1H3-2

新しいタグの出現とソーシャルタギングシステム

Appearance of new tags and Social Tagging System

佐藤晃矢 *1 Koya SATO　岡瑞起 *1 Mizuki OKA　橋本康弘 *1 Yasuhiro HASHIMOTO　池上高志 *2 Takashi IKEGAMI　加藤和彦 *1 Kazuhiko KATO

*1筑波大学システム情報系 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

Why people make a new kind of tag? This is a very big question on social tagging system. In this paper, we consider the mechanism of new tag creation from the micro aspect like the simultaneous use of tags. We discover that user's tagging motivations affect the micro aspect of the new tag creation mechanism.

1. はじめに

Delicious，Flickr，Instagram，Twitter，Facebook などのオンラインコンテンツ共有サービスでは，ユーザーが任意の文字列（e.g., タグ）を付与することによって投稿されるコンテンツの管理を行う Social Tagging System（STS）が採用されている [Gupta 10, Golder 06]．新たなタグはどのように，あるいはどの程度生み出され，どのように使われているのかといったタグの振る舞いを知ることは，効率的なデータベース設計や情報ナビゲーションを実現するために大切な課題であり，これまでにいくつかの研究が行われている [Gupta 10, Strohmaier 12]，オンラインサービスにおけるタグの振る舞いをモデル化する最初の試みは，Golder と Huberman による「Polya urn」モデルである [Golder 06]．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」があることを実データから発見し，それを説明するモデルを提案した．また，Catutto らはソーシャルネットワークサイトにおけるタグ使用頻度の順位と頻度数，さらには，新しい種類のタグの増加する傾向がベキ分布に従うことを示した [Cattuto 07b, Cattuto 07a]．彼らは，壺モデルの代わりに，新しい語彙が生み出される効果を取り入れた Yule-Simon 過程を導入することで説明した．

Yule-Simon 過程とは，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ために，Yule により使用された [Willis 22, Yule 25]．その後 Simon により生物以外のシステムに観測できるベキ分布を説明することにも応用できることが示された古典的なモデルである [Simon 55]，Yule-Simon 過程が説明するのは新しい種類のタグの増え方を一定と仮定し (α)，確率 ($1-\alpha$) でこれまでに使われたタグから選択するという，既存のタグの選び方である，Social Tagging におけるタグの振る舞いは Yule-Simon 過程により現象論をうまく記述することが可能である．然しながら，Yule-Simon 過程の枠組みでは，新たなタグの発生確率が一定の確率，あるいは時間減少する確率により記述されていることから，突然変異のようなランダムな振る舞いにより一定の確率で引き起こされると考えられており，新たなタグが生み出される機構論的なメカニズムに言及しているとは言えない，

そこで，本研究では新たなタグの生成確率を決定するメカニズムを明らかにすることを目的に，ミクロなタグのシーケンスである，タグの同時利用に注目し分析を行い，Yule-Simon の枠組みとのズレを議論する．その際に各サービスでは独特の傾向が現れることが考えられる．それはミクロな視点ではユーザーのタグをつける動機が効果を及ぼすためである．そこで，ユーザーのタグをつける動機に注目し，それぞれのサービスにおけるユーザーの傾向がどのように新規タグの生み出され方に影響を及ぼすのかを確かめる．

表 1: 各サービスの基本統計

Service	Users	Tags	Tag assign	Posts
Delicious	532,924	2,481,108	140,126,555	47,257,452
Flickr	319,686	1,607,879	112,900,000	28,153,045
Instagram	2,110	271,490	8,201,542	1,047,774
RoomClip	32,852	194,881	3,141,524	692,459

2. データ

本研究では STS を採用している代表的な SNS である Delicious，Flickr，Instagram，RoomClip のタグつけデータを利用した．タグ付けの対象となるリソースは Delicious ではウェブページ，Flickr，Instagram，RoomClip では写真である．後者の 3 つのサービスは写真共有という点では共通しているが，Flickr に比べて，Instagram と RoomClip は写真のアーカイブという役割以上に，コミュニケーション手段の一つとして写真を使うという SNS 的な側面が強い，また，Delicious は個人的な情報管理という側面が強い [Golder 06]．

