F-001

ソーシャルブックマーク分析のユーザへの影響と先行性による解決
Effect of SBM Analysis on Users' Adoption and Solution by Bookmarking order

太田 飛鳥† 松澤 智史† 武田 正之†
Asuka OHTA Tomofumi MATSUZAWA Masayuki TAKEDA

1. はじめに

Web における情報量は，近年の Blog や SNS，マイクロブログ等の CGM(Cunsumer Generated Media)の浸透によってますます増加し，その膨大な情報の中から有用な情報を見つけ出すことが求められている．

日々生み出される情報の中には，ユーザの嗜好，興味につながる情報も多く，これらの分析や，協調フィルタリング[1]等を用いた推薦の技術は日々研究が行われている．

ソーシャルブックマークも，ユーザの嗜好，興味を反映した情報を持つ Web サービスの一つである．

ソーシャルブックマークでは，ユーザは興味を持った Web ページに対し，タグと呼ばれる自由な文字列で表現できる分類名を与え，ブックマークとして公開，共有することができる．ソーシャルブックマークにブックマークされた Web ページは，ユーザによって精査され，何らかの価値が認められた Web ページであると考えられるため，有益な Web ページの情報を多く持つサービスとして利用者を獲得している．

また研究領域からも，同じ理由に加え，ユーザに紐づけられた嗜好，興味の情報を多く持つサービスであることから注目され，ソーシャルブックマークを利用した研究が盛んに行われている．

山家ら[2]はソーシャルブックマーク上でページをブックマークしたユーザの数を Web 検索の尺度として，PageRank[3]と組み合わせたランキング手法を提案した．

丹羽ら[4]はソーシャルブックマークから Web ページ推薦を行うことについて，特にタグとユーザ間の関係，Web ページとタグ間の関係を利用した手法を提案した．

推薦の手法によく用いられる協調フィルタリングでは，ユーザの嗜好に合うアイテムを見つけるため，嗜好の近しいユーザが好むアイテムを提示する．これをソーシャルブックマークに適用する場合，ソーシャルブックマークにおけるブックマークが，広大な Web 空間上から発見した Web ページに価値を認め，選択する行為だと言えるため，嗜好とはブックマークした Web ページの傾向となる．

ただし，多くのソーシャルブックマークでは，一定期間でのブックマーク数によるランキング等の簡単なブックマーク分析を提供し，ユーザの利便性を高めている．

こうした情報の提供がユーザのブックマークに影響することは十分に考えられる．Cho ら[5]の報告によると，検索エンジンにおいて上位にランキングされるページは，よく人の目につくことを理由として，高い順位を保持する傾向にあるとされている．同様のことがソーシャルブックマークでも発生することが考えられる．すなわち，ランキングに提示され，よく人の目につくようになった Web ページがよくブックマークされるようになり，多くのブックマークを集めるという構造である．そして同時に，よく目につくランキングに登場する Web ページばかりを優先的にブックマークするユーザの登場も予見される．

先述のとおり，ソーシャルブックマークに協調フィルタリングを適用する際に利用する嗜好情報は，Web 空間からの興味，嗜好にかなった情報の選択であり，推薦するアイテムの根拠は，近しい嗜好のユーザに精査され，価値を認められた Web ページであることにある．ランキングから選択するユーザの行動は，必ずしもそれに合致するものではなく，同列に扱われるべきものではない．

そこで本研究では，ソーシャルブックマークにおけるユーザの振る舞いについて，そうしたブックマーク数ランキングのようなソーシャルブックマーク分析の影響を探り，またそうしたユーザの影響に対処するため，より早期にブックマークするユーザを優位に評価する，先行性という指標を導入することを行った

1.1 対象とするデータ

本研究ではソーシャルブックマークのブックマークログを対象とした分析を行う．今回の実験に用いる分析対象としては，株式会社 Livedoor の提供するソーシャルブックマークである Livedoor クリップ[1]のブックマークログを利用した．同サービスではブックマークログが研究利用を前提として公開されている．データセットの仕様は以下の通りであるとされている．

