統計分析ツール tailstat で本棚.org から選ぶ「僕らの大切な本」

栗原　一貴[†]　　土谷　洋平[‡]

本論文では統計分析ツール tailstat を用いた事例として，書籍情報の共有サービスである本棚.org から「あるコミュニティにより大切にされている書籍」を発掘するサービスの開発について報告する．tailstat 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分析に適する統計分析を可能にするツールである．具体的にはたくさんのコンテンツにユーザがアクセスし，金銭的対価を支払ったりコメントやレーティングなどを行うような Web サービスにおいて，アクセス数ランキング上位ランカーなどのいわゆる「ベストセラー」の陰に埋もれている，「小規模ではあるが特異な人気を博す」コンテンツの発掘を可能にする．

“Our Precious Books in Hondana.org” Selected by Statistical Analysis Tool Tailstat

KAZUTAKA KURIHARA[†]　　YOHEI TUTIYA[‡]

In this paper we report our development of “our precious books in Hondana.org” that searches books loved by certain communities in a book-related information sharing service Hondana.org using a statistical analysis tool tailstat. The tailstat facilitates statistical analyses with small sample sets from given populations, which render the central limit theorem inapplicable. It suits for searching extraordinarily popular contents among some small communities and thus often hidden beneath so called “best-seller” contents in usual best-seller rankings of web services where the visitors are supposed to pay money or spend effort of giving comments or ratings.

1. はじめに

近年普及してきている CGM(Consumer Generated Media)コンテンツ共有サービスやインターネットショッピングサービスでは，トップページに売れ筋やアクセス数のランキングが表示され，一般的な人気を博したコンテンツがより注目を集める仕組みを備えている場合が多い．これは，商店であれば「売れ筋」の商品をもっとも客の目に届く場所に置き，さらに売り上げを加速させるしくみとして一般的な戦略である．しかし多様化の時代である現代においてこのようなサービス上のコンテンツの売上（閲覧回数も含む）はいわゆるロングテールな分布を持ち，そのロングテールの裾野部分の売上がサービス全体の売上に無視できない影響を与えている[1]と言われているため，単純な売れ筋提示だけでは戦略として不十分である．従ってこの裾野を有効活用できるようなデータマイニング，コンテンツ推薦技術の開発が望まれている．特定のユーザやコンテンツの特徴に基づき類似のユーザやコンテンツを推薦する協調フィルタリングはその一例である．

我々はこれまでに統計分析ツール tailstat[2]を開発し，「希少度」という指標により特定のユーザやコンテンツの視点にとらわれず，俯瞰的に「規模を問わず特異な人気を博している」コンテンツを発掘したり，そのような特異さをコンテンツ同士で比較可能にする手法を提案している．本論文ではこの手法をユーザ参加型書籍情報共有サイトである本棚.org[3]に適用し，ある非明示的なコミュニティにより大切にされている書籍を発掘することが可能な web サービス「僕らの大切な本」を開発したことを報告する．

2. 本棚.org のデータ構造と来訪・対価支払いモデルの適用

先行研究[2]に示したように，tailstat によるコンテンツの希少度算出には，対象とする Web サービス上における全てのユーザ活動履歴が既知であり，かつその履歴データが「来訪・対価支払いモデル」にマッピングできることが前提条件である．[4]において公開されている本棚.org の 2010 年 10 月 13 日版データベースでは，entries というテーブルにおいて(book_id, score)という列が存在し，直観的にはこれがそのまま来訪・対価支払いモデルにおける(コンテンツ ID, 対

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

価量)としてマッピング可能であると考えられる．しかし対価量に対応する score は全データにわたって一貫性がなく，またほとんどの場合未入力であったため，ここでは別の対価量として「大切度」を定義し導入する．すなわち，本棚への登録本数が少ないほど１冊１冊への思い入れが強いという仮定をおき，ある本棚の主が本棚の本を大切にしている度合いとして大切度を「その本棚に登録されている本の数の逆数」で表現する．本棚への本の登録を「来訪」，そしてその本棚における本の大切度を「対価支払い」とすることで，tailstat による希少度計算が可能となる．