今回扱う 4 つのデータはそれぞれ異なる方法で取得した．Delicious と Flikcr のデータは PINTS の EXPERIMENTS DATA SETS を利用した [Görlitz]．クローラーにより取得された結果のデータセットである．Instagram のデータは Emilio らが収集したデータを利用する [Ferrara 14]．これらのデータは以下のように集められた：毎週金曜日にイベントが開かれているのだが，その参加者の中から無作為に 2000 人ほど選び，それぞれのユーザーが投稿した記録を全て取得したデータとなっている．RoomClip のデータはサービスを運営する Tunnel 社から直接提供を受けた。これらのタギングデータは ”投稿された時間，ユーザー id，リソース id，タグ名 ”という形式で表現する．Delicious の場合はブロードタギングシステム [Wal 05] を採用しているため，個々のタグ付けの行為はユーザ ID とリソース iD の組で区別される．一方，ナロータギングシステムを採用する他の 3 つのサービスでは，リソース ID のみで区別される．これらのデータの基本統計をに示す．

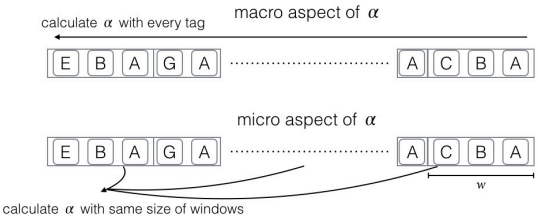

図 1: ミクロなタグ付けとマクロなタグつけの概念図.

2.1 ミクロ，マクロ

Yule-Simon 過程では，タグは 1 つずつシーケンシャルに発生し，一つの投稿に対してタグが複数使われるようなソーシャルタギングに典型的なケースは考慮されていない．そこで，本研究では新たなタグの生み出され方とタグの同時利用に関する分析を行う．そのために今回取得したタグ付けのシーケンスデータから，ユーザーとリソースを結びつけた複数のタグがつけられる投稿を以下のようにして定義する．$window_i$ = (user, resource, {tags}, time)．これにより，ある投稿 (i) に対してどのようなタグがつけられ，また同時利用の個数が幾つなのかがわかる．このようなデータのセットで表現される投稿一つ一つを以降はウィンドウと呼び，$window_i$ に対してつけられるタグの個数をウィンドウサイズ（w_i) と定義し，分析を行う．また，本研究におけるマクロは投稿された全てのタグのシーケンスであり，ミクロはウィンドウサイズごとに集計をとった値である．その概念図を図 1 に示す.

3. 解析手法

3.1 ユーザーのタグをつける動機．記述者，分類者

これまでの研究から，ユーザーがタグをつける動機はユーザーにより異なり，その分布も各サービスで異なることがわかっている [Strohmaier 12]．同時に付けられるタグの数はユーザーのタグをつける動機を理解するうえで重要な指標となる．人から検索されることを意識したサービスでのタグつけの場合には，その平均値は大きくなることが知られている．これは Overtagging と呼ばれ，様々な表現により情報を記述することで，検索にヒットしやすくすることを目的としている [Heckner 08]．つまり，w の平均値が大きくなるようなサービスが情報を共有しようとする振る舞いの強いサービスであり，逆に小さい値を撮る場合がそのような傾向の弱いサービスであると言える．

また，この概念をより正確に分類するために記述者，分類者と呼ばれる指標が提案されている．分類者は情報を主観的に分類することに価値を見出すユーザーであり，自分の個人的な情報管理を主な目的としてタグ付けを行う．また，記述者は情報を詳細かつ正確（客観的）に記述し，他者からの投稿を検索されることに価値を見出すユーザーであり，その後の検索を容易にすることを目的にタグ付けを行う．そのため，記述者的なタグ付けを行うユーザーのつけるタグの方が，サービス全体で合意がなされているタグである．

今回は，両者の 2 つのタグ付けのモチベーションの分布をサービスごとに見るために Strohmaier らが提案した手法を用いることにより分析を行う [Strohmaier 12]．Strohmaier らは記述者的な側面と分類者的な側面からユーザーのタグ付けに注目し，以下のような値を定義した.

$$M_{desc} = \frac{|\ \{t :|\ R(t)\ | \leq n\}\ |}{|\ T\ |}, n = \lceil \frac{|\ R(t_{max})\ |}{100} \rceil \quad (1)$$

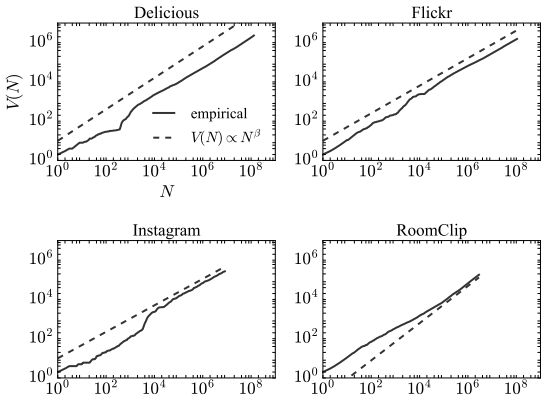

図 2: Heaps' law.