- 3 つ以上の公開クリップがついているページへのクリップで，3 ヶ月以上前から存在する公開クリップ
- スパマや R18 の除外などはしない（書き出し時点ですでに削除されているデータは除外される）
- ファイル形式は utf8 の csv
- フィールドは必ず ″ ″ でエンクローズされる。値自体に″が含まれる場合はエスケープされる
- フィールドは順に user_id，対象 url，クリップした時刻，タグ（図 1）
- user_id はもともとの id ではなく，シャッフル済の整数
- タグが複数ある場合は空白区切りで結合。タグに改行やタブが含まれる場合は除去済
- 6 ヶ月毎に新しい csv ファイルを書き出す。（以前のデータも引き続きダウンロード可能）

UserID	URL	Timestamp	Tags
1	http://clip.livedoor.com/	2006-06-27 17:24:54	sbm clip これはすごい web2.0
2	http://clip.livedoor.com/	2006-06-27 17:27:46	sbm
3	http://clip.livedoor.com/	2006-06-27 17:46:44	sbm livedoor

図 1 データセットの書式

† 東京理科大学 Tokyo University of Science

[1] http://clip.livedoor.com/

今回は 2009 年 12 月時点で書き出されたものを利用した．データ量は約 246 万行存在し，その中に 33 万 URL，4 万 5 千ユーザのログを含む．

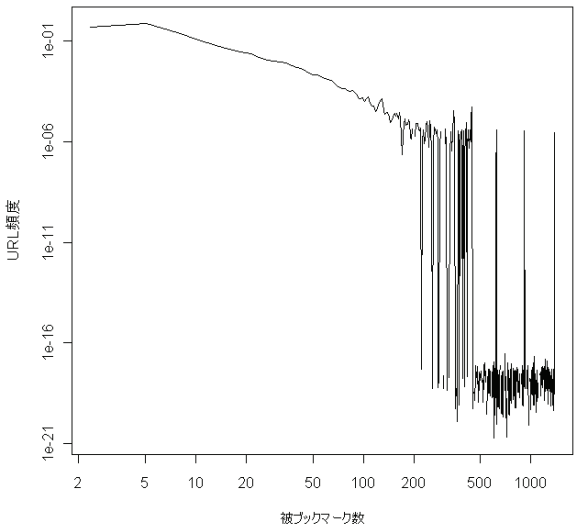

図２　URL の被ブックマーク数分布

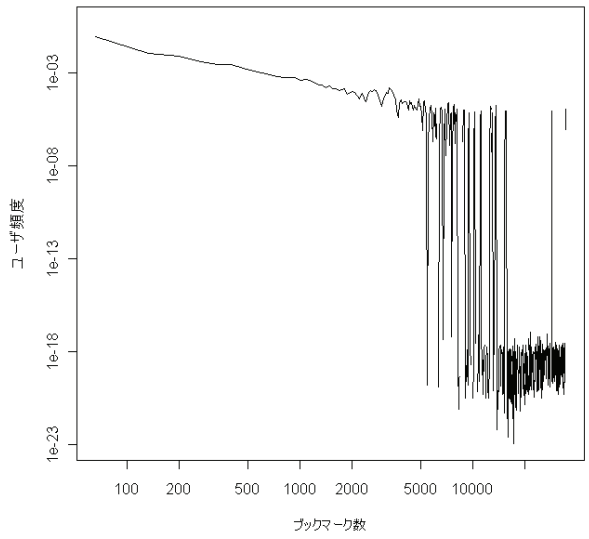

図３　ユーザのブックマーク数分布

2. ソーシャルブックマークの特徴分析

ユーザ分析を行うにあたり，データセットの特徴を知るため，以下の二つについて調べた．

- URL の被ブックマーク数分布
- ユーザのブックマーク数分布

これらを求めたものがそれぞれ図２，図３になる．図２，図３は横軸に URL の被ブックマーク数（ないしユーザのブックマーク数）を取り，縦軸に頻度を取っている．ともに両対数グラフであり，べき乗則に従っていることが見て取れる．

ここから以下の構造が読み取れる．

- 多くの URL は，少ないブックマーク数を獲得する
- 少数の URL は，多くのブックマーク数を獲得する
- 多くのユーザは，少ないブックマークを行う
- 少数のユーザは，多くのブックマークを行う

これらのことから，ソーシャルブックマークを用いた推薦について，先述した「人気ページ」などの影響に加え，多くのブックマークを持つユーザ，多くのブックマークを持つ URL が大きな影響を持つのではないかと推測できる．この推測は次の分析で確認する．

