1H3-4

ソーシャルタギングにおけるタグ出現数のゆらぎ

橋本康弘 *1 Yasuhiro Hashimoto　佐藤晃矢 *2 Koya Sato　岡瑞起 *1 Mizuki Oka　池上高志 *3 Takashi Ikegami

*1筑波大学システム情報系情報工学域
University of Tsukuba, Division of information Engineering

*2筑波大学大学院システム情報系
University of Tsukuba, Graduate School of Systems and Information Engineering

*3東京大学大学院総合文化研究科
University of Tokyo, Graduate School of Arts and Sciences

We found that the typical statistical behavior observed in some actual social tagging systems are well-explained by the Yule–Simon process. The behavior followed Zipf's law in a frequency-rank distribution of tag usage and Heaps' law in the growth of overall vocabulary size. However, the experiment exhibited another unique behavior that can't be explained by the ordinary Yule–Simon process, that is, a large fluctuation in the growth of individual word occurrences. We show how such large fluctuation behaves and discuss on what should be incorporated into the model to reproduce such anomalous behavior.

1. ソーシャルタギングシステムとは

ソーシャルタギングシステムとは，写真や動画，ウェブページ，科学論文などの情報リソースをウェブ上で他者と共有する際に広く用いられている，現代的な情報検索システムの一種である．ソーシャルタギングシステムでは，人々は “タグ” と呼ばれる任意の文字列のセットを将来的な検索のためにメタ情報として共有リソースに与える．これは専門家や管理者がトップダウン的に行ってきた伝統的な分類システムとは対照的に，サービスのユーザ自身が自由に分類の語彙を生み出すところに大きな特徴がある．タグの語彙は，我々の日々の生活や環境，注意，問題意識，その表現の多様性を反映した “コード” であり，ソーシャルタギングシステムは多様なコードの生成と発生によって発展する一種の集団現象としての振る舞いを示す．

以上の観点から，以下の 2 つの問いはその言葉以上の意味を持つ：

1. いつ，どのような頻度で新しい語彙は生み出されるのか？
2. それぞれの語彙はどのような頻度で用いられるのか？

最初の問いは，新しい概念の生成を通した我々の認知空間の成長則に対する問いでもある．そして二つ目の問いは，我々は現実にどのような意味的空間を生きているのかという問いである．新しい語彙は生活の新しい意味や新しい可能性を提供する一方で，人々の選好を反映した淘汰圧に曝され，あるものは成長しあるものは衰退し，そしてときに人々の新しい行動の引き金となる．言うなれば，ソーシャルタギングのダイナミクスとは，人々の行動と，人々の認知空間の外化としての語彙の共進化システムと捉えることができる．

2. Yule–Simon 過程

前節で挙げた 2 つの問いからそのまま一つの数理モデルを立てることができる．時間を離散的とし，各時刻にタグが一つずつ出現するタグの時系列を考える．出現するタグは以下のルールに従うものとする：

連絡先: 橋本康弘，筑波大学，hashi@cs.tsukuba.ac.jp

1. 確率 α で新しい語彙が出現する．
2. 補確率 $1-\alpha$ で時系列の中の既存のタグが選択される．

これは Yule–Simon 過程 [Simkin 11] と呼ばれる古典的な数理モデルそのものであり，2 つのルールは前節の 2 つの問いにそれぞれ直接答えたものになっている．すなわち，新しい語彙は一定確率でランダムに生成され，各語彙は過去に出現した回数に比例した出現確率を持つ，という描像である．この単純な数理モデルは実際のソーシャルタギングの振る舞いを調べる際のよい比較対象になるだろう．我々の関心は，現実がどこまで Yule–Simon 過程に従い，あるいはどの点で逸脱するのかを明らかにし，その機構を考察することである．

3. 実験結果

実際のソーシャルタギングのデータとして Delicious，Flickr の公開データ [Görlitz 08] と，Tunnel 社より提供された RoomClip のデータを分析する．Yule–Simon 過程では，その定義により語彙の増加はアノテーション数のベキに比例し（Heaps 則），語彙の出現数の累積確率分布は指数 $\alpha-1$ のベキ分布（Zipf 則）に従うことが知られている．これは人間の言語活動に見られる典型的な統計的特徴でもある．上記 3 つのサービスについて調査した結果を図 1 と 2 に示す．語彙の増加についてはいずれも Heaps 則に従い，また Delicious と RoomClip については Zipf 則にも従う．得られる Zipf 則の指数は Heaps 則の指数と Yule–Simon 過程の枠組みからよく説明することができ，全体的な振る舞いを見る限りにおいて，実際のソーシャルタギングシステムの振る舞いは Yule–Simon 過程によく合致している．

次に新しい観点として，“個別のタグについて” 出現数とその期待値の関係を見る．つまり，ある語彙が初めて出現してから一定期間経過後までに出現した累積回数について，平均場近似が予測する期待値と比較し，その逸脱のスケールについて確率分布を求める．具体的には，語彙 i が生成された時刻を t_i，観察の時間スケールを $\lambda > 1$ とし，時刻 λt_i における累積出現数 $n_i(\lambda)$ をカウントする．仮に α 一定とした場合の平均場近似が予測する期待値は $\lambda^{1-\alpha}$[Barabási 99] であるから，両者

