Growth Dynamics of Pairwise Novelty in Social Annotation

Yasuhiro Hashimoto and Takashi Ikegami

The University of Tokyo, Tokyo 153-8902, Japan
{hashi,ikeg}@sacral.c.u-tokyo.ac.jp

Abstract. In social annotation, the vocabulary of tags continues to increase following so-called Heaps' law. However, it has not been extensively studied how the variation of combinatorial usage of tags increases as the web service grows. We introduce the idea of "combinatorial novelty" and investigate how it emerges in both the baseline mathematical model and the empirical web data.

Keywords: Combinatorial Novelty · Social Annotation · Yule–Simon process · Preferential Attachment · Heaps' Law.

1 Introduction

Analogous to the biological evolutionary systems, we will study the evolution of web services in terms of selection, mutation and adaptation. Instead of random processes behind the biological systems, the web services have human activities behind. By studying a large data of a web service, we reveal the essential difference between human activities and random processes. At the same time, we hope to derive new evolutionary concepts from the web evolutionary systems that are applicable to biological phenomena. In case of social tagging systems, we know that the development of new keywords follows Heaps' law, and the size distribution of tag usage follows Zipf's law. [1] An approach of this short paper is to extend Heaps' law to examine the laws of new pair creation; we discuss how new combinations of keywords emerge as the web service grows.

2 Model

The Yule–Simon process is a stochastic process which has been proposed to explain the growth process of evolutionary systems in general [2, 3]. The evolutionary process is described in terms of mutation and selection. In the previous works, we have analyzed a large data of the "RoomClip" service. In this service, users upload their photos with a set of tags, e.g. a photo of kitchen with "kitchen", "white" and "breakfast" tags. By applying YS model to this dataset, we (Y.H.) proved that the process shows Heaps' law and the exponent explains well the resulting size distribution of individual tag usage. Here, we focus more on the evolution of the new pairs rather than of a tag itself. Namely, how many

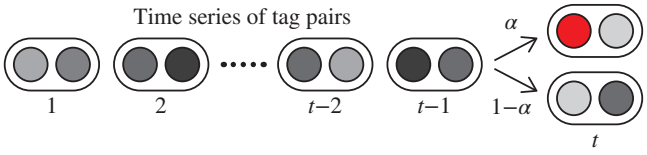

Fig. 1. Evolution of the time series of tag (circles) pairs. Given the current time t, gray circles have been used once or more before t. With a probabilistic trial at every time step, a novel tag (the red one) is introduced to the time series at novelty rate α.

number of new pairs are created in each submission of a photo is the target of our study here. We name it the evolution of *combinatorial vocabulary*.

The original YS model generates a series of symbols $\{\ \sigma_1, \sigma_2, \sigma_3, \ldots, \sigma_t\ \}$, where σ_t designates a tag drawn from a pool of tags with a probability $1-\alpha$. A new tag is introduced with a probability α. We now view a series of symbols as a series of pairs of symbols. The length of the time series extends one by one at each time step; e.g. (ab)(ab)(ac)(dc)..., where two symbols in each parenthesis composes a pair of tags. The schematic diagram of how this abstracted tagging action proceeds is shown in Fig. 1. This "pairwise Yule–Simon process" displays the power-law usage distribution, known as Zipf's law, and the linear growth of the vocabulary size, obviously. However, there is one unique thing, i.e. a *pairwise novelty*, which is defined as a new pair of tags that has never happened in the current time series. A creation of new pair does not require a new tag to be introduced. A combination of already existing tags can provide new pairs. At what rate the size of the pairwise vocabulary grows? Do we find any empirical law, namely, the "second-order Heaps' law" there?

3 Analytical view

There can be two mechanisms for creating pairwise novelty; one is generated by coupling with a novel tag and existing tags and the other one is combination of existing tags. Let us consider two kinds of tags i and j existing in the current time series, and the birth time of i—t_i precedes t_j. We denote the number of co-occurrences of i and j at time t by $e_{ij}(t)$. Then, the probabilities that $e_{ij}(t)$ is equal to zero and becomes one are written, respectively, as follows:

$$P\left[e_{ij}(t)=0\right] = \left\{1-\frac{n_i(t_j)}{\mathcal{A}(t_j)}\right\} \prod_{\tau=t_j+1}^{t} \left[\alpha + (1-\alpha)\left\{1-\frac{n_i(\tau)}{\mathcal{A}(\tau)}\frac{n_j(\tau)}{\mathcal{A}(\tau)}\right\}\right], \quad (1)$$

$$P\left[e_{ij}(t)\to 1\right] = P\left[e_{ij}(t-1)=0\right] \times \frac{n_i(t-1)}{\mathcal{A}(t-1)}\frac{n_j(t-1)}{\mathcal{A}(t-1)}, \quad (2)$$

where $n_i(t)$ is the number of total usage of i at t. The first $\{\cdots\}$ part of Eq. (1) means the probability that tag j does not co-occur with existing i when j is used