2.1 ユーザを用いた推薦に関する分析

協調フィルタリングの一般的な手法は以下のようになる．ユーザ集合を $A=\{a_1,a_2,a_3,...\}$ アイテム集合を $B=\{b_1,b_2,b_3,...\}$ ユーザ間の類似度を $s(a_i,a_j)$ ユーザのアイテムへの評価値を $r(a_i,b_j)$ で表すとする．

$$score(a_i,b_j)=\sum_k r(a_k,b_j)\times s(a_i,a_k)$$

とする $score(a_i,b_j)$ が得られる．このスコアをユーザ a_i に関して全てのアイテム b_j で求め，ランキングする．ランキングの結果得られたスコアの高いアイテムを優先的に推薦する．以上が，協調フィルタリングの手法である．

ここで，ユーザ a_i の全ユーザへの推薦の影響度を示す指標 $sims(a_i)$ を考える．

協調フィルタリングのアルゴリズムから，あるユーザへの推薦について，そのユーザとの類似度が高いユーザほど大きな影響を与えることがわかる．そのため，$sims(a_i)$ はユーザ a_i の全ユーザに対する類似度の和で表現できる．

$$sims(a_i)=\sum_k s(a_i,a_k)$$

実験では，$sims(a_i)$ を求めるための類似度の計算式 $s(a_i,a_j)$ は Dice 係数を用いて求めることとした．Dice 係数は集合の類似度を求める計算式である．ユーザ a_i がブックマークしている URL の集合を $bm(a_i)$ と表すとして，Dice 係数は以下のように求めることができる．

$$s(a_i,a_j)=Dice(a_i,a_j)=\frac{\left|bm(a_i)\cap bm(a_j)\right|}{\left(\left|bm(a_i)\right|+\left|bm(a_j)\right|\right)/2}$$

