総体が価値を持ってきた様子をとらえて「Web2.0」や「ロングテール」といった言葉で表現され，時代を象徴する言葉となっている[10]．

このような時代背景にあって，我々が社会分析，経営分析等に用いている統計手法，特にある母集団から得られた標本を分析する手法は依然として限定されている．従来の一般的な統計手法は（1）一部の特権的な分析者が（2）膨大なデータに対して（3）平均的・全体的な振る舞いを推測する　ために多く用いられてきた．マスメディアが行う世論調査がその代表例である．これは独自の情報網もしくは資金力を生かし（たとえば千人規模の国民に無作為に電話をかけ），アンケート結果から平均的な日本人の思考を推測するものである．また，経営学における「客単価」の推定もこの範疇であり，たくさんの来客数とそれに対する総売り上げから，平均的な顧客の振る舞いを推測するものである．これらは冒頭に述べたような大規模画一的かつ一方通行的な社会構造においては重要な役割を果たしてきた．

しかしロングテールの時代において，これら（1）（2）（3）の仮定が適用できない事例が多く見られる．たとえば，オンライン書店の「あまり売れない本」がいま3冊売れたとき，それはどのくらいの珍事件であり，市場価値を持つのか，という問題を考えてみる．このような分析興味は，一般市民による情報発信への参画が強い意味を持つ現代において重要なものであるが，従来は（1）の観点からそもそも「売れない本」として切り捨てられがちであった．また，このような分析においては（2）および（3）に反して，サンプル数が少ないため平均的な振る舞いを想定できない．なぜならば，従来の統計手法は基本的にサンプル数が多いという条件を仮定する中心極限定理とそれに伴う正規分布標本論に支えられて成立しているからである．

一方で，インターネットの普及は統計分析に変革をもたらした面もある．従来はそもそも分析対象としている事象の母集団分布を得ることが難しかったため，十分な数のサンプルを得て推測統計学を用いた分析をすることが一般的であった．しかし近年においては，情報技

†1 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology
†2 神奈川工科大学 (基礎・教養教育センター)
Kanagawa institute of technology

© 2010 Information Processing Society of Japan

術の恩恵により母集団分布が比較的容易に得られ，またそれを公開していく機運が高まっている．たとえば，ニコニコ動画[16] においては動画の閲覧数，動画に関連する商品の購入者数などが公開されており，また livedoor は自社のソーシャルブックマークサービスのほぼすべてのデータセットを公開している[8]．これらは，あるサービスという経済圏における活動のすべてが包含されているため，母集団分布と呼んで差し支えないだろう．

我々は，このような時代背景にふさわしい統計手法の再構築を研究している．本論文ではその第一段階として，母集団分布が既知である場合の標本和および標本平均の分布の分析手法を提案する．従来，母集団分布が未知である場合，標本和および標本平均の分布が漸近的に正規分布であることを保証する中心極限定理が強力なツールとして多用されてきた．一方で母集団分布が既知である場合は，小標本・大標本を問わず母集団分布のたたみこみを計算することにより標本分布をより直接的に，厳密に計算可能である．たたみこみは計算量の多い演算であるが，近年発展の著しい計算機の計算能力を活用することで実用的な性能を達成可能である．このことを用いて，既知の母集団分布を入力として各種分析を行う統計分析パッケージ tailstat を開発する．

tailstat を用いることで，個々のコンテンツ発信者およびサービス提供者双方に利益がもたらされる可能性がある．個々のコンテンツ発信者は自分の発信するコンテンツの価値を定量的に見積もれるようになり，よりコンテンツの価値を高める投資をすべきかどうかの判断などを行うことが可能となる．一方サービス提供者は「今それほど有名ではないが，今後注目のコンテンツ」という新しい観点でのデータマイニングや情報推薦，もしくは異常事態の自動検知などが可能となり，サービス全体の価値の向上に貢献できる．

本論文では，まず関連研究を挙げ，その後 tailstat の実装と機能を紹介し，tailstat を用いた分析例を挙げてその有効性を議論する．

2. 関連研究

現代の情報化社会における統計データおよび統計手法の一般市民への開放について，関連する試みを挙げる．web サイトの訪問者数や訪問履歴がそのサイトの価値を測る上で重要な指標とみなされるようになり，自分の運営するサイト上におけるサイト閲覧者の行動を分析する手法が発達してきている．Google Analytics[4] は代表的なものであり，これは任意のサイトについて，そのような統計の取得と分析とを無料で代行するサービスである．また，総務省統計局のウェブサイト[20] のように国が管轄の統計情報を公開することは旧来より行われてきているが，経済データのグラフ作成・共有サイトである vizoo[11] や，Public Data Sets on Amazon Web Services[9] などのように，民間の様々な統計情報を民間主導で共有していくというコンセプトのサービスも現れ始めている．このような動きから，統計を行う主体が少数の管理的立場の人間から一般市民へと広がってきているとともに，そのような分析を行う上で必要なデータも豊富になってきているといえる．