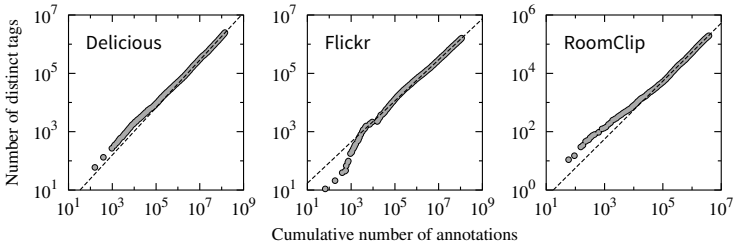

図 1: 語彙数の増加．横軸はアノテーションの総数，縦軸は語彙数．点は実験結果，点線は 10^5 アノテーション以降（RoomClip のみ 10^4 以降）のベキ曲線へのフィッティング．指数は左からそれぞれ 0.81，0.70，0.99．

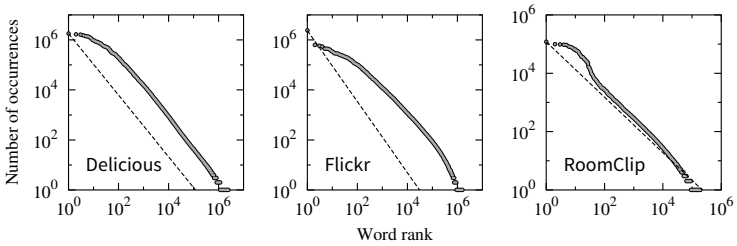

図 2: 頻度–ランク分布．横軸は出現数の降順で並べたタグのランク，縦軸は出現数．点は実験結果，点線は左からそれぞれ $y \propto x^{-1/0.8}$，$y \propto x^{-1/0.7}$，$y \propto x^{-1/0.95}$ を参考として引いた．これらの指数は Yule–Simon 過程から導かれる値．

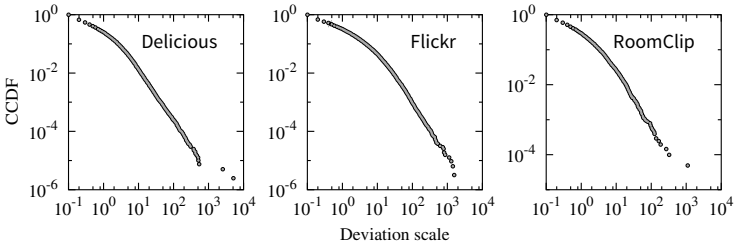

図 3: 期待値からの逸脱のスケールの確率分布（$\lambda = 10$ の場合）．横軸が逸脱のスケール，縦軸はその補累積確率．横軸の値が 1 のとき期待値と等しいことを意味する．

の比 $n_i(\lambda)/\lambda^{1-\alpha}$ を逸脱のスケールと定義する．得られた確率分布を図 3 に示す．いずれのサービスも逸脱のスケールはベキ分布を示しており，これは Yule–Simon 過程が期待する指数分布 [Hashimoto 15] とは明らかに異なる．

4. 議論

分析に用いたすべてのサービスで，個々の語彙の出現数のゆらぎが期待される指数分布ではなくベキ分布を示したことは，そこに大きなゆらぎを生成する何らかの普遍的機構の存在を示唆している．例えば，Simon は Zipf 則を与える必要条件として，語彙の出現確率が過去の出現回数に比例する関係よりも緩い条件を示しており [Simon 55]，これに倣えば，Zipf 則や Heaps 則といった全体的な性質を維持したまま，語彙の選択機構を変更することは可能である．出現数のゆらぎに見られるタグの個性を支配する要因は語彙に固有の（例えば *fitness* のような）ものなのか，あるいは Yule–Simon 過程とは異なる何らかのフィードバック機構によるものか，あるいは新しい語彙の生成とも相関を持っているのか，新たな視点と実験が必要である．

謝辞

貴重なデータを提供いただいた Tunnel 株式会社に感謝いたします．

参考文献

[Barabási 99] Barabási, A.-l., Albert, R., and Jeong, H.: Mean-field theory for scale-free random networks, *Physica A*, Vol. 272, No. 1, pp. 173–187 (1999)

[Görlitz 08] Görlitz, O., Sizov, S., and Staab, S.: PINTS: Peer-to-Peer Infrastructure for Tagging Systems, in *IPTPS2008, Proceedings of the Seventh International Workshop on Peer-to-Peer Systems*, Tampa Bay, USA (2008)

[Hashimoto 15] Hashimoto, Y.: Growth fluctuation in preferential attachment dynamics, *arXiv:1509.05590v2* (2015)

[Simkin 11] Simkin, M. V. and Roychowdhury, V. P.: Re-inventing Willis, *Physics Reports*, Vol. 502, pp. 1–35 (2011)

[Simon 55] Simon, H. A.: On a class of skew distribution functions, *Biometrika*, Vol. 42, No. 3–4, pp. 425–440 (1955)