2H3-4

Zipf's Law and Heaps' Law in Social Tagging

橋本 康弘 *1　佐藤 晃矢 *2　岡 瑞起 *1
HASHIMOTO, Yasuhiro　SATO, Koya　OKA, Mizuki

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

*2筑波大学情報学群情報科学類
College of Information Science, School of Informatics, University of Tsukuba

We briefly review the relationship between Zipf's Law and Heaps' Law observed in our linguistic activities such as natural language or online social-tagging, then introduce and discuss on a new analytic idea that provides another aspect of the relationship between the both laws.

1. はじめに

ウェブ上で共有された動画や写真などのリソースに対して複数のユーザが自由な文字列（タグ）を付与する行為をソーシャルタギングと呼ぶ．ソーシャルタギングシステムは YouTube[*1] や Flickr[*2] といったモダンなオンラインコンテンツ共有サービスにおいて広く提供されており，その利用実態の統計は自然言語に似た Zipf 則と Heaps 則を示す事例が報告されている [Kornai 02, Cattuto 09]．Zipf 則とは語彙の出現数とそれに基づく順位がベキ則に従う関係を指し，定められた期間内の語彙の出現数に関するスタティックな経験則について述べている．また，Heaps 則とは語彙数とコンテンツの増加がベキ則に従う関係を指し，時間とともに新しい語彙が生み出されるダイナミックな経験則について述べている．この 2 つのベキ則は人の言語活動や社会活動において広く観察され，両法則の関係性については様々な議論が行われている [Serrano 09, Tria 14]．本発表ではこの Zipf 則と Heaps 則が同時に現れるメカニズムについてのオーソドックスな考え方についてレビューし，それらとは異なった視点からのメカニズムを示しながら，両ベキ則の関係について考察する．

2. Zipf 則と Heaps 則

離散時刻 t における語彙の総数を $K(t)$，t 以前に出現した語彙 i の数を $n_i(t)$ と表記する．ただし時刻を所与のものとする場合，t は省略する．さらに n_i の降順で語彙にランク（$1 \leq r \leq K$）を与え，r_i と表記する．Zipf 則は以下の関係を指す：

$$n_i \propto r_i^{-\gamma}. \tag{1}$$

また Heaps 則とは以下の関係を指す：

$$K(t) \propto t^{-\beta}. \tag{2}$$

ある語彙の個別のインスタンスをワードと呼ぶなら，本来 Heaps 則は時間に対してではなく読み込まれたワード数に対する語彙の増加を指す．ここでは時間に対して線形にワードが読み込まれると考える．また，β の値は一般に 1 より小さく，これは "sub-linear" 則と呼ばれている．

連絡先: 橋本 康弘, hashi@cs.tsukuba.ac.jp

*1 https://www.youtube.com

*2 https://www.flickr.com/

3. 2 つのベキ則の関係

2 つのベキ則を関係付ける典型的な説明の一つは以下に示す Baeza らによるものである [Baeza-Yates 00]：

1. Zipf 則が成り立つならば $n_i = Ar_i^{-\gamma}$．ただし $A = N/\Sigma_{i=1}^{K} r_i^{-\gamma}$，$N = \Sigma_{i=1}^{K} n_i$.
2. 最も稀な語彙（ランク K）が 1 回程度しか出現しないならば $AK^{-\gamma}$ は $O(1)$ でスケールし，したがって K は $O(A^{1/\gamma})$ でスケールする．
3. $\Sigma_{i=1}^{K} r_i^{-\gamma}$ が適当な範囲に収まるならば（例えば $\gamma > 1$），A は $O(N)$ でスケールする．よって，K は $O(N^{1/\gamma})$ でスケールする．

つまり，Heaps 則と Zipf 則の指数が互いに逆数の関係になることを主張する．

もう一つの説明は，ある種の確率過程に基づいたものである．今次々とワードが出現し，あるワードが出現する際には確率 α で新しい語彙を生成，$1-\alpha$ で既存の語彙から選択するという試行を考える．既存の語彙から選択する際，もし以下の条件で語彙が選び出されるなら，それを Simon 過程と呼ぶ [Simon 55]：

> 過去に n 回使われた語彙をまとめてクラス $[n]$ の語彙と呼び，その語彙数を k_n とすると，$[n]$ に属する語彙が選択される確率は nk_n に比例する．

これは過去に多く出現した語彙ほど次の試行でも選択されやすい傾向を記述している．しかし，$[n]$ の中から具体的にどの語彙が選択されるかについては任意性が残されている．このモデルから導かれる語彙の確率分布は

$$p(n) \propto n^{-\rho}, \quad \text{ただし} \quad \rho = \frac{1}{1-\alpha} + 1 > 2 \tag{3}$$

となる．補累積分布にして縦横軸を反転すれば，Zipf 則の指数 $\gamma = 1 - \alpha < 1$ が得られる．また，新規語彙の生成が時間に対して線形であることは定義から明らかであるので，Heaps 則も指数 1 で成立する．つまり，この説明では Zipf 則と Heaps 則の指数は逆数の関係にならない条件を作ることができる．

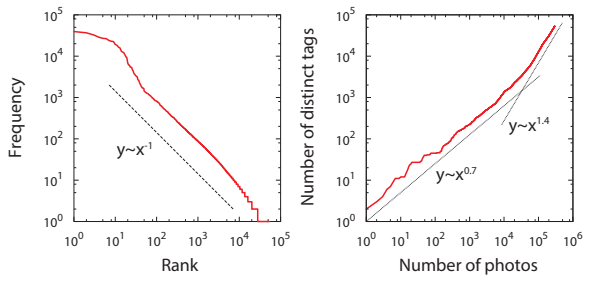

図 1: Tunnel 株式会社が運営するサービス “RoomClip” で使われるタグが示す Zipf 則と Heaps 則.