次に，本論文で提案する統計手法が対象にしているような，いわゆる web2.0 的なサービスについて，一般ユーザのサービスへの関わり方を分類軸にして代表的なものを挙げる．まずは，いわゆるネットショッピングのサービスである．楽天[21] のように個別に管理された小さなネット商店を束ねる構造のものも見受けられるが，書店の amazon.com[1] のようにサイト運営者が一括して商売を管理している場合もある．これらにおいては基本的には利用者がサイトを訪問し，金銭を支払い商品を購入しているが，購入者による商品の使用感などが商品ごとに投稿・掲載されていることもあり（「口コミ」とよばれることがある），新規訪問者の購買活動に影響を与えていると考えられる．

一方で，商品・コンテンツを企業だけでなく，一般市民から募集し公開することで成立しているサービスも多い．商品取り引きを行う例では Yahoo!オークション[13] などのネットオークションが挙げれらるが，サイト訪問者が無料でコンテンツを視聴できる例も多く，動画共有サービスである YouTube[14] やニコニコ動画[16]，写真共有サイト flickr[2]，また料理レシピ共有サービスであるクックパッド[15]，百科事典 Wikipedia[12] などが挙げられる．これらにおいてもネットショッピングの例のように，コンテンツに関する閲覧者の感想・コメント・レーティングなどが許されている場合が多く，総視聴回数とともにそのような閲覧者の反響の程度がコンテンツの価値を決める重要な要素となっている．

また，このようなサービスにおいてはタグと呼ばれる文字列をコンテンツに付与することも盛んである．タグにより人力によるコンテンツの要約が行われ，コンテンツがテキスト検索により発見される機会を増加させる作用がある．タグは近年ではそれぞれのサービスに独立して実装されてきているが，任意の URL に対しタグを付与することで検索に生かすサービスとして，livedoor クリップ[7] などのソーシャルブックマークが挙げられる．

タグのように，一般ユーザの知恵と労力を用いて計算機システムのみでは実現が困難な問題を解決するアプローチは「集合知」などと言われることがあるが，集合知をうまく活用するためには，一般ユーザが知恵や労力を貢献しやすい環境を準備する必要がある．たとえば上記に挙げたようなタグの実装法では，コンテンツ提供者もしくは閲覧者が意識的にタグをつける作業を想定することが通常である．しかし一般にタグをつけるという作業は労力を伴うため，タグを集めるために特別な工夫を行った事例もある．Google Image Labeler[5] は，

© 2010 Information Processing Society of Japan

画像データにタグを付与する作業にゲームの要素を加えたものである．ユーザはゲームをしているだけにすぎないが，実は背後でそれが Google の画像検索の性能向上に寄与している．一方，音声・映像等の配信手段である Podcast のコンテンツを収集，音声認識し，テキストによる検索を可能にした PodCastle[6] では，音声認識の誤りを訂正する作業を一般ユーザに解放し，音声認識システムの精度の向上を図っている．訂正作業においてユーザは「間違い探しゲーム」のような感覚で臨んでおり，Google Image Labeler と同様エンターテイメント性が作業への貢献を促進していると考えられる．また，自分の好きな有名人の音声が間違って認識されていることを放置することができない，という心理も作業への貢献を促していると考えられる．

エンターテイメント性以外の要素を持ち込んだ事例としては，人力検索はてな[19] のように，金銭もしくは金銭とほぼ同等の価値を持つサービス内貨幣により，情報提供が売買されることで一般ユーザの貢献が促進される例もある．

我々はこのように一般ユーザが様々な形で参画することで成立している近年のインターネット上におけるサービスについて，個々のコンテンツやユーザの挙動，ひいてはその総体であるコミュニティや集合知の性質などの分析に適用可能な統計手法を提案する．ソーシャルネットワークを対象としてタグ付け行動を分析したり，モデル化した先行研究事例として[3] などが挙げられるが，そのような研究に対し我々は記述統計学的な新しい分析ツールを追加提供することが可能である．また，本論文で扱う分析事例では従来とは別の切り口からコンテンツを価値づけする指標を定義するが，そのような先行研究として，入江ら[18] によるもの，大石ら[17] によるものなどが挙げられる．入江ら[18] の研究は，動画共有サイトの動画のランキング方法として「編集がされた程度」という観点を加えたものである．また大石ら[17] の研究は，ソーシャルブックマークのタグを「新規性」という観点で価値づけし，サイトの重要度を算出するものである．我々の提案手法は動画内容やタグの文字列内容といった個々のサービス特有の情報に依存せず，幅広くこれらのサービスに共通するデータ構造のみから結果を出せる汎用性を備えている．そのため，そのような既存手法と競合せず，むしろ併用することによる相乗効果が期待される．