| T | はあるユーザーが利用したユニークなタグの数を表し，| $R(t)$ | はあるユーザーが使用したタグ t がつかわれたリソースの数を表す．t_{max} は最も多くのリソースに付けられたタグを指す．つまり，M_{desc} は記述者的な孤児のようなタグが増えるという側面に注目した値となっており，ユニークなタグを多く使う記述者的なユーザーの場合に大きな値をとる．

分類者的な側面に注目した値として，

$$M_{cat} = \frac{H(R|T) - H_{opt}(R|T)}{H_{opt}(R|T)} \quad (2)$$

という値がある．$H(R|T)$ は，あるタグである際に，あるリソースである場合の条件付き情報量であり，ユーザーがタグにより投稿を符号化する際のやり方を捉えた値となっている．また，正規化するために $H_{opt}(R|T)$ という値を利用している．これは対象とするユーザーの持つユニークなタグの数と，リソースの数，リソースに対して同時に付けられるタグの数の平均値を保持したまま，$H(R|T)$ の値が小さくなるよう，つまり，全てのタグが等しく情報を区別可能にした際の，理想的な分類者的な振る舞いを示している．

以上の 2 つの値の平均値をとり，

$$M_{combined} = \frac{M_{desc} + M_{cat}}{2} \quad (3)$$

という値により，記述者的振る舞いと分類者的振る舞いを説明する．$M_{combined}$ は M_{desc} と M_{cat} の平均値であり，この値が 0 に近いユーザーが分類者的なユーザーである．逆に大きな値を示すユーザーが記述者的なユーザーである．

w の平均値と $M_{combined}$ の値の分布を各サービスに対して計算することで，各サービスにおけるタグをつける動機の傾向を特徴付けることが可能となる．両者が小さい値を撮る場合には記述者的な情報を共有するためのタグつけであり，両者が大きな値をとる場合が分類者的な個人的な利用が主なタグつけであると言える．

4. 結果

4.1 Heaps' law

まずはじめに，すべてのサービスが Yule-Simon 過程によりモデル化できることを確かめる．ソーシャルタギングで使われる新しい語彙の数は Heaps 則 [Heaps 78] に従って増加することがよく知られている [Cattuto 07a]。Heaps 則の指数は新し

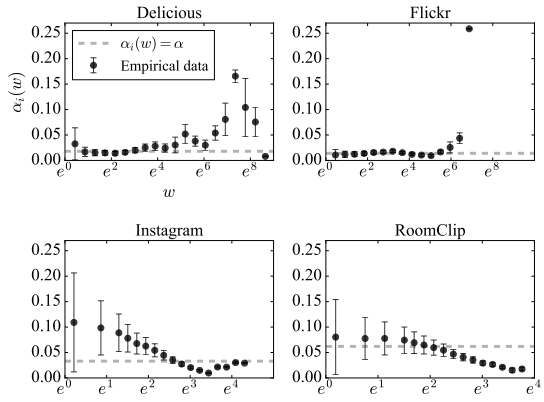

図 3: ウィンドウサイズと新規タグの生成確率の相関.

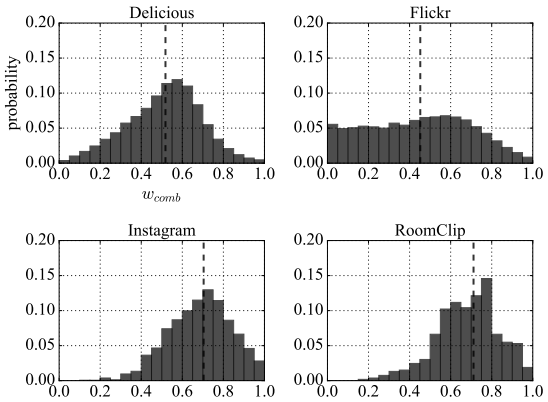

図 4: 各サービスにおける $M_{combined}$ の値の分布.

