Vol.2017-HCI-171 No.14
2017/1/23

SNS にみる新規語彙の生成・選択・揺らぎの現象

岡瑞起[†1]　西川仁将[†2]　佐藤晃矢[†2]　橋本康弘[†1]　池上高志[†3]

概要：進化生物学やその数学的枠組みを用いると、言語やソーシャルタギングシステム（Social Tagging System, STS）などの Web サービスに、人と人工システムとの相互作用に関する新しい様相が見えてくる。進化的解析の対象は、サービス開始とともに増加するユーザの数、どのようにタグがつけられ、どういうタグが共起するか、どのようなタグが進化するか、といった点である。加えて、サービスそのものの構造も時に進化する。本研究では、こうしたタグの生成と選択メカニズムについての分析を報告し、オリジナルな Yule Simon では捉えられなかった個々のタグ使用頻度の揺らぎを再現する新しいモデルについて後半で議論する。

キーワード：ソーシャルタギング，SNS, Yule-Simon 過程, 進化生物学, 優先的選択

Generation and Selection Mechanism on Social Tagging Systems

MIZUKI OKA[†1]　YOSHIMASA NISHIKAWA[†2]　KOYA SATO[†3]
YASUHIRO HASHIMOTO[†1]　TAKASHI IKEGAMI[†3]

Abstract:
Evolutionary biology and its mathematical framework are beginning to be widely used in evolutionary analysis of non-biological systems such as languages and web services adopting social tagging system (STS). In general, as the service starts, the number of users using the service increases, and the structure of the service itself evolves at time. As the number of users increases, the aspect of how tags are tagged, what tags co-occur, and what kind of tags evolve is very STS as it can be analyzed like biological evolution Interesting. In this research, we report on generation of tags and analysis on selection mechanism.

Keywords: Social Tagging, SNS, Yule-Simon Process, Evolutionary Biology, Preferential Attachment

1. はじめに

インターネットのソーシャル化による、個人の行動や発言が大量に記録が蓄積されている。こうした大量の記録を使って、これまで定性的にしか議論できなかった人間行動や社会現象を定量的に理解しようとする「計算社会科学」や「ウェブサイエンス」といった分野が国内外で注目され始めている。特に、著者らは、インターネットを使ったウェブサービスの作り出す疑似生命的な進化に注目し、自然科学と情報科学の交点としての分野形成を目指した研究活動を進めている [1]。

ウェブサービスでは通常、そのサービス固有のコンテンツ（テキスト、写真、音声、動画、タグなど）に、時間とともにバリエーションが生まれ、選択され淘汰されていく。それは生命システムの進化発展の特徴でもある。実際、進化生物学やその数学的枠組みは、言語やソーシャルタギングシステム（Social Tagging System, STS）において徐々に用いられ始めている [2], [3], [4]。STS とはオンラインコンテンツ共有サービスにおいてユーザが任意の文字列（e.g., タグ）を付与することでコンテンツの管理を行うシステムのことであり、Flickr, Twitter, Instagram, Facebook などがこれにあたる。一般に、サービス開始とともにユーザ数は増え、サービスそのものの構造も、同時に進化する。ユーザ数の増加とともに、新たなタグはどのように、あるいはどの程度生み出され（i.e., 突然変異）、どのように使われているのか（i.e., 淘汰）、といった観点はソーシャルタギングがあたかも生物進化のように解析できるという点で非常に興味深い。

オンラインにサービスにおけるタグの振る舞いを記述する最初のモデルは Golder と Huberman により提案された「Polya urn」モデルである [5]。彼らは、これまでに使われた回数の多いタグほどより選ばれやすいという「優先的選択性」が働いていることを示した。Catutto らは、既存タグの選択が優先的選択が働くことだけでなく、新しいタグの増加はベキ分布に従うことを示し、Yule-Simon 過程がこ

†1 筑波大学システム情報系
Faculty of Engineering, Information and Systems, University of Tsukuba
†2 筑波大学
University of Tsukuba
†3 東京大学大学院総合文化研究科
Graduate School of Arts and Sciences, College of Arts and Sciences, The University of Tokyo