4. 事例

Cattuto らは Zipf 則と Heaps 則がソーシャルタギングに存在し，その指数が逆数の関係にあることを実際のソーシャルブックマークサービスから得たデータから示した．[Cattuto 09] この結果は Baeza らの見積もりをサポートする．我々は異なる事例として，Tunnel 株式会社が運営する写真共有サービス *3 からソーシャルタギングのデータを得て，その振る舞いについて分析を行った．結果は図 1 に示す通り Zipf 則と Heaps 則が見られたものの，その指数は Zipf 則がほぼ 1，Heaps 則はサービス初期が 0.7，中期以降が 1.4 であった．両指数は逆数の関係になく，また Heaps 則には 2 つのドメインが存在し，さらに中期以降では 1 を超えるという特徴が見られた．これらの結果はここまでの議論では説明ができない．

5. もう一つの考え方

一つの思考実験として Heaps 則を明示的に与えたうえで，Yule 過程 [Bacaër 11] や Barabási らのグラフモデル [Barabási 99] で用いられた “優先的選択ルール” を導入することで，どのような Zipf 則が生成可能かを検討する．

語彙 i の出現数の時間発展を以下で定義する：

$$n_i(t+\Delta t) = n_i(t) + \frac{k n_i(t)}{\sum_i n_i(t)}\Delta t. \tag{4}$$

k は一つのコンテンツが含む語彙の数を表し，ここでは定数とする．つまり，あるコンテンツがポストされたときに語彙 i が選択される確率は，その語彙の過去の出現数 $n_i(t)$ と 1 つのコンテンツが含む語彙の数 k の積に比例する．今，$\sum_i n_i = kt$ であることに注意すれば，$n_i(t)$ についての連続極限を仮定して

$$\frac{dn_i}{dt} = \frac{k n_i(t)}{kt} = \frac{n_i(t)}{t} \tag{5}$$

が得られ，

$$\int^{\infty} \frac{1}{n_i} dn_i = \int^{\infty} \frac{1}{t} dt \tag{6}$$

より，語彙の出現数についての成長則

$$n_i(t) = t/t_i \tag{7}$$

が得られる．ここで t_i は語彙 i が初めて出現した時刻を指し，$n_i(t_i) = 1$ という境界条件を用いた．n_i が任意の n よりも小さい確率 $P(n)$ は

$$P(n) = P(n_i(t) < n) = P(t_i > t/n) \tag{8}$$

となる．これは語彙 i が初めて出現した時刻が t/n 以降である確率を表しており，Heaps 則 $K(t) = at^{\beta}$ を明示的に導入すると，

$$P(n) = 1 - \frac{K(t/n)}{K(t)} = 1 - \frac{a(t/n)^{\beta}}{at^{\beta}} = 1 - n^{-\beta} \tag{9}$$

が得られ，語彙の出現数の確率分布

$$p(n) = dP(n)/dn = n^{-(\beta+1)} \tag{10}$$

が得られる．したがって，Zipf 則の指数は $1/\beta$，つまり Heaps 則の指数の逆数となる．Baeza らの考え方と前提とする条件は異なるが，帰結は同じとなる．この考え方を用いれば Heaps 則の指数を実データに合わせることは可能だが，依然，前節に示した両指数が逆数の関係にない現象は説明できない．

6. 議論

- Zipf 則の指数が変化する条件について．
- Heaps 則の指数が変化する条件について．
- Heaps 則の指数が 1 を超える場合の条件について．

謝辞

We would like to show great appreciation to Tunnel Inc. for providing their data.

参考文献

[Bacaër 11] Bacaër, N.: *Yule and evolution (1924)*, pp. 81–88, Springer Science & Business Media (2011)

[Baeza-Yates 00] Baeza-Yates, R. and Navarro, G.: Block addressing indices for approximate text retrieval, *Journal of the American Society for Information Science*, Vol. 51, No. 1, pp. 69–82 (2000)

[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)

[Cattuto 09] Cattuto, C., Barrat, A., Baldassarri, A., Schehr, G., and Loreto, V.: Collective dynamics of social annotation, *PNAS*, Vol. 106, No. 26, pp. 10511–10515 (2009)

[Kornai 02] Kornai, A.: How many words are there?, *Glottometrics*, Vol. 4, pp. 61–86 (2002)

[Serrano 09] Serrano, M. A., Flammini, A., and Menczer, F.: Modeling Statistical Properties of Written Text, *PLoS ONE*, Vol. 4, p. e5372 (2009)

[Simon 55] Simon, H. A.: On a class of skew distribution functions, *Biometrika*, pp. 425–440 (1955)

[Tria 14] Tria, F., Loreto, V., Servedio, V. D. P., and Strogatz, S. H.: The dynamics of correlated novelties, *Scientific Reports*, Vol. 4, p. 5890 (2014)

*3 http://roomclip.jp/