情報処理学会論文誌 Vol. 52 No. 2 477–487 (Feb. 2011)

ロングテール時代のための中心極限定理によらない統計分析手法

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手法が現代のインターネット情報社会では重要性を持っていることを示す．具体的には，適用範囲の広い「来訪・対価支払い」型ウェブサービスのモデルを定義し，サービス上の各コンテンツの価値や異常性を表現する「希少度」を計算する．ソーシャルブックマークサービスである livedoor クリップを分析事例として，手法の汎用性と有効性を示す．

Non-central Limit Theorem Statistical Analysis for the "Long-tailed" Internet Society

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

This article presents a statistical analysis method, 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. We define "visitor/compensation" type web service model and "rarity" as a metric of value or abnormity for each content of the web services. A large-scale case study on livedoor clip, a social bookmark service, demonstrates the effectiveness and versatility of the proposed method.

†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
Kanagawa Institute of Technology

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 3 分の 1 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[15)]．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[19)] や Yahoo!オークション[18)] 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まった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[15)]．

このような時代背景にあって，我々が社会分析，経営分析など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ともな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[21]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[11]．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分析手法を研究している．本論文ではその第 1 段階として，ロングテール的（Web2.0 的）なウェブサービスに広く見られるコンテンツやユーザの挙動をモデル化し，これまであまり注目されてこなかった統計手法による分析を適用する．分析事例を示すことで，手法の汎用性，有効性を示す．

我々が用いる統計分析手法は，母集団が既知である場合の標本和および標本平均の分布を取り扱うものである．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．

本論文で示すウェブサービス分析手法を転用することにより，類似のウェブサービスにおいて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，まず関連研究をあげ，次に我々が分析対象とするウェブサービスのモデルの定義と分析手法の定式化を行い，実際の分析事例を示すことでその一般性と有効性を議論する．

2. 関 連 研 究

一般に統計分析の理論において，既知の母集団およびそこから得られた標本の特性を調べる分野を記述統計学と呼び，未知の母集団を標本から推測する分野を推測統計学と呼ぶ．本研究は Web 上で公開されているデータセットなど，既知の母集団から得られた標本が作る統計量の挙動，特に標本和の挙動を利用した分析手法を提案するため記述統計学に分類される．

記述統計学では個々の標本が従う母集団分布が所与である．そして各種統計量は標本の変数変換であるから，その確率分布は原理的には初等的な確率論で計算できる．しかし標本数が大きくなるとこの計算は計算量の観点から実用上実行不能となる．そのような場合には標本数が無限大のときに従う漸近分布を求めておいて，それで近似するということが行われる．特に重要な例は標本和および標本平均の分布に対する，中心極限定理による正規近似である．

1 章の (1)，(2)，(3) の性質を持つような対象においては大標本から求めた統計量によるマクロな分析が中心となるため，このような近似は非常に有効に機能した．そのため 20 世紀以降多くの統計量の漸近分布が求められてきた．数学的な興味がそれを後押ししたということもあるだろう．

しかしながら本研究で扱うような小規模標本に対しては，統計量の確率分布を用いた分析手法は統一的にはあまり探求されてこなかった．既知の母集団から得られた小標本は，各現場ごとにノンパラメトリックテストなどの推測統計の理論を，アドホックに応用しながら分析されてきた．1 章でも述べたことであるがロングテールに代表されるような小規模標本の記述統計が分野横断的な普遍性を持って重要であると認識されるようになったのはごく最近のことであるから，これは当然のことといえる．

なおジップ則やロングテールに代表されるパレート型の母集団分布が形成される原因を数理モデルを用いて説明する研究は 1980 年代から始まっている．本研究の問題意識はこのような数理記述統計学とはまた違ったところにあるということには注意していただきたい．

次に，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，関連する試みをあげる．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うえで重要な指標と見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[6] 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[25]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[16]や，Public Data Sets on Amazon Web Services[13]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うえ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あげる．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[26]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」と呼ばれることがある），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[18]などのネットオークションがあげれらるが，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[19]やニコニコ動画[21]，写真共有サイト flickr[3]，また料理レシピ共有サービスであるクックパッド[20]，百科事典 Wikipedia[17]などがあげられる．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さかんである．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[10]などのソーシャルブックマークがあげられる．