れら両方の性質を満たすことを示した [2]。Yule-Simon 過程は、もともと生物の属に含まれる種の数がベキ分布になる、という性質を説明する古典的なモデルである [6] [7]。Yule-Simon 過程は、新しい種類のタグの増え方を一定（α）、あるいは時間減少する確率（α^{-t}）と仮定し、確率（$1-\alpha$）で既存のタグから選択する。ソーシャルタギングにおけるタグの振る舞いは Yule-Simon 過程で現象論を上手く記述できることが報告されている [7]。しかし、新たなタグが生まれるメカニズムや既存のタグが選択されるメカニズムは、きちんと議論されていない。そこで本稿では、ソーシャルタギングシステムのデータ分析と通して、Yule-Simon 過程からのズレに注目し、タグ生成と選択のメカニズムを探る。

2. ソーシャルタギングシステム

2.1 分析データ

３つのソーシャルタギングシステム（Delicious[], Flickr[], RoomClip）を分析する。Delicious はブックマークを共有するサービスである。共有されたブックマークページに、複数のユーザがタグ付けを行うことができる。Flickr と RoomClip は写真を共有するサービスである。写真を投稿したユーザのみが写真にタグを付けることができる。各データの語彙数、アノテーション総数、分析対象時期を表１に示す。

表 1　ソーシャルタギングデータ
Table 1　Three social tagging data sets in number.

	語彙数	アノテーション総数	時期
Delicious	2,480,996	140,116,358	2003-01 ～ 2006-11
Flickr	1,607,879	112,900,000	2004-01 ～ 2005-12
RoomClip	194,890	3,776,535	2012-04 ～ 2015-05

2.2 タグの生成・選択と時間発展

タグはユーザによって新しく生成され、使用（選択）され、淘汰されていく。「男前」タグの時間発展をみると、生成からジャンルとして定着するまでの様子良く分かる（図１）。このような時間的発展のモデルとして、Yule-Simon 過程を採用し、モデルの予測と実データを比較する。モデルからのズレを詳細に分析することで生成と選択のメカニズムに言及する。

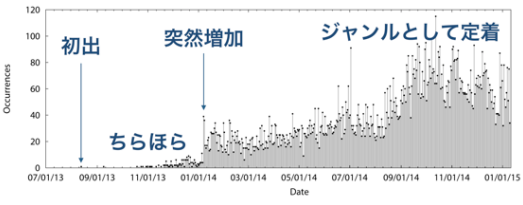

図 1　「男前」タグ使用の時間発展
Figure 1　Population Dynamics of “Otokomae” tag.

3. Yule-Simon 過程と実データ

3.1 モデル

Yule-Simon 過程は試行ごとに新たな種類のタグを確率αで生成、またはこれまでに使われたタグの中から確率$1-\alpha$で選択する。時刻tでのアノテーション数をN、タグiが出現している回数を $n_i(t)$ とすると、$N+1$回目の試行でタグiを選択する確率$P(i)$は、

$$P(i) = \frac{(1-\alpha)n_i(t)}{N(t)}$$

で与えられる。出現回数（k）が多いタグほど選ばれる確率が高くなる優先的選択性を持つ。Yule-Simon 過程は実データをどの程度、記述できるかを以下でみていく。

3.2 語彙生成確率と既存語彙の選択

まず、実データのタグ生成確率をみる。図 2 は横軸に累積アノテーション数、縦軸に語彙生成確率／日のプロットである。Delicious、Flickr、RoomClip 全てのデータにおいて一定確率、あるいは時間減衰する確率でタグが生成されていることが分かる。

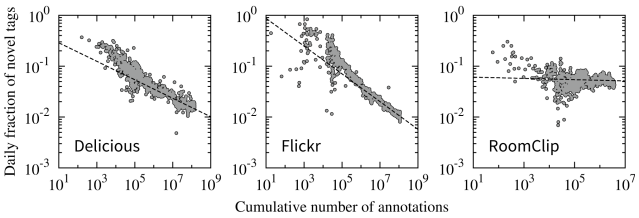

図 2　語彙生成確率分布
（左）Delicious・（右）Flickr・（右）RoomClip
Figure 2　Tag Creation Rate on Delicious, Flickr, RoomClip.