3. 統計パッケージ tailstat

本節では，前節で扱ったような統計分析を行うソフトウェアパッケージ tailstat の実装について詳細に述べる．tailstat は，C#を用いて開発されたソフトウェアである．コマンドラインから動作するアプリケーションと，GUI アプリケーションからなるが，基本的な機能は同等である．

tailstat における基本的な作業は，確率分布を読み込みそれをたたみこみ計算することである．読み込む確率分布は，csv ファイルとして以下のフォーマットを持つ：

$$\begin{array}{l} category_1, probability_1 \\ category_2, probability_2 \\ \qquad\vdots \\ category_k, probability_k \end{array} \tag{1}$$

ここで $category_i$ は確率変数の i 番目の値，$probability_i$ はそれに対応する確率を表す．これらのペアは $category_i$ を key とし，$probability_i$ を value とするハッシュテーブルに登録される．

$$value\ for\ key(category_i)\ =\ probability_i \tag{2}$$

読み込まれた確率分布のたたみこみは，新しいハッシュテーブルを用意し，すべての value をゼロに初期化したのち，すべての $i, j(1 \le i \le k, 1 \le j \le k)$ の組について，

$$value\ for\ key(category_i + category_j)\ +=\ probability_i \times probability_j \tag{3}$$

を求めることで実現される．

本手法により n 回たたみこむ場合は，四則演算の回数および必要なメモリは最大で k^{n+1} のオーダで増大するため，k および n が大きい場合は計算が困難となる．しかし，よほど分布が左右に偏っていない限り，n がそれほど大きくなる前に中心極限定理が成立し，以降は正規分布による近似が有効になる．また，$category_i$ の各値が a を定数として $0, a, 2a, 3a, 4a, \ldots, (k-1)a$ のように等間隔である場合は現実問題として多くみられるが，その場合たたみこみ後に得られる key の総数は $(k-1)(n+1)+1$ 程度であり，メモリが大幅に消費されることはない．さらに，そのような場合には FFT（高速フーリエ変換）を用いたたたみこみ計算の高速化が可能である．

たたみこみ後に必要な機能として，tailstat には

- 任意の階級幅による度数分布表の作成
- 歪度，尖度に基づく正規分布らしさの判定（中心極限定理の成立度合いの確認）
- 分布における右側確率，左側確率の導出，および中心極限定理が成立するとみなしたときの正規分布による同種確率の導出
- 確率変数の線形変換

などが実装されているが，個別の機能の説明については自明なので割愛する．

© 2010 Information Processing Society of Japan

4. 分 析 例

本章では tailstat を用いて分析を行った事例を示し，その有効性を議論する．

4.1 livedoor クリップの分析

実際的で応用範囲の広い事例として，livedoor クリップの分析例を述べる．livedoor クリップは，ソーシャルブックマークサービスである．これは，通常各人がブラウザのブックマークとして記録しているサイトを公開することにより価値あるサイトをみんなで発掘，共有しようという発想のサービスである．ここで重要になるのがタグの概念で，先述のようにユーザがサイトをブックマークとして登録する際，そのサイトを象徴する任意の文字列をキーワードとして任意登録したものである．タグがあることにより，タグによる検索や，あるサイトに付与されているタグから連想される関連サイトの推薦などが行われるようになる．タグが付与されることはサービス運営者側にとってみるとサービス全体の質の向上につながり，また各サイトの管理者にとってみてもサイトの注目度が高まるため双方に利益がある．

ここで，この構造はインターネット上の様々なサービスで類似点が多い点に注目する．タグ機能はソーシャルブックマークサービスだけではなく，写真共有サイト flickr やニコニコ動画などの投稿動画サイトでも存在する．また，料理レシピ共有サイトであるクックパッドの「つくレポ」のように，あるコンテンツに対する返信コメントの数でそのコンテンツの質が評価される場合，コメント数はタグ数と類似の価値を持つ．より直接的には，無数にあるインターネットショッピングサイトにおける各来客者（ページ閲覧者）に対する商品購入額が，そのコンテンツの価値を決める有効な指標であることは明らかである．