以上より，$sims(a_i)$ の分布を求めると，図４の分布が得られる．図４は横軸に $sims(a_i)$ の値，縦軸にユーザの出現頻度を取る．

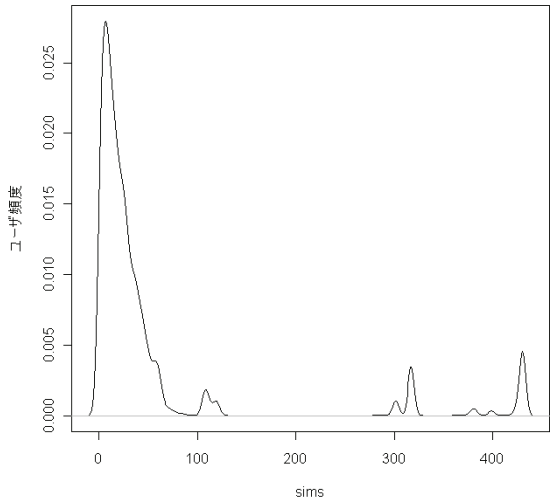

図４　推薦影響度分布

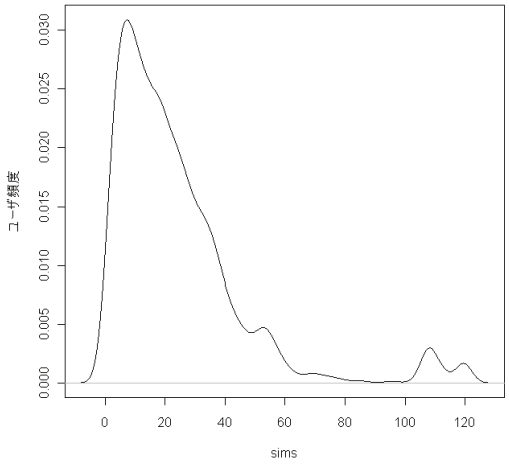

図５　ユーザブックマーク数条件 20

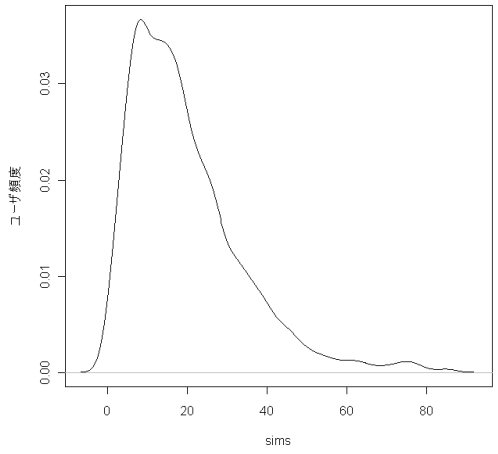

図６　ユーザブックマーク数条件 50

ただし，図４を求めるにあたって，あらかじめ以下の条件を与えている．

- ブックマーク件数 10 件以上のユーザのみ扱う
- 被ブックマーク件数 10 件以上の URL のみ扱う

そのためこのグラフにおける分析対象のユーザは 9523 人まで減少する．

図４には最も高い左側の山に続いて，３つの山が存在している．これは $sims(a_i)$ が高く，かつ近しい値を持つユーザのグループが存在していることを示す．

このグループについて知るため，以下の実験を行う．

2.1.1 ユーザのブックマーク数と影響度

分布を求める計算に用いるユーザの条件を，ブックマーク件数 20 件，50 件と引き上げ，同様に分布を求める．

その結果，図５，図６がそれぞれ得られる．

条件により分析対象ユーザ数がさらに減少する．図５では 6656 人，図６では 4046 人に減少した．

実験の結果，ユーザの条件をブックマーク数 20 とした図５では図４の右にあった二つの山が消え，またユーザの条件をブックマーク数 50 とした図６では最も大きな山の隣にあった山もほぼ消えている．

これらの山を構成するために，ブックマーク数 20 から 50 件以下程度の，比較的ブックマーク数の少ないユーザが大きな影響を及ぼしていたことがわかる．

2.1.2 URL の被ブックマーク数と影響度

同様に，分布を求める計算に用いる URL の条件を変えて実験を行う．図７は URL の被ブックマーク件数条件を 50 件まで引き上げた結果のグラフである．

分析に利用できる URL 数が減り，それによりブックマークを持たない扱いとなったユーザが分析対象から除かれ，分析対象のユーザ数は 7590 人になった．図２から，被ブックマーク件数 50 件以上は多くの URL を分析対象から除外する厳しい条件であることがわかるが，図７を図４と比べると右端に存在する山が依然残っていることがわかる．

そこで，さらに次の条件のもとで実験を行う．

- ブックマーク件数 10 件以上のユーザのみ扱う
- 被ブックマーク件数 10 件以上の URL のみ扱う
- 被ブックマーク件数 200 件（100 件）以下の URL のみ扱う

図４を得た際の条件に加え，URL の被ブックマーク件数に上限を与えた．

実験の結果図８，図９を得た．分析対象ユーザ数は図８では 9131 人，図９では 8682 人である．それぞれ，図８では図４の右端にあった二つの山が消え，図９ではさらにもう一つ山が消えている．

これにより，これらの山を構成するために大きな影響を及ぼしたユーザが，主に被ブックマーク件数 100 から 200 件以上程度の，比較的被ブックマーク数の多い URL を主にブックマークしていたことがわかる．