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」などとい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．たとえば上記にあ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ともなうため，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[7]は，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・映像などの配信手段である 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[8]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[24]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コミュニティの成長や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文献 2)，4)，5)，9)，14) などがあげられるが，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手法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[23]によるもの，大石ら[22]によるものなどがあげられる．入江ら[23]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[22]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. 「来訪・対価支払い」型ウェブサービスとコンテンツの希少度

本章では，我々が分析対象とするウェブサービスについてモデル化を行い，コンテンツの

価値（希少さもしくは異常度）を表す希少度という尺度を定義する．

3.1 「来訪・対価支払い」モデル

今日のインターネット上には，商品や観賞用コンテンツなどが膨大に集積され公開されているサービスが多数見られる．その 1 例として，ソーシャルブックマークサービスがあげられる．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，これらのサービスにおいては，ユーザがあるコンテンツを訪問・閲覧し，そのコンテンツに反応し，何らかの対価を支払うという行為が基本単位として抽出できる．対価は商品の購入金額であったり，そのコンテンツの価値を賞賛・批判・要約するためのテキスト（コメントやタグ）であったりする．そのコンテンツの価値（利益性，話題性など）を，訪問者により支払われた対価の総計で判断することは自然な行為であるといえよう．

不特定多数のユーザがこの基本単位を繰り返すことで成立するウェブサービスを，「来訪・対価支払い」型ウェブサービスと定義する．このモデルが適用できるサービスにおいては，インターネット上にサーバを配置し一元的に運用しているのであれば，サーバ上に蓄積される履歴から（コンテンツ ID，対価量）を一組とするデータを，不特定多数のユーザの来訪の数だけ取得可能である．よって，このようなデータ構造を持つサービスを「来訪・対価支払い」型と定義しても同値である．なお，対価量にはゼロも許される．

このデータ構造から，以下の代表的な 2 つの分布が得られる．1 つは，「来訪」の分布 f である．これは，それぞれのコンテンツがどれほどの数のユーザの来訪を受けるかを表した分布である．「ロングテール」なウェブサービスについては，ごく一部の有名なコンテンツが多くのユーザの来訪を受け，それ以外の大部分を占める無名なコンテンツはわずかな来訪者数しか得られないため，この分布はジップ則に従う右に裾野の広い（long-tailed）分布となる．

もう 1 つは，「対価支払い」の分布 g である．これはあるコンテンツを訪れたユーザが支払う対価の量を表した確率分布である．対価の尺度や単位は対象となるサービスによって異なる．ショッピングサイトであれば円が単位となり，タグやコメントなどのテキストが対価となる場合はその絶対数，文字数などが単位となりうる．分布の形状も多用であり，特にユーザが「まず対価を支払うかどうかを判断し，その後に支払額を決定する」というような場合は，「対価を支払う/支払わない」の 2 値をとる確率変数と，「対価を支払う場合の支払額」を表す確率変数をかけ合わせた確率変数が従う分布となる．

3.2 希　少　度

3.2.1 定　　義

「来訪・対価支払い」型のサービスにおいて，自明な分析の 1 つは経営学におけるごく初歩的な理論により，各コンテンツについて (客単価) = (対価合計) ÷ (来訪者数) を算出し比較することである．来訪者数が一定の時，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来訪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．分布 f の形状に見られるように，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どおりの額の買い物をすることもむしろ珍しいことが多い．

そこで我々は，すべてのコンテンツで計算可能かつ互いに比較可能な新たな価値尺度として「希少度」を導入する．これは，各ユーザの行動が独立試行であり，それぞれの来訪時の対価支払い量が確率分布 g に従う確率変数であるとしたとき，「あるコンテンツが得た対価支払い量以上の対価支払いを得る確率」で定義される．

たとえばあるコンテンツが n 人の来訪を受け，総対価支払い量 t を得た場合の希少度 p_t は，以下のように計算する．

(1)　確率分布 g^{*n} を求める．ここで，記法 h^{*i} は確率分布 h を $i-1$ 回たたみこんだものと定義する．g^{*n} は「n 回の訪問があった場合に，支払われる対価量の合計 T の確率

分布」を表す．

(2)　この分布において総対価支払い量が t 以上となる確率 $p_t = P_{g^{*n}}(T \geq t)$ を求める．これがこのコンテンツの希少度となる．

p_t は小さければ小さいほど，そのような数の対価量が得られるのは稀である，ということができる．

なお，n が大きい値の場合，希少度は中心極限定理を用いても近似計算可能である．分布 g の期待値を μ，分散を σ^2 とすると，T は正規分布 $N(n\mu, n\sigma^2)$ に近似的に従う．この分布において総対価支払い量が t 以上となる確率 $p_z = P_{\mathrm{norm}}(T \geq t)$ は，希少度の近似値として活用できる．