これらを一般化すると，このようなサイトにおいては，ユーザがあるコンテンツ（web ページ）を訪問し，そのコンテンツに反応し，なんらかの対価を支払うというモデルが適用できる．対価はタグであったり，返信コメントであったり，商品の購入金額であったりする．そのコンテンツの価値は，訪問者により支払われた対価の総計で判断される．

このモデルによる自明な分析の一つは経営学におけるごく初歩的な理論により，(客単価) = (総売上) ÷ (訪問者数) を算出することである．訪問者数が一定の時，客単価の高い店および商品は優良であると判断され，それに基づく販売戦略が立てられることであろう．しかし，このような戦略が有効なのはそもそも訪問者数が十分多い場合，すなわち中心極限定理や大数の法則が成立する場合である点に注意が必要である．ロングテールなインターネット商店街（コンテンツ群）は基本的に大部分の商店が閑散としているのであり，稀に訪れた客が客単価通りの額の買い物をすることもむしろ珍しいことが多い．

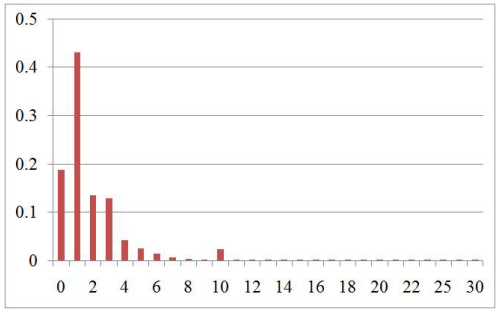

図 1 ブックマークあたりの登録タグ数の頻度分布
Fig. 1 Frequency distribution of number of registered tags for each bookmark.

我々は tailstat を用いて，そのような領域において意味のある分析を試みる．livedoor クリップを例に話を進めるが，基本的にそれは他の様々な「来訪・対価支払い」のモデルで表されるサービスに読み替えても一般性を失わない点を強調する．

4.1.1 手 法

livedoor クリップ 2008 年 12 月版データセット[8] は，1572742 件のブックマークについて（ユーザ ID，対象 url，クリップした時刻，タグ）を網羅したものである．このデータセットには，データセット収録 3 か月前から存在し，3 人以上のユーザに登録されているブックマークのみが収録されている．本分析ではブックマーク登録者を「来訪者」，登録タグ数を「対価」と考えて分析を行う．具体的には,「ブックマーク登録者数のわりに登録タグ数が多い（少ない）サイトはどれか」を探す分析である．なお，より発展的にはサイト閲覧者のうちどのくらいの割合でブックマーク登録がされたのか（これも一種の対価支払いとみなせる）が分かればさらに詳細な議論ができるが，これは閲覧者数がデータセットに収録されていないため割愛する．

まず図 1 に，ブックマークごとにどのくらいの数のタグが登録されているのかを表す頻度分布を示す．これは，それぞれのブックマーク登録が独立試行であると仮定したとき,「来訪者 1 人当たりの対価支払い額の確率分布」に相当し，その期待値 1.818 がゆくゆくは（大数の法則により）客単価になる．左右非対称であり，タグ数 1 をピークに減衰していく形状をしている．また，タグ数 10 のカテゴリが特別多いが，これは livedoor クリップが許すタグの最大表示数が 10 であり，ユーザの間でタグを 10 個つける文化が発達しているから

© 2010 Information Processing Society of Japan

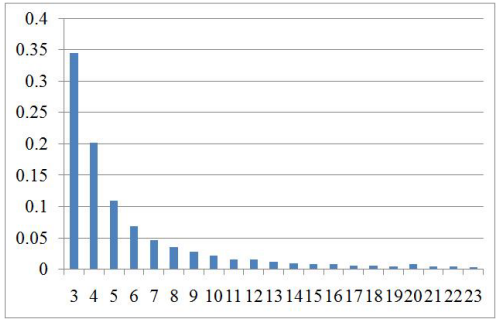

図 2 ブックマークあたりの登録者数の頻度分布（実際は 960 人まで裾が伸びている）
Fig. 2 Frequency distribution of bookmark registrants. Note that this visualization cut off at 23 people but actually continues to 960 people.