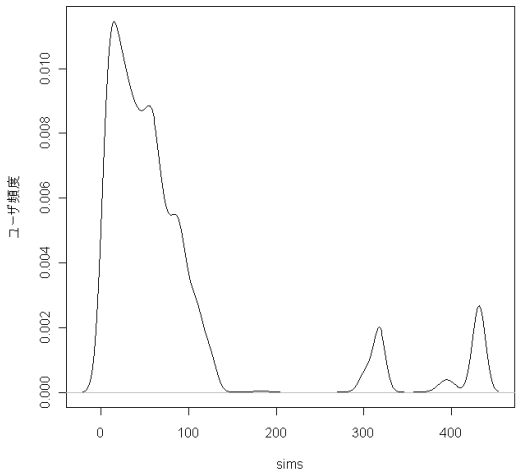

図７　URL 被ブックマーク数条件 50

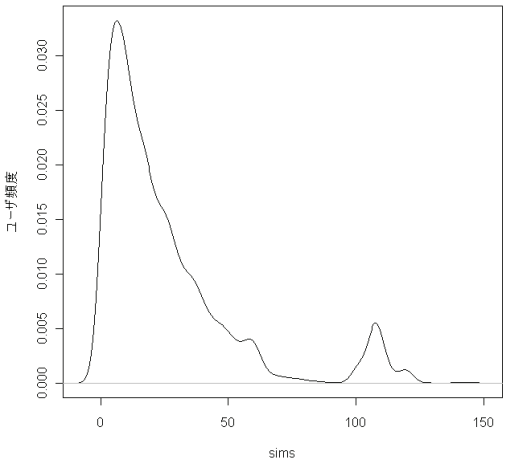

図８　URL 被ブックマーク数条件上限 200

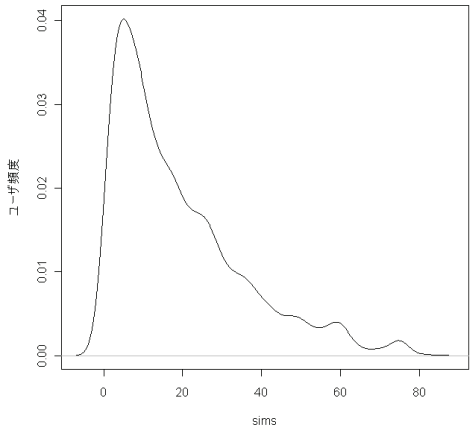

図９　URL 被ブックマーク数条件上限 100

2.1.3 考察

以上の結果から，図４において３つの低い山を構成した，比較高い $sims(a_i)$ 値を持ち，$sims(a_i)$ 値の似通ったユーザグループ（以下グループ）について考察する．

グループがグラフ上で構成する山は，以下のどちらかを満たすことで出現しなくなった．

- 多くのブックマーク数を持つユーザのみ扱う
- 多くの被ブックマーク数を持つ URL を扱わない

ここから以下のことが推測できる．

ブックマーク数が少なく，かつ多くの被ブックマーク数を持つ URL ばかりをブックマークしているユーザが存在する．彼らは相互に同種のアイテムをブックマークしているので相互の Dice 係数が高く，また彼らがブックマークしている URL は多くのユーザがブックマークしているので，他のユーザとの Dice 係数も比較的高い．こうした集団が存在し，グループとなっている．

このことを確かめるため，ブックマーク数と $sims(a_i)$ の散布図を取ると，図１０が得られる．

図１０は横軸を $sims(a_i)$ 値，縦軸をブックマーク件数とした，ユーザの散布図である．図１０から，ブックマーク数が少なく，$sims(a_i)$ 値の高いユーザがまとまって存在していることを確認できる．

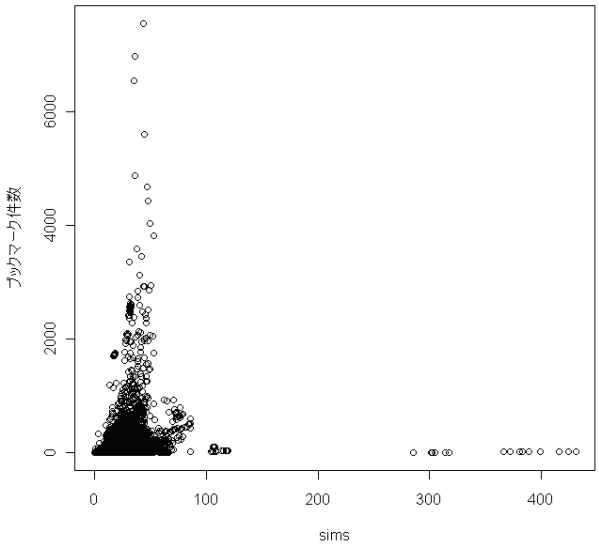

図１０　推薦影響度とブックマーク件数

2.2 先行性を考慮した推薦影響度

ここまででブックマーク数が少なく，かつ多くの被ブックマーク数を持つ URL をブックマークしているユーザについて論じた．ここで，あるユーザが多くの被ブックマーク数を持つ URL を選択的にブックマークしている場合，次の二つの可能性が考えられる．

１．ブックマークした後に URL が多くの被ブックマークを集めた
２．既に多くの被ブックマークを獲得している URL をブックマークした

ここで，１のブックマークをしたユーザは２のブックマークに影響を与えていると考えることができる．Song ら[6]は商品の購入行動について次のような仮説を唱えている．ユーザの購入行動に関する情報は，何らかの形によってユーザ間を伝搬し，ユーザの購入行動に影響を与える．Song らはその伝搬可能性を同一のアイテムを先に購入した頻度に結び付けて表現した．つまり，ユーザ a_i がユーザ a_j よりも先に商品を購入した頻度が高いほど，ユーザ a_i の購入行動がユーザ a_j に影響を与えて同一の商品を購入させる可能性は高い．