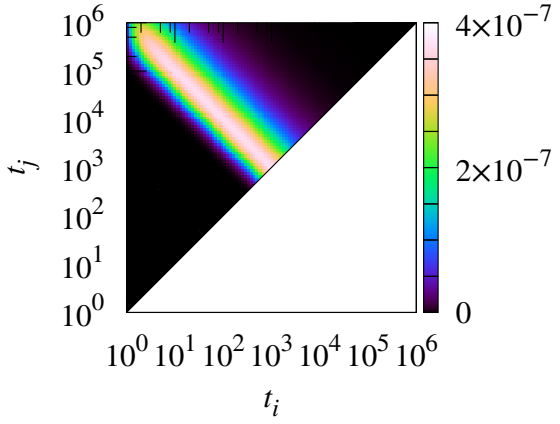

Fig. 2. Probability configuration of Eq. (3) for $t = 10^6$. $t_i < t_j$ and they are bounded by the upper limit t. The color hue indicates the value of the probability for given t_i and t_j.

for the first time. The second part, the product, means the probability that tag i and j never co-occur after the introduction of j throughout to time t.

Using the mean-field solution of individual tag growth, $n_i(t) \approx 2(t/t_i)^{1-\alpha}$[4], and in the asymptotic limit of small α, we obtain an apparently simple form of the probability (2) as follows:

$$P\left[e_{ij}(t) \to 1\right] \approx \frac{1}{t_i t_j}\left(1 - \frac{1}{t_i}\right)\left(1 - \frac{1}{t_i t_j}\right)^{t - t_j - 1}. \tag{3}$$

Figure 2 shows the example of the probability configuration of Eq. (3) in terms of t_i and t_j for $t = 10^6$.The prerequisite $t_i < t_j$ blanks the right-bottom triangular area. Let us remind that tags having a smaller birth time tend to be used more frequently than those having a larger birth time by preferential attachment [5]. So, if both i and j would be such old and well-established tags, it is unlikely to observe that they never co-occur until t ($\gg t_i, t_j$) (see the dark left-bottom area). On the other hand, if both i and j are rather recent (close to t) ones, it is also difficult for them to co-occur at t (see the dark right-top area); because they have not grown sufficiently to be chosen simultaneously among other competent tags.

In order to estimate the growth rate of the combinatorial vocabulary of the model, we need to sum up Eq. (3) for all pairs of i and j as follows:

$$P\left[e_{**}(t) \to 1\right] = \sum_{i}^{K(t)} \sum_{j>i}^{K(t)} P\left[e_{ij}(t) \to 1\right]. \tag{4}$$

This is the rate at which an arbitrary pair of existing tags is used for the first time—*combinatorial novelty rate*, which is what we want to quantify. Integrating

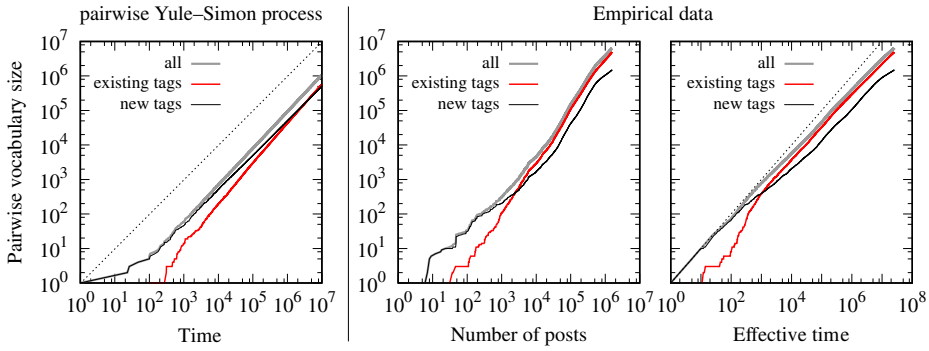

Fig. 3. Growth of the pairwise vocabulary size in the pairwise Yule–Simon process (left) and the empirical data (middle and right). Gray, red, and black curves are the total pairwise vocabulary size, the contribution by the combination of existing tags, and the contribution by the introduction of new tags. Dotted lines are the upper bound of the total size, mentioned in the main text.