次に、実データの既存語彙の選択確率をみる。図 3 は、過去に使われた回数（横軸）とその選択確率（縦軸）の分布である。出現回数が多いほど、選ばれる確率が高くなる優先的選択が働いていることが分かる。

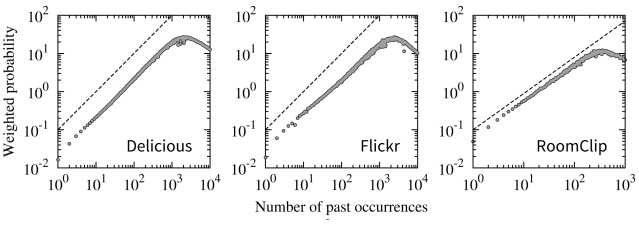

図 3　既存語彙選択確率分布
（左）Delicious・（右）Flickr・（右）RoomClip
Figure 3　Distribution of Number of past occurrences.

3.3 Heaps 則と Zipfs 則

生成確率と選択確率から、Yule-Simon 過程の仮定は実データと一致することが分かった。Yule-Simon 過程の仮定は実データと一致することが分かった。そこで、実際に語彙数の増加を分析し、語彙生成確率の実測値と Yule-Simon の予測値との一致度合いを調べて見る。総アノテーション数を t、語彙総数を$V(t)$とすると、Yule-Simon 過程の語彙数の増加は、$V(t) \propto t^{\beta}$ の関係が成り立つことが知られている。これを Heaps 則と呼ぶ。Yule-Simon 過程において、語彙生成率αが時間減衰する場合 $\alpha(t) \propto t^{-\gamma}$, $\beta = 1 - \gamma$ の関係が成り立つ []。図 4 に、総アノテーション数（横軸）と語彙総数（横軸）として実測値と、理論値（実データで得られた値をαに設定）を示す。実データ、Yule-Simon 過程どちらも Heaps 則に従い、実測値と理論値のベキ指数 β がよく一致する結果が得られた1。実データのβ, γを表 2 に示す。

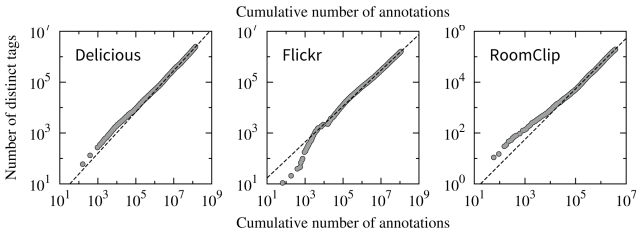

図 5　語彙数の増加：実測値（実線）と理論値（点線）
（左）Delicious・（右）Flickr・（右）RoomClip
Figure 5　Heaps Law.

表 2　実データにおける β, γ
Table 2　Empirical Results of β, γ.

	β	γ
Delicious	0.812±0.110%	0.238±1.50%
Flickr	0.704±0.066%	0.370±1.23%
RoomClip	0.993±0.120%	0.012±44.1%

次に、ランク出現数分布を考察する。語彙出現回数をk_r、ランクをrとすると、Yule-Simon 過程のランク出現分布は、$k_r \propto r_{\lambda}$ の関係が成り立つことが知られている。これを Zipfs 則と呼ぶ。Yule-Simon 過程において、αが一定の場合、$\lambda = 1 - \alpha$、時間減衰する場合$\frac{1}{\beta - t^{-\gamma}}$となる。図 5 に、ランク（横軸）と出現回数（縦軸）として、実測値と理論値を示す。実データ、Yule-Simon 過程どちらも Zipf 則に従い、実測値と理論値のベキ指数λが一致する結果が得られた。Heaps 則のベキ指数βと Zipf 則のベキ指数λの関係は、逆数の関係になっていることが知られており、3 つのデータセットについても逆数の関係となり、そのことが証明される2。

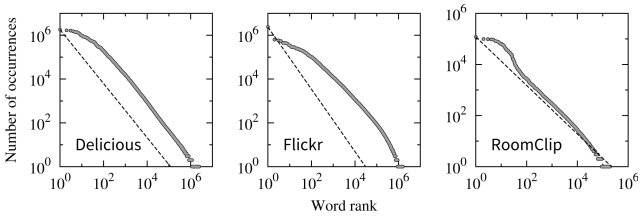

図 5　ランク出現数分布 : 実測値（実線）と理論値（点線）
（左）Delicious・（右）Flickr・（右）RoomClip
Figure 5　Zipfs' Law.

