2J1-2

ソーシャルタギングシステムにおける語彙の個性

Individuality of vocabularies in social tagging systems

橋本康弘 *1 Yasuhiro Hashimoto　佐藤晃矢 *2 Koya Sato　岡瑞起 *3 Mizuki Oka　池上高志 *4 Takashi Ikegami

*1*2*3筑波大学システム情報系 The University of Tsukuba　*4東京大学総合文化研究科 University of Tokyo

In social tagging systems, users create a number of tags, each of which is used according to people's preferences. In other words, the diversity of the tags continue to increase as they are exposed to selection pressure from cultural preferences. This is analogous to living ecosystems in nature. Such population dynamism, yielding so-called Zipf's law, is mathematically described by a simple stochastic model—the Yule–Simon process, which models the manner of emergence and the growth of tag vocabularies. However, in actual web services, we observed that large temporal fluctuation emerges in the growth of the cumulative occurrences of vocabularies, which cannot be explained by the usual Yule–Simon process. We derive the theoretical form for the fluctuation of the process, show how the fluctuation in real data deviates from the prediction, and discuss what should be incorporated into the model to reproduce such anomalous vocabulary individuality.

1. Introduction

A *social tagging system* is a kind of information retrieval system that has been widely adopted in modern web services where individuals share information resources such as photos, movies, music, and webpages. The system allows people to associate their resources with a set of arbitrary short texts, called *tags*, as annotation for future retrieval. Consequently, the vocabulary of tags reflects the diversity of our activities of daily living, including attentions, awareness, preferences, and creations, as if we have coded them democratically into computable entities. From such a point of view, the following two questions are connotative rather than a literal meaning: (1) When and at what rate do we create a new vocabulary? (2) How frequently do we use each vocabulary seasonally and cumulatively? The first question implies a sense of the growth of our cognitive space through the creation of novelty ideas, and the second question implies a sense of what form of the universe each of us dwells in. New vocabularies provide us with new concepts and possibilities in life, and the behaviors triggered by new ideas could lead to the creation of other new ideas [1]. In other words, social tagging dynamics is considered a co-evolutionary system between human behaviors and vocabularies, wherein vocabularies are exposed to selection pressure from the cultural preferences of the time.

2. Yule–Simon process

The two ecological questions posed above may remind us of one mathematical model—the Simon process [2], which describes the stochastic growth of the vocabulary size and the frequency of each vocabulary. In this study, we address the question of what is going on in actual social tagging systems, focusing on what types of behavior can or cannot be explained by the well-defined mathematical model. Several existing studies report social tagging analyses based on the Simon process [3]. The major interest in our study, however, is on the individuality of the vocabularies that appear due to the large fluctuation in the growth of tag usage—the increase in the cumulative occurrences of tags. Deriving an analytical solution for the fluctuation of tag growth for a specialization of the Simon process—the Yule–Simon process, which adopts a *preferential attachment* rule in Simon's framework—we evaluate such individuality observed in real data, and discuss the origin of the deviation from the prediction.

連絡先: 橋本康弘: hashi@cs.tsukuba.ac.jp

3. Temporal fluctuation

The derivation of the theoretical form of the fluctuation is carried out following the idea given by [4]. We denote the number of cumulative occurrences of tag i at time t as $n_i(t)$, and assume a sufficiently large number of tags (i.e., $t \gg 1$) and a sufficiently small value of the novelty rate (i.e., $\alpha \sim 0$), which is the probability of introducing a new vocabulary at every time step. Then, the probability that $n_i(t)$ is n and the probability that $n_i(t)$ becomes n right at time t are defined, respectively, as follows:

$$P[n_i(t) = n] \sim \frac{t_i}{t}\left(1 - \frac{t_i}{t}\right)^{n-1} \quad (t \gg n), \tag{1}$$

$$P[n_i(t) \to n] = P[n_i(t-1) = n-1]\left(\frac{n-1}{t-1}\right), \tag{2}$$

where t_i is the time at which the vocabulary of tag i was used for the first time. Substituting Eq. (1) into Eq. (2), we obtain

$$P[n_i(t) \to n] \sim \left[\frac{t_i(n-1)}{t^2}\right]\left(1 - \frac{t_i}{t}\right)^{n-2}. \tag{3}$$

Note that this joint probability for i, n, and t is marginalized over possible values of n, resulting in $1/t_i$ for the arbitrary t; i.e., $\sum_{n=2}^{n_{\max}} P[n_i(t) \to n] = 1/t_i$. Now we consider

a temporal scale factor, say x, for the growth curve given by the mean-field approximation, $n_i(t) = t/t_i$, and define x as $t \equiv nt_i/x$. This definition means that the timepoint t to achieve the cumulative number of occurrences n is x-times faster than the time that the mean-field approximation predicts. Introducing this scale factor into Eq. (3), we obtain

$$P\left[n_i\left(\frac{nt_i}{x}\right) \to n\right] \sim \frac{x}{t}\left(\frac{n-1}{n}\right)\left(1-\frac{x}{n}\right)^{n-2}. \quad (4)$$

