

ソーシャルタギングにおけるタグ出現数のゆらぎ

Generation and Selection Mechanism on Social Tagging Systems

西川仁将^{*1}
Yoshimasa Nishikawa岡瑞起^{*1}
Mizuki Oka橋本康弘^{*2}
Yasuhiro Hashimoto池上高志^{*2}
Takashi Ikegami^{*1}筑波大学大学院システム情報工学研究科
Department of Computer Science, University of Tsukuba^{*2}東京大学大学院総合文化研究科
Graduate School of Arts and Sciences, The University of Tokyo

1. はじめに

近年, Delicious や Flickr, Twitter, などのソーシャルメディアサービスが普及している. これらのサービスではユーザーがタグと呼ばれる任意の文字列を付与することによってコンテンツの管理を行うソーシャルタギングシステムが採用されている.

ソーシャルタギングシステムを対象とした研究は様々に行われている. たとえば Golder らは「Polya urn (壺モデル)」に基づいてソーシャルタギングシステムを理解しようとしている [1]. 壺モデルでは過去に使われた回数の多いタグほど選ばれやすいという優先的選択性が働いている. また, Catutto らは既存のタグの選択に優先的選択性が働くことだけでなく, 新しいタグの増加傾向がベキ則に従うことを示した [2]. このように新しいタグがどのように生み出されるか, また, それぞれのタグがどのように使用されていくかはソーシャルタギングシステムを理解する上で重要な問いである.

これらの問いを 1 つの数理モデルとして記述できるモデルとして Yule-Simon 過程がある [3][4]. Yule-Simon 過程は古典的な数理モデルの一つであり, 新しいタグの増える割合を一定と仮定して, そうでなければ既存のタグを選択する, というルールに従う. Yule-Simon 過程は実際のソーシャルタギングシステムにおける全体的な振る舞いに関してはうまく記述することが可能である.

一方, 個々のタグの累積出現回数について注目してみると Yule-Simon 過程が予測する値から逸脱するタグが存在して, Yule-Simon 過程ではその逸脱の分布が指数関数的減衰を示す [5]. しかし, 実際のデータからは Yule-Simon 過程では捉えきれない大きさの逸脱をもつタグが多く存在することがわかっている. そこで本研究では実データにおける Yule-Simon 過程からの逸脱の大きさを確認し, Yule-Simon 過程を拡張したモデルを提案し比較する.

2. Yule-Simon 過程

本研究ではソーシャルタギングの新規生成と既存タグの選択のモデルとして Yule-Simon 過程を用いる. Yule-Simon 過程は, 出現回数に比例した確率でランダムに既存の単語が選択される優先的選択性をもつ. Yule-Simon 過程では, 新しいタグを追加するごとに, 新たな種類の単語を新規タグ生成確率 α で生成し, これまでに使われた単語の中から確率 $1 - \alpha$ で選択する. ここで t は時刻である. 時刻 t におけるアノテーション数を $N(t)$, タグ i が出現している回数を $n_i(t)$ とすると, 時

刻 t にタグ i が選択される確率 $P_i(t)$ は,

$$P_i(t) = \frac{(1 - \alpha)n_i(t)}{N(t)} \quad (1)$$

で与えられる.

Yule-Simon 過程では以下のような 2 つの性質を過程している.

- 新規タグ生成確率は一定でランダムに新規タグが生成される
- 出現回数に比例した確率でランダムに既存タグが選択される (優先的選択性)

また, Yule-Simon 過程では, 定義からタグの種類の増加はアノテーション数のベキに比例 (Heaps 則) し, タグ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.

3. Yule-Simon 過程からの逸脱

実データにおける個々のタグの出現回数が Yule-Simon 過程における予測値からどの程度逸脱するかを計測する. ここでは実データとして写真共有サービスである RoomClip^{*1}を用いる. 図 1 は RoomClip におけるタグの累積出現数の増加曲線の一部で, 実線は実データで破線が Yule-Simon 過程が予測する出現数である. たとえば「ラウンジ」というタグではモデルの予測値と実データの振る舞いがほぼ一致している. 一方, 「地球儀」ではモデルの予測値に比べて増加が遅い. 逆に, 「男前」では予測値を大きく上回っている. この実測値とモデルの予測値の間のズレの大きさを x とするとタグ i が初めて使われた時刻 t_i から時刻 s_i までの累積出現数の予測値と実測値の比によって次のように記述できる.

$$x = \frac{n_i(s_i)}{n_i^*(s_i)} \quad (2)$$

ただし, $n_i^*(s_i)$ は時刻 s_i における予測値で次のように定義される.

$$n_i^*(s_i) = \left(\frac{s_i}{t_i}\right)^{(1-\alpha)} \quad (3)$$

タグの成長が Yule-Simon 過程に従う場合, x の確率密度分布は, 指数関数的減衰を示すことが示されている. たとえば予

連絡先: yoshimasa@websci.cs.tsukuba.ac.jp

^{*1} <http://roomclip.jp/>. Tunnel 株式会社が運営するサービス. 日本最大の部屋のインテリア実例共有サイトである.