このように希少度は単に標本和分布におけるある上側確率であり，その概念自体は新しくない．付録 A.1 に，たたみこみと正規近似による希少度の計算を行う分析補助ツール tailstat について具体的なアルゴリズムとともに示す．本論文で我々が主張するのは，母集団既知という仮定が成立しやすい今日の「来訪・対価支払い」型ウェブサービスにおいて，希少度がコンテンツ分析に有効な指標である可能性があるという点である．

3.2.2　解　　釈

希少度の解釈は分析対象のサービスによって異なる．ショッピングサイトの購入金額などを対価として考えている場合は希少度はいわゆる「希少価値」として肯定的に取り扱うことができるが，観賞用コンテンツに対するコメントテキストの量を対価として考えている場合などには，悪意や故意により有害なテキストを大量に登録する行為も想定される．その際は希少度は「通常起こりえない異常事態」の指標となる．これらの区別を行うにはコンテンツや対価の実際の内容を吟味することが必要であり，本論文のスコープ外であるが，自動化には自然言語処理，およびコンテンツ自動理解技術との併用が効果的であると考えられる．

具体的には，テキスト量を対価とする場合は個々のテキストが特定の「ブラックリスト」に含まれないかどうか，もしくはより高度にはスパムメールの判定と同様の原理で悪意や故意を判断する手法が有望であろう．さらに，文献 22)，23) のようにコンテンツそのものを入力として何らかの「価値」あるいは「悪意・故意の度合い」を計算する指標と希少度を合議，あるいは適切な重み付けにより運用することで，より確実な価値判断の自動化が可能となるだろう．

3.2.3　ノンパラメトリック手法との関係

離散値をとる小標本の希少さを確率を用いて評価する場合，ノンパラメトリックな分析法の 1 つである Exact Test（多項検定）を応用することも考えられる．Exact Test を応用する場合，本質的には同時確率関数 $L = \prod_{j=1}^{m} p_j^{k_j}$ を用いて P 値（「上側」確率）を計算し，希少さを判定することになる．ただし p_j はユーザが 1 つのコンテンツに対して j 単位の対価を支払う確率であり，k_j は j 単位の対価を支払ったユーザ数である．また，m は対価の最大値である．ここで問題となるのは，標本 $(k_1, k_2, \cdots, k_m)$ は m 次元の空間に値をとっているので，どの領域を標本の「上側」と見なすのかは分析者がアドホックに設定しなければならないことである．この閾値の設定に恣意性があるという問題は，Exact Test の運用においてはつねにつきまとう．今回，「来訪・対価支払い」型ウェブサービスにおいて希少さを判定したいのは総対価量 $\sum_{j=1}^{m} jk_j$ である．したがって，m 次元空間の中で，総対価量が標本のものよりも大きい領域での L の和を P 値と定義するのが自然であろう．この場合，P 値は本論文で定義した希少度と完全に一致する．換言すれば，自然な用い方をする限り希少度は Exact Test の考え方を含んでいる．

また P 値を用いないまでも，同時確率 L や，その変形である $L' = \prod_{j=1}^{m}(p_1 + \cdots + p_j)^{k_j}$ など尤度に相当するものを直接希少さを表す指標として利用することも考えられるだろう．しかしこのような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ないことである．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 つ目の問題は来訪者数についての標準化ができていないことである．3 人の来訪者を持つコンテンツに対価が 27 単位支払われているという事実と 4 人の来訪者を持つコンテンツに対価が 37 単位支払われているという事実の希少さ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，我々の希少度の定義では両者は比較可能である．

4. 分　析　例

本章では「来訪・対価支払い」型ウェブサービスとしてソーシャルブックマークサービスである livedoor クリップに注目し，希少度に基づいた分析を行い，その有効性を議論する．具体的には以下の項目を検証する．

(1)　livedoor クリップが確かにロングテールな分布 f を持っている．

(2)　希少度は中心極限定理を用いず，たたみこみにより計算すべきである．

(3)　客単価では得られないコンテンツを希少度により発掘可能である．

4.1　livedoor クリップ

livedoor クリップは，ソーシャルブックマークサービスである．livedoor クリップ 2008 年 12 月版データセット[11] は，1,572,742 件のブックマークについて（ユーザ ID，対象 url，