であると考えられる[⋆1]．以後, この図 1 の頻度分布を g と表すことにする．次に図 2 を参照されたい．これは個々のサイトにブックマークを付ける登録者数の頻度分布のうち，主たるもの（登録者数 23 人以下）を示したものであり,「来訪者数の頻度分布」に相当する．2 人以下のデータについてはそもそもデータセットに収録されていないためここでは扱わない．分布の期待値は 7.218，標準偏差は 10.59 であり，まさに long-tailed（裾の長い）分布である．ここから，多くのサイトにとってブックマーク登録をしてくれるユーザ数は通常それほど多くなく，大数の法則や中心極限定理を適用するには程遠いことが明らかである．

ここで tailstat を用いて，図 2 の分布に現れた 3 人から 960 人までのブックマーク登録者数 $n(n = 3, 4, \ldots, 960)$ に対応して，図 1 の分布を $n-1$ 回たたみこんだ確率分布 g^{*n} を求める．これは「n 回ブックマークが登録された場合に，付与されるタグの数 T の確率分布」を表す．図 3 および図 4 に，それぞれ g^{*20} および g^{*200} のヒストグラムを示す．g^{*20} ではまだ分布は右に裾が長く，一方 g^{*200} ではほぼ正規分布となっていることが確認できる．次にデータセット中のあるサイトに着目し，ブックマーク登録者数 n と付与された総タグ数 t を観測する．ブックマーク登録者数 n に対応する確率分布 g^{*n} を呼び出し，総タグ数が t 以上となる確率 $p_t = P_{g^{*n}}(T \geq t_k)$ を求める．また，比較のために同じ条件で図 1 の分布を用いて中心極限定理が成立すると仮定した場合の正規分布 $N(n_k\mu, n_k\sigma^2), \mu = 1.818, \sigma = 2.008$

⋆1 実際にはタグの数が 10 を超えるブックマークも 2764 個存在したが，この理由は不明である．10 以上のタグも受け付けた時代があったか，そのような非公開手法が存在したのではないかと推測される．

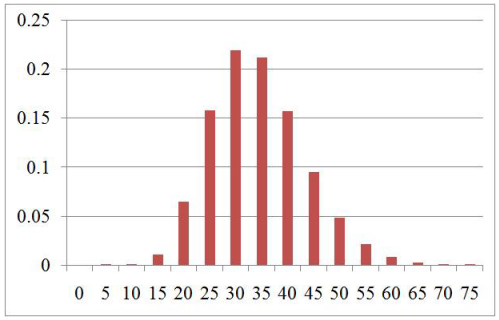

図 3 g^{*20} のヒストグラム (歪度 = 0.5379, 尖度 = 0.3668)
Fig. 3 Histogram for g^{*20} (skewness = 0.5379, kurtosis = 0.3668).

からの確率 $p_z = P_{\mathrm{norm}}(T \geq t)$ も求めておく．以上，p_t および p_z を求める作業を，データセット中のすべてのサイトに対して行う．p_t や p_z は小さければ小さいほど，そのような数のタグが得られるのは稀である，すなわち「希少価値がある」あるいは「異常事態である」ということができる．

なお，総タグ数の異常度の評価基準として尤度 $L = \prod_{i=1}^{10} p_j^{k_j}$ を用いるという考え方もありうるだろう．ただし p_j は個人が 1 つのサイトに対して j 個のタグをつける確率であり，k_j は n 人のブックマーク付与者のうち j 個のタグをつけた人の人数である．L そのものに対して適当な棄却域を設定する多項検定はこの考え方に基づいた評価法の最たるものである．また L の変形として $L' = \prod_{i=1}^{10}(p_j + \cdots + p_{10})^{k_j}$ なども手軽に計算できる指標の候補になりうるかもしれない．しかし総タグ数 $\sum_{j=1}^{10} jk_j$ の異常度を検知するのに L や L' のしきい値をどのように設けるかは恣意性がある．汎用ツールとしては $\sum_{j=1}^{10} jk_j$ の値そのものの確率を評価する tailstat のやり方が一番標準的であると言わざるを得ない．また積確率を用いた指標には実用上の問題点が 2 つある．1 つはその値自体に自然な実用上の解釈が無い事である．尤度が 0.01 違うことにどのような意味があるのかは自然には解釈できない．また 2 つ目の問題は来客数についての標準化ができていないことである．3 人の訪問者を持つサイトにタグを 27 個ついているという事実と 4 人の訪問者を持つサイトに 37 個のタグがついているという事実の異常度の違いが各々の尤度の比較からは明らかにならない．しかし, tailstat の上側確率では両者の異常度は比較可能である．

4.1.2 tailstat v.s. 客単価

ここから，まず p_t と客単価（総タグ数 $T \div$ ブックマーク登録者数 n）の関係を分析する．

© 2010 Information Processing Society of Japan

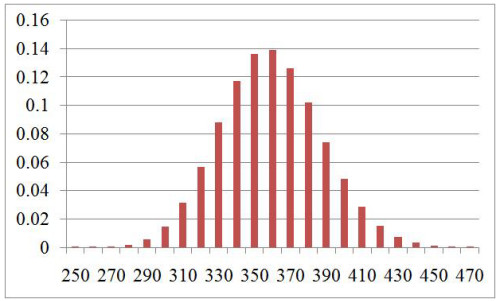

図 4 g^{*200} のヒストグラム (歪度 = 0.1705, 尖度 = 0.03244)
Fig. 4 Histogram for g^{*200} (skewness = 0.1705, kurtosis = 0.03244).