い語彙が出現するタイミングによって決定され、例えば新規語彙の生成レートが時間的に現象する場合、Heaps 則はサブリニア則を示す．図 2 に今回のデータセットに対する語彙の成長カーブを示す。いずれも指数 1 以下の Heaps 則に従っていることがわかる．

このようにサービス全体の平均でみると，α は時間減衰する関数，または一定となり，これまでのさまざまな研究で示される通り，各サービスでその増加則は同様のメカニズムに従っており，Yule-Simon 過程によりシステムの振る舞いをモデル化は妥当であることが分かる．

4.2 Relationship between α and w

各サービスにおける，あるウィンドウでのウィンドウサイズと新規タグの間の関係を図 3 に示す．横軸がウィンドウサイズ (w) である．横軸は $log(w_{max})/10$ の瓶で区切っている．w_max は各サービスでのウィンドウサイズの最大値である．これはウィンドウサイズの分布が非常に広い分布であり，べき分布を示すため，このようなウィンドウサイズごとに幅の変わる瓶サイズで区切った．縦軸はあるウィンドウサイズの瓶での新規タグを含む割合の平均値である．また，エラーバーは標準偏差である．

実データのウィンドウサイズと新規タグの割合の相関に注目すると，各サービスでは異なる傾向が現れていることがわかった．Delicious では正の相関が現れることがわかった．Instagram と RoomClip の場合には負の相関が現れる．また，Flickr には両者の関係は無相関であることがわかった．

4.3 ユーザーのタグをつける動機

各サービスにおける w の値をそれぞれみてみる．表**??**に，各サービスにおける w の平均値と中央値を示す．他者から検索されることを意識したタグ付けの場合には w の値が大きくなることが考えれ，自分自身の情報管理のためのタグ付けの場合には w の値は小さくなることが考えられる．各サービスにおける w の平均値は Deliciosu，Flickr，RoomClip, Instagram の順に大きくなっており，Delicious の w の値が最も小さかった．これは，Delicious が個人的な情報管理のために利用されていることを示し，人から検索されることを意識したタグ付けではないことが考えられる．

次に各サービスにおけるユーザーが記述者的な振る舞いを示すのか，分類者的な振る舞いを示すのか，その分布を確かめる．図に各サービスにおける 200 回以上ポストしたユーザーを対象にした際の，$M_{combined}$ の値の分布を示す．横軸が $M_{combine}$ の値であり，縦軸がユーザーの割合である．また，0-1 までの値を表示しており，0 に近い値が分類者的な振る舞いを表しており，大きな値が記述者的な振る舞いを表している．それぞれ 20 個の等間隔の瓶により $M_{combine}$ の値を区切っている．点線で表現される垂線が各サービスにおける $M_{combine}$ の値の平均値である．

表 2: Median and Mean of w on each services

Service	Median of w	Mean of w
Delicious	2	2.96517
Flickr	3	4.01022
Instagram	3	7.82759
RoomClip	4	4.53677

図を見ると，4 つのサービスでは Instagram と RoomClip の分布の中心は記述者側に存在することがわかり，Instagram と RoomClip を利用するユーザーは記述者的な振る舞いを示すユーザーが多いということがわかった．Delicious と，Flickr では先の 2 つのサービスと比べると分類者側に分布の中心が存在しているということがわかった．また，Flickr ではその分布の形が他の 3 つのサービスと異なり，記述者的なユーザーの分布と分類者的なユーザーの分布の混合分布のような形となり，記述者と分類者の混在しているサービスであることがわかった．

以上のことから，Delicious は分類者的な個人の情報管理のためのタグ付けを強調するようなシステムであり，Instagram と RoomClip は記述者的な他者からの検索されるためのタグ付けを強調するようなシステムであり，Flickr は記述者的な振る舞いと分類者的な振る舞いも受け入れる，3 つのサービスの中間に位置するサービスであると言える．

5. 考察と議論

本研究ではマクロな視点では Yule-Simon 過程により新規タグの生成メカニズムを記述することが可能な 4 つのそれぞれ異なるサービスに対して，ミクロな視点から解析を行った．その結果，それぞれのサービスではそれぞれ異なる傾向が現れた．つまり，ミクロな視点での新規タグの生み出され方は Yule-Simon 過程の枠組みでは説明できないということがわかった．Delicious では新たなタグの生成確率と同時に付けら

れるタグの数の間の関係には正の相関が現れ，RoomClip と Instagram では負の相関が現れた．また，Flickr では相関はあらわれなかった．

正の相関は同時に付けられるタグが増えるのは新たなタグが生成されるためであるということを表している．負の相関は同時に付けられるタグが増えるのは既存のタグが選択されるためであるということを表している．その後の情報検索という観点においては，新たなタグというのはシステムにおいてはこれまでに存在していないタグであり，他のユーザーの情報検索を助けるようなタグではないことが考えられる．つまり，あまり他者を意識していないようなタグつけの場合には正の相関が現れることが考えられる．逆に負の相関が現れる場合には既存のタグが選択されることを意味し，それは他のユーザーからの検索にヒットしやすくなることを意味する．つまり，他者からの検索を意識したタグつけの場合には負の相関が現れることを意味する．