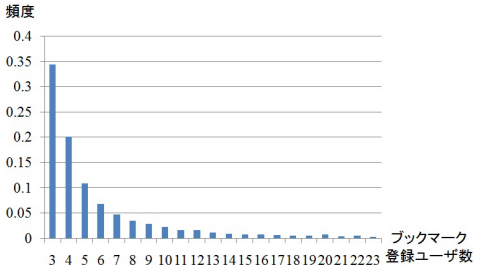

図 1　ブックマークあたりの登録者数の頻度分布（実際は 960 人まで裾が伸びている）
Fig. 1　Frequency distribution of bookmark registrants. Note that this visualization cut off at 23 people but actually continues to 960 people.

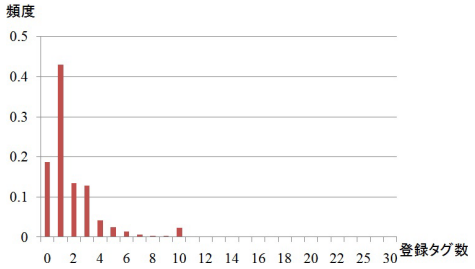

図 2　ブックマークあたりの登録タグ数の頻度分布
Fig. 2　Frequency distribution of number of registered tags for each bookmark.

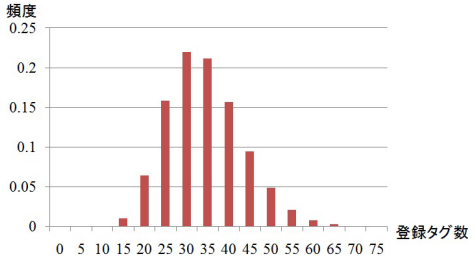

図 3　g^{*20} のヒストグラム（歪度 = 0.5379，尖度 = 0.3668）
Fig. 3　Histogram for g^{*20} (skewness = 0.5379, kurtosis = 0.3668).

クリップした時刻，タグ）を網羅したものである．ブックマーク登録を行う際，タグは 10 個以下の任意の数で登録できる．タグを登録することは必須ではないので，ゼロであることもある．このデータセットには，データセット収録 3 カ月前から存在し，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，希少度により「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見なせる）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．

4.2　来訪と対価の分布

図 1 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 f」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 である．ここから，確かに f は long-tailed（裾の長い）分布であることが分かる．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ほど遠いことが明らかである．

次に**図 2** 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 人あたりの対価支払い額の確率分布 g」に相当する．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であると考えられる[⋆1]．

4.3　希少度の計算

tailstat（付録 A.1 参照）を用いて，図 1 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 2 の分布 g 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

⋆1 実際にはタグの数が 10 を超えるブックマークも 2,764 個存在したが，この理由は不明である．10 以上のタグも受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．

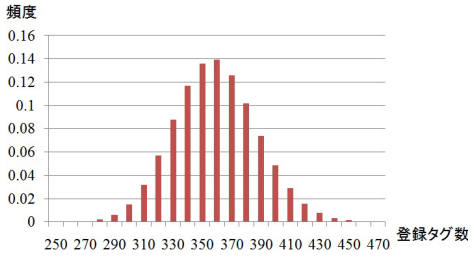

図 4　g^{*200} のヒストグラム（歪度 = 0.1705，尖度 = 0.03244）
Fig. 4　Histogram for g^{*200} (skewness = 0.1705, kurtosis = 0.03244).

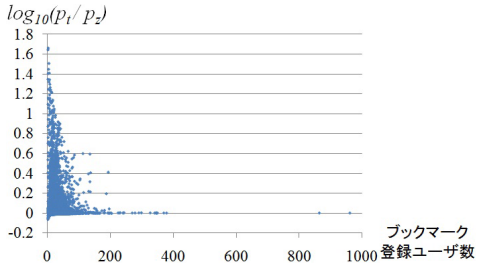

図 5　ブックマーク登録ユーザ数と $\log_{10}(p_t/p_z)$
Fig. 5　Number of bookmark registrants vs. $\log_{10}(p_t/p_z)$.