We now pass over to the new measure from discrete time to scaled continuous time. Hence, the probability density function for x should be considered as follows:

$$\begin{aligned} p(x) &= t_i P\left[n_i\left(\frac{nt_i}{x}\right) \to n\right]\frac{dn}{dx} \\ &= x\left(\frac{n-1}{n}\right)\left(1-\frac{x}{n}\right)^{n-2}, \end{aligned} \quad (5)$$

where we used $dn/dx = n/x$ and the sum value of $P[n_i(t) \to n]$ mentioned above. This result tells us that the probability density of x is independent of individual tags. We regard this tag-independent relationship in the growth fluctuation as one of the interesting aspects of the Yule–Simon process as well as another noteworthy aspect, Zipf's law.

4. Empirical analysis

Using this theoretical result, we evaluate the empirical datasets gathered from the actual web services—Delicious and Flickr [5]. In the data, each annotation is represented as a pair of the time stamp of the annotation created and a string of the added tag, and all annotations are sorted temporally in physical time. Here, the physical time is irrelevant and we focus only on the temporal sequence of the annotations. Considering $P[n_i(nt_i/x) \to n]$ for fixed n, we measure x for t_i (i.e., each vocabulary) and compose a probability density distribution with the bin size 0.05. A comparison of the empirical and theoretical results for $n = 2$, 10, and 100 is shown in Fig. 1. We see a large deviation from the Yule–Simon process in both datasets for all values of n. Simply put, the number of vocabularies exhibiting *very fast* and *very slow* growth is significantly larger in the actual data than in the model predictions.

5. Discussion

We speculate that this deviation is related to the combination of the *fitness* of each vocabulary and the effect of the *attention decay* [6]. Large parts of vocabularies are used intensively right after their creation, however, they rapidly decay. On the other hand, the attention on a certain number of vocabularies persists, creating a fat tail as in the probability density distribution with large n, irrespective of when the vocabulary was created. This sort of individuality is not considered in the Yule–Simon process, however, we may incorporate some related mechanism into the Simon process using alternatives to the preferential attachment rule. In the next step, we will extend the idea of measuring fluctuation proposed here to identify the boundary between short-decaying and long-persistent vocabularies, and discuss the prospective alternatives in the tag-selection rule.

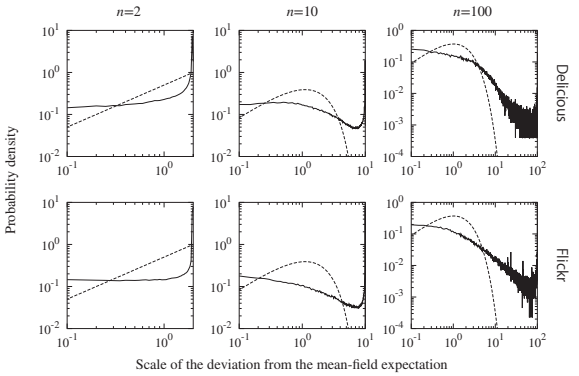

Figure 1: Probability density distribution of the growth fluctuation. Solid and dashed lines show the empirical results and the theoretical curves of the Yule–Simon process, respectively.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 16K00418.

References

[1] Francesca Tria, Vittorio Loreto, Vito Domenico Pietro Servedio, and Steven H. Strogatz. The dynamics of correlated novelties. *Sci. Rep.*, 4:5890, 2014.

[2] Herbert A. Simon. On a class of skew distribution functions. *Biometrika*, 42(3–4):425–440, 1955.

[3] Ciro Cattuto, Alain Barrat, Andrea Baldassarri, Gregory Schehr, and Vittorio Loreto. Collective dynamics of social annotation. *Proc. Natl. Acad. Sci. USA*, 106(26):10511–10515, 2009.

[4] Yasuhiro Hashimoto. Growth fluctuation in preferential attachment dynamics. *Phys. Rev. E*, 93(4):042130, 2016.

[5] Olaf Görlitz, Sergej Sizov, and Steffen Staab. Pints: Peer-to-peer infrastructure for tagging systems. In *Proceedings of the 7th International Conference on Peer-to-Peer Systems*, IPTPS'08, pages 19–19, Berkeley, CA, USA, 2008. USENIX Association.

[6] Pietro Della Briotta Parolo, Raj Kumar Pan, Rumi Ghosh, Bernardo A. Huberman, Kimmo Kaski, and Fortunatom Santo. Attention decay in science. *J. Informetr.*, 9(4):734–745, Sep 2015.