このように、Heaps 則と Zipf 則の分析から、Yule-Simon 過程は語彙の生成と選択のマクロな現象を上手く記述する。そこで、次章以降で、個々のタグの振る舞いはどの程度モデルに従うのか、というミクロな視点による分析結果をで述べる。

4. 語彙生成・選択のミクロな分析

4.1 語彙生成確率

これまでみてきた通り、Yule-Simon 過程での語彙生成確率はランダムに一定αあるいは時間減衰$\alpha(t) \propto t^{-\gamma}$ すると仮定してきた。しかし、実際のサービスではコンテンツ（ウェブページや写真）には、同時に複数のタグが使用される。これら複数のタグが使われ方には相関があるとしたら、語彙生成率 α はランダムではなく、各コンテンツpに付けられるタグ数w_pとの相関が期待される。

そこで、タグ間の相関を取り除くために、データをランダムにシャッフし（ランダムデータ）し、オリジナルデータと比較する。図 6 に、タグ数w_p（横軸）と語彙生成率α（縦軸）の相関関係を示す。この相関は、各サービスでユーザがどのようにタグセットを選ぶかを表す。正の相関は、使うタグ数を増やすとき、タグを新たに生成することを示す。一方、負の相関は、タグ数を増やすとき、既存のタグから

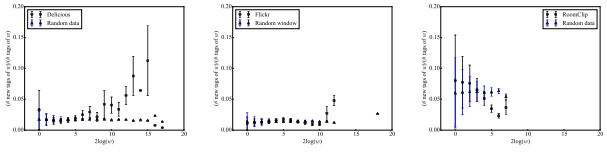

図 6　タグ数と語彙生成率の相関：オリジナルデータ（黒線）とランダムデータ（青線）
（左）Delicious・（右）Flickr・（右）RoomClip
Figure 6　Correlations between w_p and α. The black plot shows original data, and the blue plot shows random data.

選んで使うことを示している。Delicious は正の相関、

1 Yule-Simon モデルの確率密度分布のベキ指数は$\gamma = -\left(1 + \frac{1}{1-\alpha}\right)$となる。これを累積確率分布$P(X < x)$にするには積分するので指数に 1 を足して$\frac{-1}{-(1-\alpha)}$、さらに Zipf 則にするには縦横軸を入れ替えるので反転して$-(1-\alpha)$となる。これが Yule-Simon 過程から導かれる（期待される）Zipf 則の指数となる。

2 Heaps 則と Zipfs 則のベキ指数の関係に関する詳しい解説は、https://dencity.jp/misc/yule-simon_process/を要参照。

RoomClip は負の相関を示す。 Flickr はほとんど相関が無く、タグ数が少ない多いに関わらず語彙生成率は一定である。

このように語彙生成率は、ミクロにみるとランダムに一定ではなく、コンテンツに付けられるタグ数によって変化する。変化の傾向はサービスごとに異なり、サービスの特徴に起因するユーザのタグ付けの動機からある程度説明される［Koya submitted］。ウェブページを共有する Delicious では、ユーザは可能な限り重複しないタグを選んでいる（Categorizer と呼ぶ）。例えば、そうすることで、容易に目的のウェブページを検索できるように工夫していると考えられる。一方、ユーザ間で写真を共有するためのサービス RoomClip では、ユーザはなるべく共通するタグを選んでいる（Describer と呼ぶ）。これにより、より多くのユーザに写真をみてもらうことが可能となる。

4.2 既存語彙選択

次に、個々のタグの振る舞いを既存語彙選択という観点から詳細に分析する。図 7 は、個々のタグの使用回数の変化を示している。実線は実測値、点線は Yule-Simon 過程が予測する成長線である。例えば、「ラウンジ」「観葉植物」「無印良品」は、理論値とほぼ同様に成長している。一方、「PC デスク周り」「地球儀」は、理論値よりも大きく減少する。反対に、「男前」は理論値を大きく超えて使用されている。