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 $p_t = P_{g^{*n}}(T \geq t)$ を求める．これがこのサイトの希少度である．また，比較のために同じ条件で図 2 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 $N(n\mu, n\sigma^2)$，$\mu = 1.818$，$\sigma = 2.008$ からの確率 $p_z = P_{\mathrm{norm}}(T \geq t)$ も求めておく．これは希少度の近似値となる．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「故意や悪意による異常事態である」ということができる．

4.4　中心極限定理についての検証

たたみこみにより得られた希少度 p_t と，中心極限定理が成立すると仮定し計算された希少度の近似値 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 3,880）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 5** 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，1 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，中心極限定理を用いた希少度の近似計算は大多数のサイトにおいて不適切であり，たたみこみによる厳密な希少度の計算が求められることが示された．

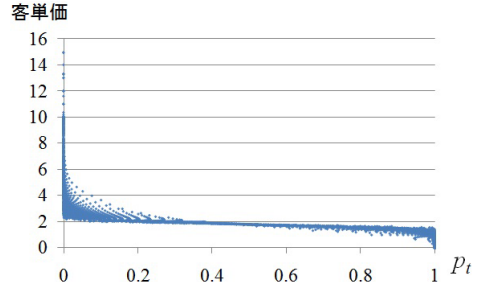

図 6　p_t と客単価の比較
Fig. 6　Average customer spend vs. p_t.

4.5　希少度と客単価の比較

希少度 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．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と考えられる．**図 6** は横軸に希少度 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，希少度 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，より希少度の高い（もしくは低い）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 7 はそれを定量的に示したものである．図 6 から $p_t < 0.1$ のものだけを抽出し，これを「希少度の高いサイトの全体」と定義する．いま，正数パラメータ a を 0 から少しずつ

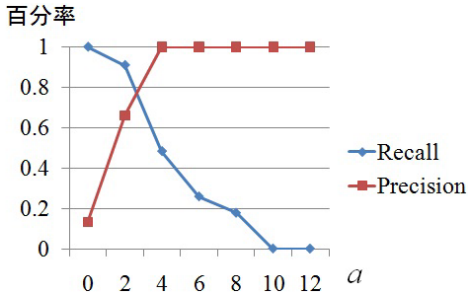

図 7　客単価をクエリとした場合の再現率と精度
Fig. 7　Precision and recall for each average customer spend as a query.

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度の高い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 7 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，希少度の高いサイト全体の 8 割弱程度を網羅できるが，同時に希少度の高くないサイトも 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度判断に限界があることを示している．

また，希少度 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．一方で希少度 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，希少度 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発掘を行えることが示された．

4.6 考　　察

本論文では livedoor クリップを例にとり分析を行った．しかし希少度による分析は，他の「来訪・対価支払い」型ウェブサービスにも広く適用できる一般性を備えている．本節では同様の希少度計算によるコンテンツ分析を別の「来訪・対価支払い」型ウェブサービスに適用する場合の注意事項について考察する．

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．
(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．

これらは，希少度を用いたコンテンツの分析を行う際の前提条件に深く関わってくる項目である．それぞれについて，詳しく述べる．

4.6.1　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見なしやすくなる[*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．

独立試行と見なせることは，希少度の計算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．

4.6.2　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

*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，厳密には独立とはいえない．

行った結果であり，重複する訂正は対価支払いと見なされない．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．希少度によるコンテンツ分析が対象とできるのは，互いに独立であり重複が許される対価支払い量のみである．

4.6.3　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 2）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，希少度の計算に必要な分布 g^{*n} を定義どおりたたみこみにより計算する必要が生じる．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 g^{*n} も既知の正規分布となり，たたみこみ計算は不要である．ただし，よくみられる 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．

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「来訪・対価支払い」型ウェブサービスのモデルを定義し，そのようなモデルにおける個々のコンテンツの価値もしくは異常性の尺度として，既知の母集団分布のたたみこみにより求められる希少度という指標を定義した．また，livedoor クリップの事例を用いて希少度によるコンテンツ分析が有効であることを示し，「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．

希少度を用いた分析で今後の研究課題として有望なのは，分析結果のサービスへのフィードバックである．たとえば livedoor クリップにおいては，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．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．これは商業一般にいえることで，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，希少度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．