ここでも同様の仮説に従って考える．ソーシャルブックマークに実装されている簡単なブックマーク数ランキングの影響や，あるいは他の何らかの形で，ある URL がブックマークされた情報が伝わり，それがブックマークに影響を与える．つまり，１のブックマークが２のブックマークを生み出す要因になっていると考える．

そこでより１のブックマークを重視して扱うため，推薦の指標に用いてきた類似度に，さらにブックマークの順序関係を加味した，先行性類似度和 $fastsims(a_i)$ を定義する．

$$fastsims(a_i) = \sum_k fs(a_i, a_k)$$

$$fs(a_i, a_j) = \frac{fc(a_i, a_j)}{(|bm(a_i)| + |bm(a_j)|)/2}$$

$fc(a_i, a_j)$ はユーザ a_i がユーザ a_j よりも早くブックマークしているアイテムの数である．$sims(a_i)$ と比べ，類似度の計算時に先行してブックマークしているアイテムのみが加味される点が異なる．

この結果図１１のようになる．図１１は横軸を $fastsims(a_i)$，縦軸をユーザの出現頻度としたグラフである．分析対象ユーザ数は 9523 人で，図４での分析人数と同等である．また，図４などのように突出した山は存在しない．

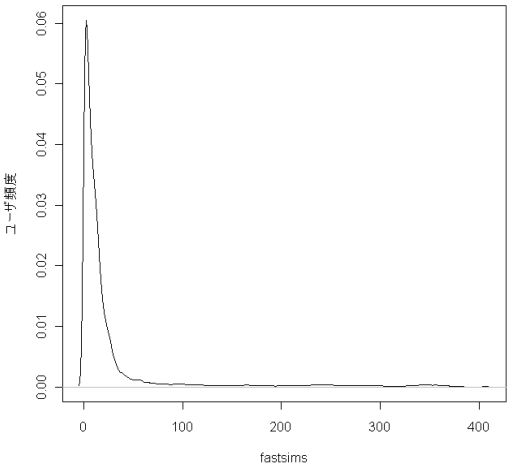

図１１　先行性類似度和分布

2.3 ユーザ先行性

先行性の指標に用いた $fc(a_i, a_j)$ について考える．$fc(a_i, a_j)$ はユーザ a_i がユーザ a_j より早くブックマークしている URL 件数を示す．これはユーザ a_i がユーザ a_j のブックマークに与える影響を示す指数であると考えることができる．そこで，ユーザ全体に対する影響度 $fcsum(a_i)$ を以下のように定義し，その分布を求める．

$$fcsum(a_i) = \sum_k fc(a_i, a_k)$$

結果は図１２のようになる．図１２は横軸を $fcs(a_i)$ 値，縦軸をユーザの出現頻度とした両対数グラフであり，べき乗則に近い推移を示している．分析対象ユーザ数は 9523 人である．

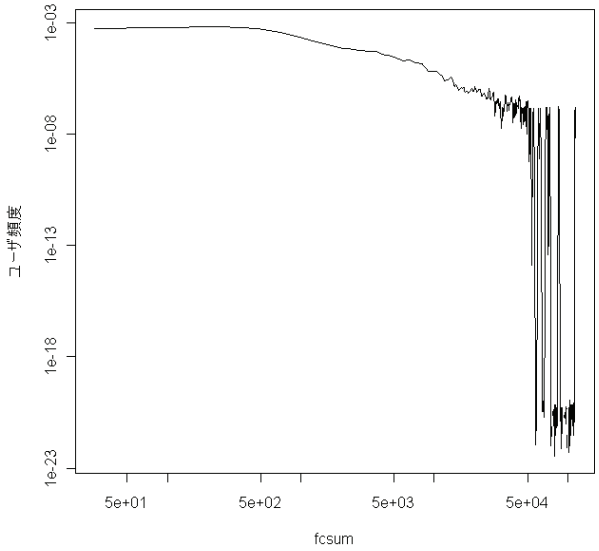

図１２　先行性影響度分布

$fc(a_i, a_j)$ に対し，さらにユーザ a_j の選択がどれだけユーザ a_i に影響を受けているのかを示す影響度の指標として $fc'(a_i, a_j)$ を定義する．