これらのユーザーがタグをつける動機を捉えた値として記述者，分類者という値を利用し，各サービスでのユーザーのタグをつける動機を分析した．記述者は情報を詳細かつ客観的なタグにより記述することで，他者からの検索を意識したタグつけである．分類者は情報を主観的なタグにより分類することにより，その後の管理を楽にすることを意識したタグつけを行う．これらの分析の結果，Delicous は分類者的なタグつけを行うユーザーの多いサービスであることがわかった．また，RoomClip，Instagram では記述者的なタグつけを行うユーザーの多いサービスであることがわかった．Flickr ではユーザーの傾向は広く分布していることがわかった．

このような，各サービスにおけるユーザーのタグをつける動機に対する解析の結果は，我々の仮定をよく補強するものであった．つまり，記述者的な他者からの検索されることを意識したタグつけの多く見られるサービスの場合には，ミクロなタグつけには正の相関が現れ，分類者的な自身のための情報管理を意識したタグつけである場合には，ミクロなタグつけには負の相関が現れ，両者の混在するサービスである場合には相関は現れないということである．

6. 結論

本研究では新規タグの生成メカニズムに注目し，ソーシャルタギングに典型的なタグの同時利用との関係を考えた．また，４つのそれぞれ異なる性質のサービスに対して分析を行い，その関係がソーシャルタギングに不変に見られる法則であるのかを確かめた，その結果，両者の関係は４つのサービスでそれぞれ異なることがわかった．それらの違いはサービスの性質により，SNS 的な側面の強いサービスであるか，SNS 的な側面が弱く，個人の情報管理的側面が強いサービスであるかにより引き起こされていることがわかった．本研究から，新規タグの生み出され方のメカニズムに影響を及ぼす要素の一つにタグの同時利用の大きさが影響を及ぼし，その相関はサービスの傾向を考えることで明らかになることがわかった．新たなタグを作るというある種創造的な振る舞いがユーザーの晒される環境により変化するというのは面白く，今後の創造的な振る舞いを促すようなシステムを構築する上で重要な知見になることが考えられる．

謝辞

解析用の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．

参考文献

[Cattuto 07a] Cattuto, C., Baldassarri, A., Servedio, V. D., and Loreto, V.: Vocabulary growth in collaborative tagging systems, *arXiv preprint arXiv:0704.3316* (2007)

[Cattuto 07b] Cattuto, C., Loreto, V., and Pietronero, L.: Semiotic dynamics and collaborative tagging, *Proceedings of the National Academy of Sciences*, Vol. 104, No. 5, pp. 1461–1464 (2007)

[Ferrara 14] Ferrara, E., Interdonato, R., and Tagarelli, A.: Online popularity and topical interests through the lens of instagram, in *Proceedings of the 25th ACM conference on Hypertext and social media*, pp. 24–34ACM (2014)

[Golder 06] Golder, S. A. and Huberman, B. A.: Usage patterns of collaborative tagging systems, *Journal of information science*, Vol. 32, No. 2, pp. 198–208 (2006)

[Görlitz] Görlitz, O., Sizov, S., and Staab, S.: PINTS: peer-to-peer infrastructure for tagging systems.

[Gupta 10] Gupta, M., Li, R., Yin, Z., and Han, J.: Survey on social tagging techniques, *ACM SIGKDD Explorations Newsletter*, Vol. 12, No. 1, pp. 58–72 (2010)

[Heaps 78] Heaps, H. S.: *Information retrieval: Computational and theoretical aspects*, Academic Press, Inc. (1978)

[Heckner 08] Heckner, M., Neubauer, T., and Wolff, C.: Tree, funny, to_read, google: what are tags supposed to achieve? a comparative analysis of user keywords for different digital resource types, in *Proceedings of the 2008 ACM workshop on Search in social media*, pp. 3–10ACM (2008)

[Simon 55] Simon, H. A.: On a Class of Skew Distribution Functions, *Biometrika*, Vol. 42, No. 3/4, pp. pp. 425–440 (1955)

[Strohmaier 12] Strohmaier, M., Körner, C., and Kern, R.: Understanding why users tag: A survey of tagging motivation literature and results from an empirical study, *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 17, pp. 1–11 (2012)

[Wal 05] Wal, T. V.: Explaining and Showing Broad and Narrow Folksonomies (2005)

[Willis 22] Willis, J. C.: *Age and Area: A study in Geographical Distribution and Origin of Species*, CUP Archive (1922)

[Yule 25] 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. Series B, Containing Papers of a Biological Character*, pp. 21–87 (1925)