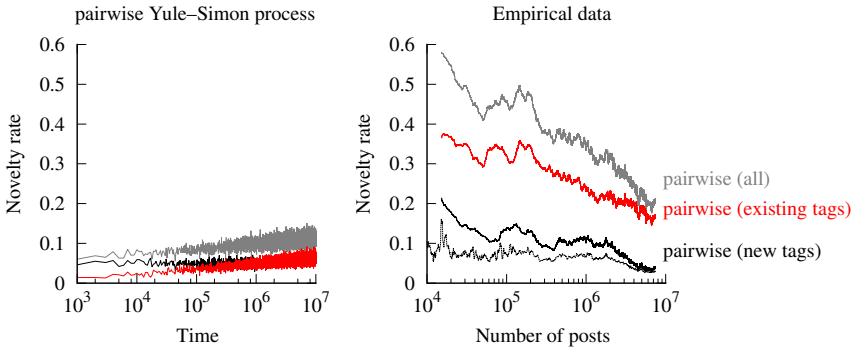

Fig. 4. Time evolution of the pairwise novelty rate—the time differential of Fig. 3. The dotted line in the empirical result shows the novelty rate of a tag itself.

Eq. (4) with respect to time, we will know how the combinatorial vocabulary grows. However, the part of $(1 - 1/t_i t_j)^{t-t_j-1}$ in the equation is apparently impractical to be approximated asymptotically. So, we perform the simulation of the pairwise Yule–Simon process directly, investigate the combinatorial vocabulary growth in the model numerically, and compare with the result from the empirical data analysis in the next section.

4 Simulation & empirical results

We set the final length of the time series to $T = 10^7$ and the novelty rate to $\alpha = 0.05$ (comparable to the empirical value in the next paragraph). As the result, the final vocabulary size $K(T)$ reached approximately 5×10^5. The left

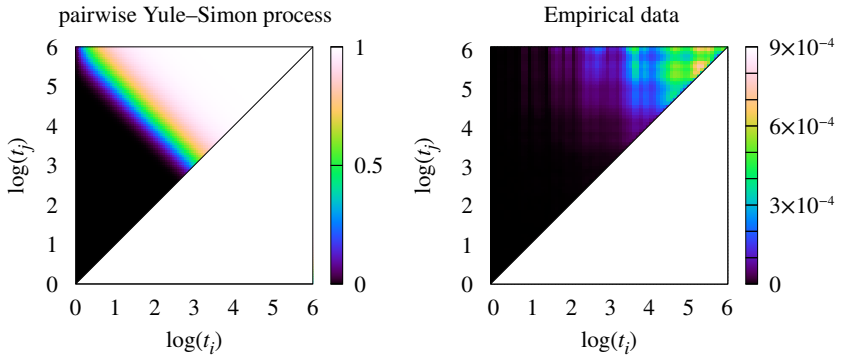

Fig. 5. Weighted probability configuration. The left panel is the result of the model. Note that we have to weight the probability (3) by $t_i t_j$ when using logarithmic bins in order to cancel the counting bias. The right panel is the empirical result, in which we counted the number of pairs used for the first time in the latest 10^6 pairs out of approximately 2.5×10^7 pairs in total.

panel of Fig. 3 shows the growth of the pairwise vocabulary against time in the model. The total pairwise vocabulary size (shown by the gray curve) is upper bounded by the number of time steps (shown by the dotted line). In the initial stage, the growth of the pairwise vocabulary is mostly dominated by the increase of tag vocabulary itself—that is, a new pair is created by introducing a new tag, whose contribution is shown by the black curve. This is in proportion to time, αt. However, after the tag vocabulary becomes large at around $10^6 \sim 10^7$ time steps, the contribution by combinatorial novelties between existing tags (shown by the red curve), which is what we expected by Eq. (4), overtakes the former one.