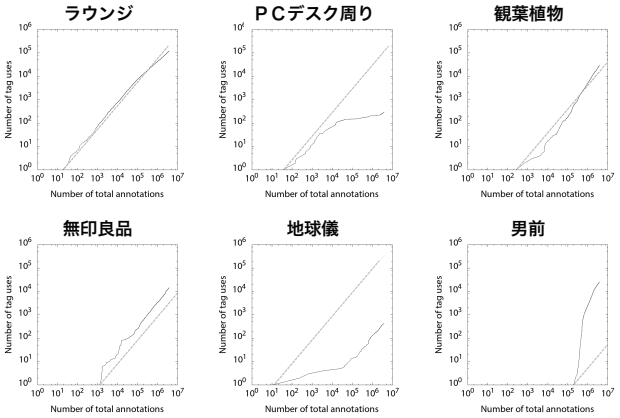

図 7　個々のタグの振る舞い

Figure 7　Growth of Individual Tag.

揺らぎの大きさは、Yule-Simon 過程が予測する値と、実測値から定量的に定義できる。図 8 にその概念図を示す。揺らぎの大きさ x は、タグが初めて使われてから時間 s_i 経過後における累積使用回数の理論値と実測値の比によって定義される。Yule-Simon 過程に従う x の確率分布は、指数関数的減衰を示し、ベキ成長からの上や下にズレる大きさが最大でも 10 倍程度に収まることが解析的に示されている [4]。

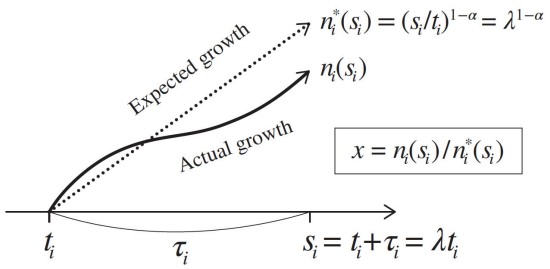

図 8　揺らぎの定義

Figure 8　Definition of deviation scale.

一方、実測値の揺らぎ分布をみると、図 9 に示すようにベキ分布となる。つまり、理論値が示すベキ成長からの上や下へのズレの大きさが、100 倍や 1000 倍、時には 10,000 倍、100,000 倍になることを示している。タグによって、何らかの大きなバイアスがかかり、大きくする成長あるいは衰退する場合があることを示唆している。

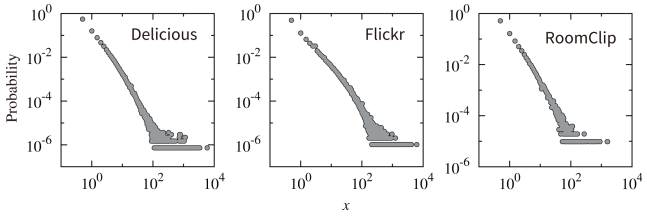

図 9　実測値の揺らぎ分布

（左）Delicious・（右）Flickr・（右）RoomClip

Figure 10　Probability distribution of the deviation scale in vocabulary growth.

5. Yule-Simon 過程の拡張

どのようなバイアスが働くと図 9 に示したような理論値からの大きなズレを生み出すのか。それを探るために、Yule-Simon 過程に修正を加えた FILO（First-In-Last-Out）モデルを考える。

5.1 Class

FILO モデルを説明するために Yule-Simon 過程に「Class」という概念を導入する。ここでは、出現回数を Class と呼ぶ。例えば、ABACDACDEB というアノテーションを考える。このとき、「Class 1」は 1 回だけ出現したタグの集団、「Class 2」は 2 回出現したタグの集団、「Class 3」は 3 回出現したタグの集団となる。この例では、A は Class 3、B・C・D は Class 2、E は Class 1 に属する。そこで、時刻 t でのタグ i が属するクラスの語彙総数を $k(i,t)$ とし、Yule-Simon 過程での、タグ i の出現する確率を次のように分解する。

$$P(i,t) = \frac{n_i(t)}{N(t)} = \frac{n_i(t)k(i,t)}{N(t)} \frac{n_i(t)}{n_i(t)k(i,t)}$$

ここで、式の後半は、「Class の中でどのタグが選ばれるか」を表している。分子が$n_i(t)$、分母が$n_i(t)k(i,t)$なので、同じ Class の中ではどのタグも同じ確率で選択される。前述の例では、Class2 に所属する B・C・D 全て$2/6$の確率で選ばれることになる。Simon は、Class の中でどのタグを選ばれるかを与える式の後半は、自由に設定してよい。つまり、式の前半をいじらなければ、Zipf 則を満足することを示している [7]。

