The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

ソーシャルタギングシステムに見られる特異的なタグの成長

Anomalous popularity growth in social tagging ecosystems

橋本康弘 *1 Yasuhiro Hashimoto　岡瑞起 *2 Mizuki Oka　池上高志 *3 Takashi Ikegami

*1*3東京大学総合文化研究科 The University of Tokyo　*2筑波大学システム情報系 University of Tsukuba

In social tagging systems, the diversity of tag vocabulary and the popularity of such tags continue to increase as they are exposed to selection pressure derived from our cognitive nature and cultural preferences. This is analogous to living ecosystems, where mutation and selection play a dominant role. Such population dynamism, which yields a scaling law, is mathematically modeled by a simple stochastic process—the Yule–Simon process, which describes how new words are introduced to the system and then grow. However, in actual web services, we have observed that a large fluctuation emerges in the popularity growth of individual tags that cannot be explained by the ordinary selection mechanism. We introduce a scaling factor to quantify the degree of the deviation in the popularity growth from the mean-field solution of the Yule–Simon process, and we discuss possible triggers of such anomalous popularity behavior.

1. はじめに

ソーシャルタギングシステム [1] とは現代のウェブサービスで用いられるアノテーションシステムの一種である．サービスを利用するユーザは，これによって自身あるいは他者が投稿した画像やテキストなどの共有リソースに「タグ」と呼ばれる短い文字列を関連付けることができる．タグとして使われる文字列はユーザの裁量に大きく任されており，その結果，タグの語彙や個々のタグの累積的な出現回数は，ユーザの文化的選好や認知的特性を反映しつつ，共に増加していく．つまり，拡大を続ける語彙空間の中で，個々のタグは人々の有限の注意を奪い合いながら，勢力を拡大していく．このタグ同士の競争的状況は，変異と淘汰が支配する生物の生態系になぞらえることができる．実際，語彙とタグの成長過程は，生物種の競争的な増加過程をモデル化した極めて単純な確率過程—Yule–Simon 過程 [2, 3]—を用いて粗く近似できることが知られている [4]．しかし，現実のウェブサービスから得られたデータを詳細に分析すると，個々のタグの成長がモデルの期待値からは大きく外れるケースが存在することが分かっている．このずれの大きさを定量化するとともに，またこのずれを生み出す社会的，あるいは認知的なメカニズムについて議論したい．

2. 準備

ここではソーシャルタギングをタグの出現シーケンスが単位時間ごとに一つずつ伸びていく過程と単純化する．そして，新たにシーケンス末尾に追加されるタグがこれまで使われたことのないタグなのか，あるいは既にシーケンス内に存在するタグなのかを確率的に決定する．前者の確率を $\alpha(t)$ と表記し，これを新規性レートと名付ける．t は離散的な時刻である．一般には t が大きくなる—サービスが成熟し，使われる語彙が大きくなる—にしたがい新規性レートは減少していく．追加するタグを既存のタグから選択する場合，どのタグを選び出すかはモデル化の重要なポイントになる．しかし，いくつかのウェブサービスでは，平均的にタグは出現回数に比例した確率で選ばれることが分かっている．これはネットワークサイエンスの分野で優先的選択 [5] として知られるメカニズムと等価である．新規性レートと優先的選択性を用いて，タグの成長（累積的な出現数の増加）は以下の時間発展式で記述できる:

$$n_i(t+1) = n_i(t) + \bar{\alpha}(t)\frac{n_i(t)}{N(t)}. \tag{1}$$

ここで $n_i(t)$ はタグ i の時刻 t における累積出現数，$N(t)$ はシーケンス長で，定義より $N(t) = t$，$\bar{\alpha}(t)$ は新規性レートの補確率 $1 - \alpha(t)$ である．十分大きな t に対して $\bar{\alpha}(t) = 1$ を仮定し，さらに t と $n_i(t)$ について連続体近似を仮定すると，近似解 $n_i(t) = t/t_i^1$ が得られる．t_i^1 はタグ i の出現数が 1 になる時刻，つまり初めて使われた時刻である．我々の興味は現実のタグがこの平均場解が示す挙動からどの程度逸脱するかにある．