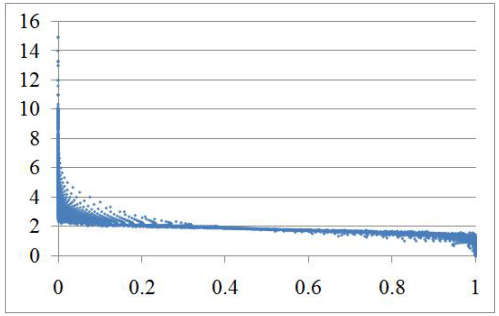

図 5 p_t と客単価の比較
Fig. 5 Average customer spend vs. p_t.

これらは両方とも「ブックマーク登録者数のわりに登録タグ数が多い（少ない）サイト」の指標として活用可能である．図 5 は横軸に p_t をとり，縦軸に客単価をとったものである．おおむね負の相関が読み取れるが，p_t がゼロに近い場合，および 1 に近い場合に，客単価には大きなばらつきが現れることが読み取れる．これは，もしも客単価上位（もしくは下位）のサイトを検索した場合，より希少価値のあるサイトを見落としてしまう可能性があることを示している．

図 6 はそれを定量的に示したものである．図 5 から $p_t < 0.1$ のものだけを抽出し，これを「希少価値のあるサイトの全体」と定義する．いま，正数パラメータ a を 0 から少しずつ大きくしていった際に，客単価 $> a$ をクエリとしてどの程度希少価値のあるサイトを検索できるかを，横軸に a，縦軸に再現率（Recall）および精度（Precision）をとって示したのが図 6 である．たとえば a が 3 の場合に再現率と精度はいずれも 8 割弱程度となり，希少価値のあるサイト全体の 8 割弱程度を網羅できるが，同時に希少価値のないサイトも 2 割強程度含んでしまう．これは，客単価という価値基準によるサイトの希少価値判断に限界があることを示している．

また，p_t は確率として得られているため，客単価よりも豊富な情報量を持っている．客単価は,「上位何%」や「上位何件」といったランキング形式での選別が行われることが多い．これは,「何点以上が合格ライン」のような具体的な意味を持つ閾値を決定することが難しいことに由来すると考えられる．一方で p_t は「そのような事象が起こりうる確率」という具体的な意味を持っているため，順序尺度的な選別に加えて「1%以下の確率でしか得られない希少なサイト」のような具体的な観点での選別が可能である．

以上の結果から， tailstat により求めた p_t により，客単価では抽出できないような観点からサイトの価値を評価することが可能であることが示された．

4.1.3 tailstat v.s. 中心極限定理

次に，tailstat により得られた結果 p_t と，中心極限定理が成立すると仮定し計算された p_z の比較を行う．p_t と p_z についてそれぞれ数値計算のノイズを除去するために 0.0001 以上のもの（その数は 3880）だけ選別した．その比の常用対数 $log_{10}(p_t/p_z)$ を縦軸にとり，横軸にサンプル数であるブックマーク登録者数 n をとり散布図をプロットしたものが図 7 である．この図から，たしかにブックマーク登録者数が多くなると（たとえば 200 人以上），中心極限定理が支配的になり p_t と p_z の差はなくなっている（ゼロに収束する）が，そのようなデータは全体の 0.6%にとどまる．逆に大部分を占めるブックマーク登録者数が少ない領域においては，p_t と p_z の差は大きくなり，一桁以上異なる（縦軸の値が 1 以上となる）場合も多く見られる．この結果から，中心極限定理を不用心に用いた分析は適切ではなく，tailstat を用いた分析が有効であることが示された．

4.1.4 考　　察

本節では分析結果を踏まえて，得られた知見を一般化するうえで留意すべき事項を考察する．今回我々は「来訪・対価支払い」モデルの適用できる例として livedoor クリップのタグ登録に着目したが，この事例には以下の特徴がある．

(1) ある来訪者の対価支払い額は他の来訪者の影響をあまり受けない．
(2) ある来訪者の支払う対価は，他の来訪者が支払う対価と重複することが多く，それが許される．

© 2010 Information Processing Society of Japan

Vol.2010-HCI-138 No.3
2010/5/14

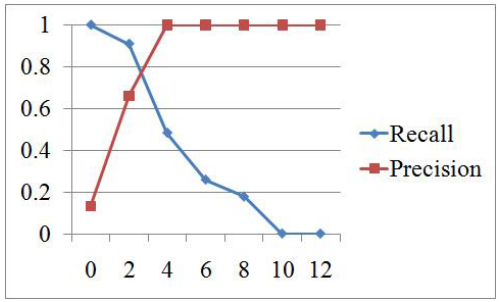

図 6 客単価をクエリとした場合の再現率と精度
Fig. 6 Precision and recall for each average customer spend as a query.