5.2 FILO モデル

提案する FILO モデルは、ある Class の中では最後に Class アップしたタグが必ず選ばれるモデルである。つまり、First-In したタグ（古いタグ）は、Last-Out（なかなか選ばれない）する。例えば、ABBCD とアノテーションが行われているときを考える。このとき、Class1 には ［A・C・D］が所属し、Cass2 には［B］のみが所属している。次の既存タグで、Class 1 と Class 2 が選ばれる確率（式の前半）は、それぞれ$3/5$と$2/5$である。Class 選択で Class 1 が選ばれた場合は D が選択され、Class 2 が選ばれた場合は B が選択されることになる。

このようなバイアスを与えたモデルを用いて揺らぎの分布をみると図 10 のようになる。黒点は FILO モデル、灰色線は実測値を示す。Simon-Yule 過程が予測する指数関数的減衰よりもテールが伸び、実測値に近い分布となった。

6. まとめ

本稿では、ソーシャルタギングは Yule-Simon 過程で比較的よく説明できることを、（１）語彙生成レートは一定確率、あるいは時間のベキで減衰する確率に従う、（２）語彙選択は優先的選択に従う、という点から示した。しかし、個々のコンテンツでの新規タグ生成レートは、必ずしも一定でないこともみえた。さらに、Yule-Simon 過程のオリジナルな枠組みでは説明できない、成長レートに大きな揺らぎを持つ語彙が存在することを示した。

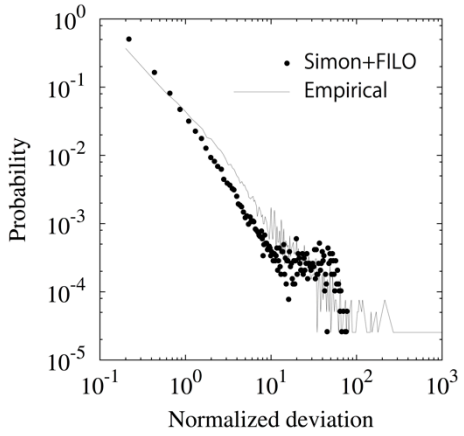

図 10　FILO モデルによる揺らぎの分布（黒線）と実測値（灰色線）

Figure 11　Probability distribution of the deviation scale in vocabulary growth with FILO model.

我々は、Class の概念を持ち込み、タグ使用の大きな揺らぎを生む Class 内の選択タグ確率をいじることで、極端な振る舞いを捉える可能性を示した。今後、さまざまなバイアスを与えることで、実データを更によく説明できるモデルの提案を行っていきたい。この改良したモデルをもう一度、ユーザの内的観点からモデル化し、実際のシステムデザインに応用できるように考えたい。

謝辞　本研究は JSPS 科研費 JP15K00420 の助成を受けたものです。

参考文献

[1] 人工知能学会 SIG-WebSci. (2015, July) SIG-WebSci. [Online]. http://sigwebsci.tumblr.com/

[2] C. Cattuto, V. Loreto, and L. Pietronero, "Semiotic dynamics and collaborative tagging," *Proceedings of the National Academy of Sciences*, vol. 104, no. 5, pp. 1461-1464, 2007.

[3] 晃矢 佐藤, 瑞起 岡, 康弘 橋本, and 和彦 加藤, "Yule-Simon 過程によるタグ共起ダイナミクスのモデル化と分析," *人工知能学会論文誌*, vol. 30, no. 5, pp. 667-674, 2015.

[4] Yasuhiro Hashimoto, "Growth fluctuation in preferential attachment dynamics," *Phys. Rev. E*, vol. 93, p. 042130, 2016.

[5] S. A. Golder and B. A. Huberman, "Usage patterns of collaborative tagging systems," *Journal of Information Science*, vol. 32, no. 2, pp. 198-208.

[6] G. U. Yule, "A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS," *Philosophical Transactions of the Royal Society of London*, pp. 21-87, 1925.

[7] H. A. Simon, "On a Class of Skew Distribution Functions," *Biometrika*, vol. 42, no. 3/4, pp. 425-440, 1955.