謝辞　この研究の遂行にあたって支援をいただいた日本私立学校振興・共済事業団，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いただいたことについて，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) Backstrom, L., Huttenlocher, D., Kleinberg, J. and Lan, X.: Group formation in large social networks: membership, growth, and evolution, *KDD '06: Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, ACM, pp.44–54 (2006).
3) flickr. http://www.flickr.com/
4) Fu, W.: The microstructures of social tagging: A rational model, *CSCW '08: Proc. ACM 2008 Conference on Computer Supported Cooperative Work*, New York, NY, USA, ACM, pp.229–238 (2008).
5) Golder, S. and Huberman, B.A.: Usage Patterns of Collaborative Tagging Systems, *Journal of Information Science*, Vol.32, No.2, pp.198–208 (2006).
6) Google Analytics. http://www.google.com/analytics
7) Google Image Labeler. http://images.google.com/imagelabeler/
8) Goto, M., Ogata, J. and Eto, K.: PodCastle: A Web 2.0 Approach to Speech Recognition Research, *Proc. Interspeech 2007*, pp.2397–2400 (2007).
9) Leskovec, J., Kleinberg, J. and Faloutsos, C.: Graphs over time: Densification laws, shrinking diameters and possible explanations, *KDD '05: Proc. 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, New York, NY, USA, ACM, pp.177–187 (2005).
10) livedoor Clip. http://clip.livedoor.com/
11) livedoor labs EDGE Datasets. http://labs.edge.jp/datasets/
12) 松原望他：統計学入門，東京大学出版会 (1991).

13) Public Data Sets on AWS. http://aws.amazon.com/publicdatasets/
14) Rader, E. and Wash, R.: Influences on tag choices in del.icio.us, *CSCW '08: Proc. ACM 2008 Conference on Computer Supported Cooperative Work*, New York, NY, USA, ACM, pp.239–248 (2008).
15) 梅田望夫：ウェブ進化論，筑摩書房 (2006).
16) vizoo. http://www.visualzoo.com/
17) Wikipedia. http://wikipedia.org/
18) Yahoo!オークション．http://auctions.yahoo.co.jp/
19) YouTube. http://www.youtube.com/
20) クックパッド．http://cookpad.com/
21) ニコニコ動画．http://www.nicovideo.jp/
22) 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).
23) 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).
24) 人力検索はてな．http://q.hatena.ne.jp/
25) 総務省統計局．http://www.stat.go.jp/
26) 楽天．http://www.rakuten.co.jp/

付　　録

A.1　分析補助ツール tailstat による確率分布のたたみこみと確率計算

本節では本論文で取り扱う，確率分布のたたみこみと，その結果得られた分布上における確率計算を簡易に行う分析補助ツール 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．

以下に二項分布を例として実用性を示す．二項分布 $Bi(n, p)$ において，実用上十分な精度を正規分布近似で得るための通常いわれている必要条件は $np > 5$ かつ $n(1-p) > 5$ である[12]．$p = 0.001$ という相当に偏った場合であっても，必要なサンプル数（たたみこみ回数 -1）n は 5,000 であり，これは手元にある普通の性能のノート PC で tailstat によるたたみこみを行っても 10 秒程度で計算が完了する．また，その際に必要な key 数はわずか 5,002 である．このような事実から，式 (3) によるナイーブなたたみこみ計算手法は十分実用的であると判断される．

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見なしたときの正規分布による同種確率の導出（本論文における「希少度」の計算に用いる）
- 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．

(平成 22 年 5 月 12 日受付)
(平成 22 年 11 月 5 日採録)

栗原　一貴（正会員）

1978 年栃木県生まれ．2007 年東京大学大学院情報理工学系研究科コンピュータ科学専攻博士課程修了．Ph.D．日本学術振興会特別研究員（DC2）を経て，同年産業技術総合研究所に入所，現在メディアインタラクション研究グループ研究員．2007 年より千葉県総合教育センター講師を兼任．2007 年から 2008 年にかけて，東京大学大学総合教育研究センター助教および特任助教を兼任．2009 年より東京大学情報学環客員研究員を兼任．ユーザインタフェース，特にプレゼンテーションツール，ICT 技術の教育応用，およびペンコンピューティングに関する研究に興味を持つ．日本ソフトウェア科学会論文賞受賞．電子情報通信学会 MVE 賞受賞．

土谷　洋平

1976 年京都府生まれ．2007 年東京大学大学院数理科学研究科数理科学専攻博士課程修了．Ph.D．同年大原大学院大学会計研究科会計監査専攻助教．2010 年神奈川工科大学基礎・教養教育センター准教授．専攻は数理物理の可積分系．JPSJ Papers of Editors' Choice（日本物理学会欧文論文誌注目論文賞）を 2007 年 8 月に受賞．