Concerning the data coverage of the actual tagging behavior in RoomClip, we noticed that the number of photo-postings, annotations, and the final vocabulary size are fitted approximately as 1.5×10^6, 7.4×10^6, and 3.3×10^5, respectively. The novelty rate is, in general, relatively higher at the early stage of the service and gradually goes down when the service gets matured. The latest novelty rate in the data is below 0.05. The middle and right panels in Fig. 3 show the growth of the pairwise vocabulary size. The number of tags used in a photo varies significantly post by post. So, we need to take such a varying increase of possible combinations of tags into account when comparing with the model. We thus re-defined the "effective time scale" as the cumulative number of possible pairs per each post (the right panel) instead of the number of posts (the middle). For example, if a post contains 5 tags, the effective time step is incremented by $\binom{5}{2}$. As a result, the effective time extends to 2.5×10^7 pairs in total. The empirical result exhibits a similar tendency to the model with respect to the crossover phenomena between the contributions by combinatorial novelties and introduction of new tags.

However, the empirical growth in the effective time scale seems to deviate downward from the upper bound, whereas the growth curve in the model scales linearly with time. At the same time, evolution of the combinatorial novelty rate depicted in Fig. 4 exhibits that the combinatorial novelty can be sustained while the creation rate of new tags is decreasing as the time elapses. So our hypothesis here is that a mechanism of creating novelty is gradually shifting from a single tag event to combinatorial tag events, in which the combinatorial tag events follows the selection bias to suppress the creation of novel tag-pairs in the actual tagging behavior. We speculate that this is attributed to that the more recent tags will be combined to use in the service. Figure. 5 shows the probability configuration, which we explained in Fig. 2, in the model (left) and the empirical data (right). As for the result of the model, we have to weight the probability (3) by $t_i t_j$ when using logarithmic bins to compare with the empirical result. The empirical result exhibits that creation of novel pairs are gathered around the combination of large t_i and t_j.

5 Discussion

The web services are artificial evolving ecosystems that teaches us how novelty develops in time, especially when we can deal with a large dataset from the beginning of a service. In the field of artificial life, novelty search dynamics and algorithm has been a big theme. So far Darwinian evolution (mutation and selection) is accepted as a main concept also for the artificial evolving systems. Genetic algorithm is an example. Recently, particle swarm optimization (called PSO [6], Swarm intelligence [7]) or evolving neural nets by using "NeuroEvolution of Augmenting Topologies" (called NEAT [8]) have slightly new aspects. But none of them have achieved open-ended evolution that creates novelties without being stacked.

Here in this paper, we have newly introduced a production of novelty by combination. Combination of tags can change the original meanings of tags to create a new meanings. This can be a new mechanism of open-ended evolution. In addition to this, we have updated the Yule–Simon process by introducing a tag-formation. This formation can be made further sophisticated to include more complex nature of producing novelties in a system.

Acknowledgements

This study is supported by MEXT as "Challenging Research on Post-K computer" - Modeling and Application of Multiple Interaction of Social and Economic phenomena (hp160264) and JSPS KAKENHI Grant Number 16K00418.

References

1. Cattuto, C., Barrat, A., Baldassarri, A., Schehr, G., Loreto, V.: Collective dynamics of social annotation. Proc. Natl. Acad. Sci. USA **106**(26), 10511–10515 (2009)

2. Yule, G.U.: A mathematical theory of evolution, based on the conclusions of dr. j. c. willis, f. r. s. Phil. Trans. Roy. Soc. B **213**, 21–87 (1925)
3. Simon, H.A.: On a class of skew distribution functions. Biometrika **42**(3–4), 425–440 (1955)
4. Hashimoto, Y.: Growth fluctuation in preferential attachment dynamics. Phys. Rev. E **93**(4), 042130 (2016)
5. Newman, M.E.J.: The structure and function of complex networks. SIAM Rev. **45**(2), 167–256 (2003)
6. de Oca, M.A.M., Stützle, T., Birattari, M., Dorigo, M.: Frankenstein's pso: A composite particle swarm optimization algorithm. Trans. Evol. Comp **13**(5), 1120–1132 (Oct 2009). https://doi.org/10.1109/TEVC.2009.2021465, http://dx.doi.org/10.1109/TEVC.2009.2021465
7. Dorigo, M., Birattari, M.: Swarm intelligence. Scholarpedia **2**(9), 1462 (2007). https://doi.org/10.4249/scholarpedia.1462, revision #138640
8. Stanley, K.O., Miikkulainen, R.: Evolving neural networks through augmenting topologies. Evol. Comput. **10**(2), 99–127 (Jun 2002). https://doi.org/10.1162/106365602320169811, http://dx.doi.org/10.1162/106365602320169811