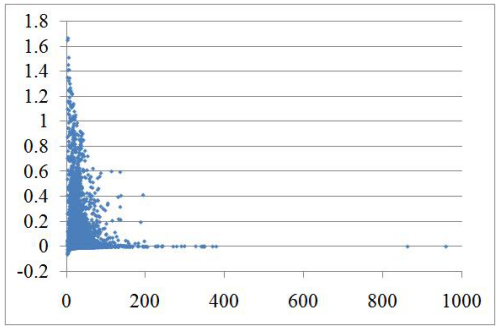

図 7 ブックマーク登録ユーザ数と $log_{10}(p_t/p_z)$
Fig. 7 Number of bookmark registrants vs. $log_{10}(p_t/p_z)$.

(3) 来訪者あたりの対価支払い額分布は，正規分布もしくは**再生性**のある既知の分布にならない．

(4) 多くの対価が支払われたコンテンツほど新たな来訪を誘発する仕組みがある．

それぞれについて，詳しく述べる．

4.1.4.1 独立試行か否か

まず第 1 の特徴は，具体的にはタグを登録する画面において，他のユーザがそのサイトに登録したタグは見えないということである．これにより，各ブックマーク登録におけるタグ登録が独立試行であるとみなしやすくなる[⋆1]．一方で PodCastle におけるある音声コンテンツの音声認識誤り訂正作業の例では，訂正作業が進めば進むほど，新たな認識誤りの発見・訂正は難しくなり，対価支払い行為が独立試行ではなくなる．同様にニコニコ動画のタグのように,「いま付与されているタグをどのように追加・削除していくか」というインタフェースであれば，独立試行の仮定は難しくなる．

独立試行とみなせることは，tailstat および主要な推測統計学を適用するうえでの重要な仮定である．よって運用にあたっては，扱う事象をよく精査する必要がある．場合によっては，独立と仮定して導いた結果から「このような事態は通常起こり得ないので独立の仮定に問題があった」と結論付け，対象となる現象の分析をより深める手法も有益であると考えられる．

4.1.4.2 対価の重複

第 2 の特徴は，具体的にはあるユーザの登録したタグ文字列が他のユーザが登録したタグ文字列と重複する場合が多く，それが許されているということである．これは第 1 の特徴と関連している．ソーシャルブックマークサービスとしてはこれは歓迎すべきことである．なぜならば，同じサイトに対し同じタグ文字列を付与するユーザ間には関連性を見出すことができ，新たな情報推薦などのサービスを展開できる可能性を秘めているからである．また，重複して登録された頻度の高いタグに重要度を増すような計算手法・可視化手法（タグクラウドなど）も可能になる．

同様な特徴は一般的なネットショッピングの事例にも現れる．すなわち，それぞれの来訪者がいくら金銭を支払おうとも，お互いに干渉しあうことはない．

一方で PodCastle においては，新たな来訪者が目にするのはこれまでの来訪者が訂正を行った結果であり，重複する訂正は対価支払いとみなされない．異なるユーザに同じ音声認識誤りを訂正させることはインタフェースデザインとしては可能ではあるが，そうしない方がサービスの質の向上につながるという判断を行っている事例である．これは,「多数のユーザの多様な意見の抽出度合い」ではなく「多数のユーザによる 1 つのコンテンツの精錬度合い」を対価として求めたものであると考えられる．

このようにサービスが対価として来訪者に何を求めているかにより，先に述べた試行の独立性，および分析したい対価の統計量の定義が異なってくる．tailstat は現状では，互いに独立であり重複が許される対価支払い量の分析を想定しており，それ以外の事例への対応に

⋆1 他者のタグを手掛かりにサイトを発見することも多いので，厳密には独立とは言えない．

© 2010 Information Processing Society of Japan

ついては今後の課題である．

4.1.4.3 母集団の分布

第 3 の特徴は，具体的にはタグの登録数分布（図 1）は正規分布ではなく，そしてたたみこみにより元の分布が再び現れる（再生性のある）ようなよく知られた分布ではないということである．これにより，定義通りのたたみこみ計算を多数実行する必要があるため，tailstat の意義が顕著化される．一方，web コンテンツに来訪者が点数をつけることで付加価値をつけるレーティングの仕組みを対価として考える場合は，母集団は正規分布と仮定できる場合があるため，その際は tailstat は不要である．ただし，よく見られる 5 段階評価のようなレーティングでは，天井効果や床効果の問題が発生する可能性もあるため，運用には慎重を期すべきである．