3. 偏差係数

逸脱の程度を定量化するため，ここでは [6] によって導入されたアイデアを拡張して以下のように偏差係数 Φ_i を定義する．

$$\Phi_i \equiv \frac{\hat{n}t_i^1 - t_i^1}{t_i^{\hat{n}} - t_i^1}. \tag{2}$$

ここで $\hat{n}$ は累積出現数の目標値であり，分子の $\hat{n}t_i^1$ はタグ i が平均場近似により目標値に到達すると期待される時刻である．分母の $t_i^{\hat{n}}$ はタグ i が実際に目標値に到達した時刻である．つまり，Φ_i とはタグ i が目標値に到達するために実際にかかった時間とその期待値の比である．これが 1 であれば，目標値に到達するまでのタグの振る舞いは平均場解と等しく，1 より大きければ成長は速い（小さければ遅い）ことを意味する．この Φ_i をデータがカバーする時間の範囲で目標値に到達したすべてのタグについて計算し，その傾向について分析する．

4. 実験結果

ここでは [7] によって収集された写真共有サービス Flickr のデータを用いる．図 1 に示した図はそれぞれ Φ_i と $t_i^{\hat{n}}$ の関係（左），および Φ_i と最終的な累積出現数の関係（右）であり，

連絡先: 橋本康弘: hashi@sacral.c.u-tokyo.ac.jp

色はタグの濃度を表している．目標値の決め方としては，タグが作られた初期の成長過程を見るために $\hat{n} = 10$，さらに，ある程度十分な出現数を獲得するまでの成長過程を見るために $\hat{n} = 10^3$ を用いた．Yule–Simon 過程の結果と比較すると，実データには偏差係数が平均場解 $\Phi_i = 1$ の周りに集まる集団と，偏差係数の非常に大きな集団が存在することが見て取れる．この偏差係数の大きな集団は $\hat{n} = 10^3$ ではほぼ消失し，この集団が成長の初期段階に固有の特徴であることを示している．また，この集団を除外してもなお平均場解の周りの裾野は広く，平均場解から大きく逸脱するタグが多く存在することを示している．さらに，最終的な出現数の大きなタグは偏差係数の大きなタグではなく，平均場解の近くに集まっていることがわかる．これは Yule–Simon 過程にも共通する特徴である．

5. 考察

仮に，優先的選択ではなく，$n_i(t+1) = n_i(t) + \bar{\alpha}c$（$c$ は定数）という固定レートの成長を仮定すると，成長の初期段階に見られる偏差係数の大きな集団をよく説明できる．つまり，成長の初期段階において，タグ集団は優先的選択に従う集団—競争的状況で使用されるタグ—と，逆に固定レートで成長する集団—競争相手の多寡に依存せず一定時刻の中で集中的に使用されるタグ—が存在すると考えられる．また，平均場解の周りの裾野が広くなる要因としては，最近使われたタグが優先的に選択される，一方でタグは時間的に忘却される，という人間の認知的な特性が考えられる．人間の注意の時間減衰やメモリー効果について扱った研究は近年少しずつ増えてきているが [8, 9]，ここでのメカニズムについては具体的には明らかではない．また，累積出現数の大きなタグが平均場解の周りに集まる点についても，実データと Yule–Simon 過程で理を一つにしているかについては検討が必要である．また，本分析では主にタグの選択を問題にしているが，それはタグの生成メカニズムとも不可分ではないはずである．新しいタグが新しい生態的ニッチを開拓し，それによって既存のタグの意味合いが変化し，タグの選択に再帰的に影響を与えることは十分に考えられる．このような問題意識を共有して，新規性の発生と選択（採用，伝播）という広い視点に基づいて議論したい．