$$fc'(a_i, a_j) = \frac{fc(a_i, a_j)}{|bm(a_j)|}$$

$fc'(a_i, a_j)$ はユーザ a_j に対する推薦の再現率に先行性を加味したものに相当する．$fcsum(a_i)$ と同様に，a_i について $fc'(a_i, a_j)$ の和を求める．

$$fcsum'(a_i) = \sum_k fc'(a_i, a_k)$$

結果は図１３のようになる．図１３は横軸を $fcsum'(a_i)$ 値，縦軸をユーザの出現頻度とした両対数グラフであり，分析対象ユーザ数は 9523 人である．

ユーザ数のピークは $fcsum'(a_i)$ 最小のときではなく，$fcsum'(a_i)$ が小さなユーザはそれほど多くないことがわかる．$fcsum(a_i)$ の例と比べると，$fc'(a_i, a_j)$ 値は比

較対象ユーザ a_j のブックマーク数が小さい時値が大きくなる傾向にある．したがってこの結果は，ブックマーク数が少なく，かつ多くのユーザに先行されているユーザの存在を示している．

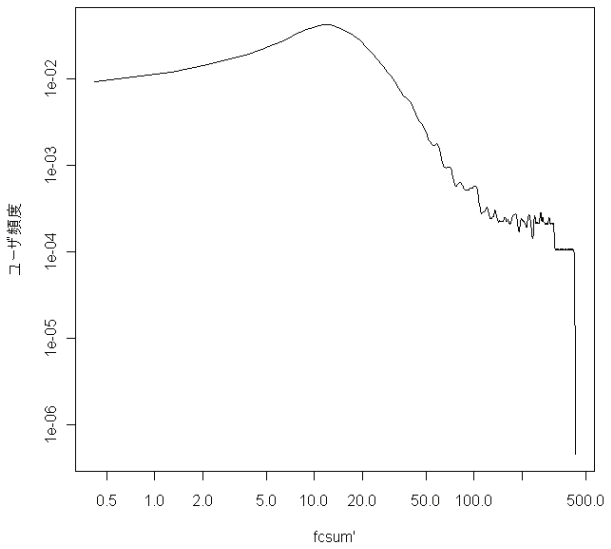

図１３　先行性再現率分布

3. 結論

本論文では，ソーシャルブックマーク上でのユーザの振る舞いについて，ソーシャルブックマークの持つブックマーク分析に大きな影響を受けるユーザの存在を示した．また，それが協調フィルタリング等の計算に影響しうることを示した．

彼らは一種のはずれ値のようなユーザであるが，彼らを排除するためには，ユーザのブックマーク件数を条件とした場合には分析対象のユーザを半減させる程度の条件をおかなければならず，また URL に関して条件を与える場合い，本来類似度の計算に有用であるはずの多くのユーザにブックマークされている URL 群を分析から取り除かなければならない．

そのため本研究では推薦指標に先行性というブックマークの時間的前後関係の要素を組み込むことを提案した．推薦指標の算出に先行性を組み込むことで，先述したはずれ値にあたるユーザの影響は小さく抑えられ，ユーザの Web 空間からの選択に基づいた推薦が行えるものと考える．

参考文献

[1] Sarwer, B., Karypis, G., Konstan, J. and Riedl, J., “Item-based Collaborative Filtering Recommendation Algorithms” , Proceedings of the 10th international conference on World Wide Web , pp285-295 (2001)

[2] Yanbe, Y., Jatowt, A., Nakamura, S., Tanaka, K., “Can social bookmarking enhance search in the web?” , Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries , pp107-116 (2007)

[3] http://infolab.stanford.edu/~backrub/google.html

[4] 丹羽 智史, 土肥 拓生, 本位田 真一,"Folksonomy マイニングに基づく Web ページ推薦システム",情報処理学会論文誌, Vol. 47, No. 5.pp. 1382-1392.(2006)

[5] Cho, J., Roy, S., Adams, R., “Page Quality : In Search of an Unbiased Web Ranking”,Proceedings of the 2005 ACM SIGMOD international conference on Management of data , pp551-562(2005)

[6] Song, X., Tseng, B., Lin, C.Y., Sun, M.T., “Parsonalized recommendation driven by information flow” , Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval , pp509-516 (2006)