さらに，本論文では母集団が離散分布となる場合に限定して議論を進めている．母集団が連続分布である場合の対応は今後の課題である．

4.1.4.4 分析結果のサービスへのフィードバック

第 4 の特徴は，具体的には多くのタグが付与されたサイトほど他のユーザに発見され，新たにブックマーク登録されやすい仕組みがあるということである．事実，タグが多くつけられるほどタグ検索にヒットする可能性が高まるので，他のユーザに発見されやすくなっている．これは商業一般に言えることで，よく売れる商品ほどよく見える場所に陳列され，さらに売れていくものであろう．従来は売上ランキング，来訪者数ランキングや客単価ランキングといった指標で商品（コンテンツ）が陳列される戦略がとられていたが，tailstat を用いて今回算出した「希少価値」という指標による商品陳列が，その後の売れ行きや希少価値指標の変化にどのような影響を与えていくかを時系列で追うことは技術的には容易であり，十分に興味深い．これは今後の研究課題である．

5. まとめと今後の展望

本論文では,「ロングテール」「Web 2.0」などの言葉で表現されるような，多様な利用者の参画が可能になったことで現れてきた新たなサービスとそこにしばしばみられるデータ構造について，母集団分布が既知であり，また中心極限定理が適用できないような小標本の統計分析が重要性を持ってきていることを示した．そしてそのような問題領域における，既知の母集団分布のたたみこみを基本とした標本和・標本平均の統計手法の再構築と，それを実現する統計分析パッケージ tailstat の開発を行った．また，livedoor クリップの事例を用いて tailstat による分析が有効であることを示し，一般的に「来訪・対価支払い」のモデルで表現される他の事例へ応用するための留意事項について考察を行った．

今後は，標本和・標本平均以外の分析手法の検討，また tailstat に拡張を行い，計算の高速化，連続分布向けの機能等を充実させていく予定である．アプリケーションの公開を通じて分析事例の蓄積も行っていきたい．

謝辞 この研究の遂行にあたって支援を頂いた日本私立学校振興・共済事業団，Zeenon，およびデータ提供を受けた livedoor に感謝します．いくつかの有益なアドバイスを頂いたことについて，東京大学数理科学研究科の Ralph Willox 氏にも感謝します．

参 考 文 献

1) amazon.com: http://www.amazon.co.jp/.
2) flickr: http://www.flickr.com/.
3) Golder, S. and Huberman, B.A.: Usage Patterns of Collaborative Tagging Systems, *Journal of Information Science*, Vol.32, No.2, pp.198–208 (2006).
4) Google Analytics: http://www.google.com/analytics.
5) Google Image Labeler: http://images.google.com/imagelabeler/.
6) Goto, M., Ogata, J. and Eto, K.: PodCastle: A Web 2.0 Approach to Speech Recognition Research, *Proceedings of Interspeech 2007*, pp.2397–2400 (2007).
7) livedoor Clip: http://clip.livedoor.com/.
8) livedoor labs EDGE Datasets: http://labs.edge.jp/datasets/.
9) Public Data Sets on AWS: http://aws.amazon.com/publicdatasets/.
10) 梅田望夫：ウェブ進化論，筑摩書房 (2006).
11) vizoo: http://www.visualzoo.com/.
12) Wikipedia: http://wikipedia.org/.
13) Yahoo!オークション：http://auctions.yahoo.co.jp/.
14) YouTube: http://www.youtube.com/.
15) クックパッド：http://cookpad.com/.
16) ニコニコ動画：http://www.nicovideo.jp/.
17) 大石剛司，亀田尭宙，深見嘉明，大向一輝，武田英明：ソーシャルブックマークにおけるユーザのタグ付与行動分析に基づくコンテンツ分類，情報処理学会全国大会予稿集，Vol.70, No.5 (2008).
18) 入江　豪，日高浩太，佐藤　隆，小島　明，相澤清晴：動画共有サイトにおける映像検索のための編集度順序によるランキング法の提案と評価，インタラクション 2009 予稿集，pp.87–94 (2009).
19) 人力検索はてな：http://q.hatena.ne.jp/.
20) 総務省統計局：http://www.stat.go.jp/.
21) 楽　天：http://www.rakuten.co.jp/.

© 2010 Information Processing Society of Japan