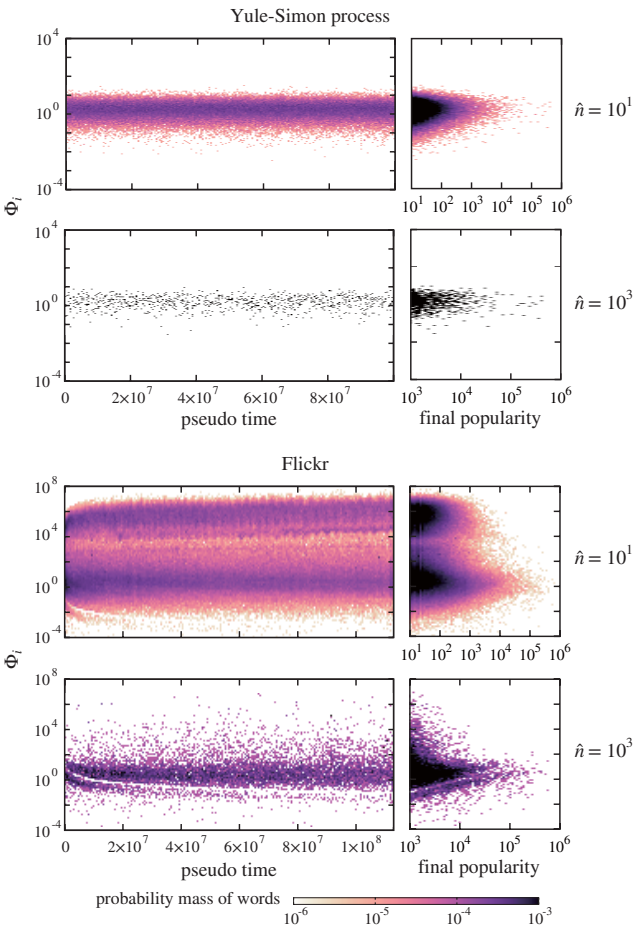

図 1: 縦軸は偏差係数，横軸は目標値に到達した時刻（左）と最終的な累積出現数（右）．縦軸はともに対数分割（左はビン数 80，右はビン数 100），横軸は左は線形分割（ビン数 200）で右は対数分割（ビン数 100）．分割したセルの色はそこに入るタグの数．全体的に Yule–Simon 過程に比べて実データの方が偏差係数の値が 1 から大きく外れている．

参考文献

[1] Daniel H. Pink. Folksonomy. `http://www.nytimes.com/2005/12/11/magazine/folksonomy.html`, 2005-12-11. Last retrived 2017-01-24.

[2] Herbert A. Simon. On a class of skew distribution functions. *Biometrika*, 42(3–4):425–440, 1955.

[3] G. Udny Yule. A mathematical theory of evolution, based on the conclusions of dr. j. c. willis, f. r. s. *Phil. Trans. Roy. Soc. B*, 213:21–87, 1925.

[4] Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero. Semiotic dynamics and collaborative tagging. *Proc. Natl. Acad. Sci. USA*, 104(5):1461–1464, 2007.

[5] Albert-lászló Barabási, Réka Albert, and Hawoong Jeong. Mean-field theory for scale-free random networks. *Physica A*, 272(1):173–187, 1999.

[6] Yasuhiro Hashimoto. Growth fluctuation in preferential attachment dynamics. *Phys. Rev. E*, 93(4):042130, 2016.

[7] Olaf Görlitz, Sergej Sizov, and Steffen Staab. Pints: Peer-to-peer infrastructure for tagging systems. In *Proceedings of the 7th International Conference on Peer-to-Peer Systems*, IPTPS'08, pages 19–19, Berkeley, CA, USA, 2008. USENIX Association.

[8] Pietro Della Briotta Parolo, Raj Kumar Pan, Rumi Ghosh, Bernardo A. Huberman, Kimmo Kaski, and Fortunatom Santo. Attention decay in science. *J. Informetr.*, 9(4):734–745, Sep 2015.

[9] Ciro Cattuto, Vittorio Loreto, and Vito Domenico Pietro Servedio. A yule-simon process with memory. *Europhys. Lett.*, 76(2):208–214, 2006.