3. 来訪の分布と対価支払いの分布

来訪・対価支払いモデルにおいて中心となる，来訪の分布 f および対価支払いの分布 g をそれぞれ図１および図２に示す．図１は１冊の本がいくつの本棚に登録されるかを示したものであり，大部分の本は数回未満の登録に留まるロングテール型の分布であることが確認できる．図２は１冊の本が１回の本棚登録の際に得る大切度（0 から 1 までの値をとる）のヒストグラムである．ここから，大多数の本は蔵書数が百～千冊規模の巨大な本棚に登録されており，我々の定義において比較的「大切にされていない」ことが分かる．

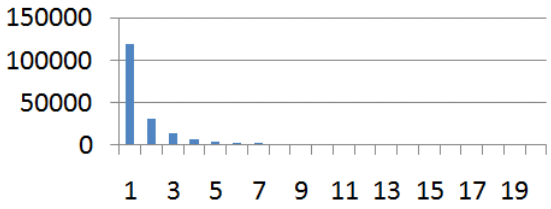

図1 来訪の分布 f．横軸は１冊の本の被登録数．実際は横軸が 239 まで続くが，その一部を切り出してある．縦軸は頻度．

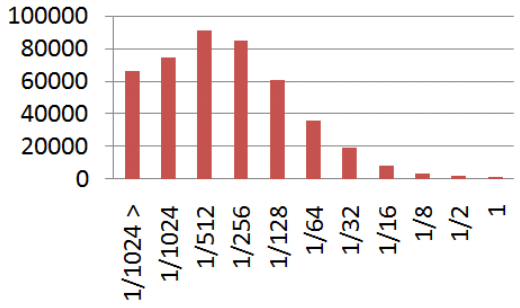

図2 対価支払いの分布 g．横軸（対数軸）は１冊の本が１回の登録で得る大切度で，0 から 1 の間の値をとる．縦軸は頻度．

4. 結果

表１に，tailstat を用いて算出した希少度に基づく書籍発掘サービス「僕らの大切な本」において評価が上位の１５件を例示する．確かに本棚登録数（商店における売上に相当）に依らず，あるそれぞれの非明示的コミュニティにおいて大切にされていると思われる多様なジャンルの書籍が発掘されている．本結果は，http://unryu.org/hondanatail/ においてさらに詳細かつインタラクティブに検索・閲覧可能である．

表1 希少度評価上位 15 件の本

タイトル	本棚登録数
ウェブ進化論	239
ジャズを聴くバカ，聴かぬバカ	26
物理学はいかにして創られたか 上	57
IT 達人の仕事術	24
鋼の錬金術師 11	38
きょうの猫村さん 1	44
竜馬がゆく 1	57
真月譚月姫 1	30
レバレッジ・リーディング	49
白夜行	94
号泣する準備はできていた	12
宇宙創成 上	10
ちくま日本文学全集 長谷川四郎	3
ドラゴン桜公式副読本 16 歳の教科書	3
海辺のカフカ 上	53

5. まとめ

本論文ではユーザ参加型書籍情報共有サイト本棚.org に統計分析ツール tailstat を適用し，書籍発掘サービス「僕らの大切な本」を開発したことを報告した．これは tailstat による来訪・対価支払い型 Web サービスの分析が汎用性をもち，かつ有効なデータマイニング手法であることを示す事例としても重要である．

謝辞 本棚.org のデータの公開をしていただいた，作者の増井俊之氏に感謝する．

参 考 文 献

1) 梅田望夫: ウェブ進化論，筑摩書房 (2006).
2) 栗原一貴，土谷洋平: 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，情報処理学会論文誌, Vol.52, No.2, (2011) in printing.
3) 増井俊之: 本棚通信：控え目なグループコミュニケーション，インタラクション 2005 予稿集, pp.135-142 (2005).
4) http://gyazz.com/本棚/MySQL データ