測値と実測値の比は、おおよそ 10 倍程度に収まる。しかし、実データにおける逸脱の大きさ x の分布を測定すると、Yule-Simon 過程が予測する成長よりも 100 倍、1000 倍の大きさになるタグが多く存在することが分析からわかっている (図 2)

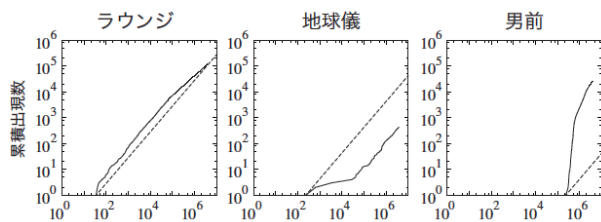


図 1: 個々のタグにおける累積出現数の増加

4. 提案モデル

図 1 に示したような予測値からの大きなズレを再現するために Yule-Simon 過程のタグの選択確率について修正を加える。実データでは Yule-Simon 過程と同様に Zipf 則が成り立っており、モデルの修正もこの性質を維持したまま行いたい。そこで Yule-Simon 過程のタグの選択確率を以下のように変形する。

$$\begin{aligned} P_i(t) &= (1 - \alpha) \frac{n_i(t)}{N(t)} \\ &= (1 - \alpha) \frac{n_i(t)k_i(t)}{N(t)} \frac{1}{k_i(t)} \end{aligned} \quad (4)$$

ここで出現回数が同じタグを Class と呼ぶこととする。Class 1 とは累積出現回数が 1 回のタグで、クラス N とは累積出現回数が N 回のタグのことを指す。 $k_i(t)$ とは時刻 t にタグ i が属する Class のタグ総数を表す。(4) 式右辺第 2 項はどの Class が選択されるかを表している、第 3 項は Class の中でどのタグが選ばれるかを表している。Zipf 則が保たれるにはクラスを選択確率の形が必要条件であり、クラスの中でのタグの選択確率の形は問わない。そこで本研究では式の後半にバイアスを与えた First-In, Last-Out (FILO) モデルを提案する。これはある Class の中で最初に Class アップしたタグが必ず最後に選ばれるというモデルである。これによって最近使われたタグが優先的に選択され、より多く使われるタグが出現することが考えられる。

5. 結果・考察

まず、FILO モデルにおけるゆらぎの大きさ x の分布を計算し、カルバック・ライブラー情報量 (Kullback-Leibler Divergence, KLD) によって実データの分布の類似度を計測した。 x の分布を図 2 に示す。

図 2 をみると FILO モデルはオリジナルの Yule-Simon 過程に比べて実測値により近い振る舞いを再現できていることがわかる。また、KLD による実データとモデルの分布の類似度は Yule-Simon が 0.833、FILO が 0.00390 という結果になった。これらの結果から提案する FILO モデルは Class 内のタグにバイアスをもった選択を与えることによって、大きな逸脱を生じ得ることを示している。

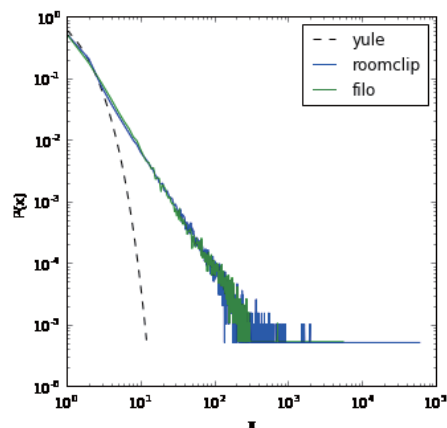


図 2: 各モデルにおけるゆらぎ x の分布

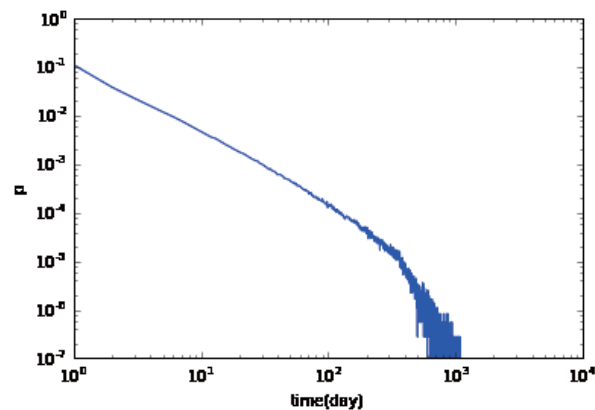


図 3: タグの再使用時間の分布

6. まとめ・課題

本研究では、Yule-Simon 過程では捉えきれない、出現数の増加曲線に大きなゆらぎをもつタグが実データに存在することを示した。また、Yule-Simon 過程の Class 内の選択にバイアスを加えることによって、タグの出現回数に大きな逸脱を生む効果を導入できることを示した。しかし、実データでは FILO モデルでも捉えきれない大きなゆらぎをもったタグが存在することが図 2 からわかった。

本研究で提案した FILO モデルでは、クラス内で最近使われたタグが必ず選ばれる、という 1 つの極端な例を考えた。このようなタグの「鮮度」に依存した使われ方は、図 3 に示すインターバルの分布を見ることでも明らかである。今後はこのような人間の注意の時間的な変化にも着目してモデル化を試みたい。

謝辞

解析に用いた貴重なデータを提供いただいた、Tunnel 株式会社の皆様に感謝いたします。また、本研究は JSPS 科研費 15K00420 の助成を受けたものです。

参考文献

- [1] Golder, S. A. and Huberman, B. A. : Usage patterns of collaborative tagging systems, *Journal of Information Science*, Vol. 32(2), pp. 198-208 (2006)
- [2] Cattuto, C., Loreto, V., and Pietronero, L.: Semiotic dynamics and collaborative tagging, *Proceedings of the National Academy of Sciences*, Vol. 104(5), pp. 1461-1464 (2007)
- [3] Yule, G. U.: A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS, *Philosophical Transactions of the Royal Society of London*, Vol. 213, pp. 21-87 (1925)
- [4] Simon, H. A.: On a Class of Skew Distribution Functions, *Biometrika*, Vol. 42(3/4), pp. 1-35(1955)
- [5] Hashimoto, Y.: Growth fluctuation in preferential attachment dynamics, *Phys. Rev. E*, Vol. 93, p. 042130 (2016)