J Econ Interact Coord (2010) 5:221–230
DOI 10.1007/s11403-010-0065-7

REGULAR ARTICLE

Macroscopic and microscopic statistical properties observed in blog entries

Yukie Sano · Misako Takayasu

Received: 22 August 2009 / Accepted: 17 May 2010 / Published online: 30 May 2010

Abstract We observe the statistical properties of blogs that are expected to reflect social human interaction. Firstly, we introduce a basic normalization preprocess that enables us to evaluate the genuine word frequency in blogs that are independent of external factors such as spam blogs, server-breakdowns, increase in the population of bloggers, and periodic weekly behaviors. After this process, we can confirm that small frequency words clearly follow an independent Poisson process as theoretically expected. Secondly, we focus on each blogger's basic behaviors. It is found that there are two kinds of behaviors of bloggers. Further, Zipf's law on word frequency is confirmed to be universally independent of individual activity types.

Keywords Blog analysis · Time series analysis · Zipf's law · Human dynamics

1 Introduction

Blogs are a new kind of social communication medium in which personal opinions can be easily uploaded on the Web. A typical blog site is maintained by an individual or a small group. Blog users are called bloggers and they post blog "entries" that are freely written texts like those in diaries. These texts include opinions on movies, evaluations of purchased items and announcements of social events. Thus, each word in blog entries may reflect social phenomena. Search engine technologies have been developed to observe the details of blog entries automatically at high speeds. In this paper, we focus on the statistics of blogs from both macroscopic and microscopic perspectives.

Y. Sano (✉) · M. Takayasu
Department of Computational Intelligence and Systems Science,
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology,
4259 Nagatsuta-cho, Midori-ku, Yokohama 226-8502, Japan
e-mail: ysano@phys.ge.cst.nihon-u.ac.jp

2 Data description

Using a search engine similar to "Google Blog Search,"[1] we analyzed Japanese blog databases that were collected from January 1st 2007 to December 31st 2008 by "Dentsu Buzz Research"[2] For given keywords, observation period, and search area, the search engine automatically lists all entries that fulfill the condition. The search engine covers 20 major blog providers in Japan that host more than 10 million blog sites. The total number of observed entries is more than 610 million, and there are about 800,000 new entries uploaded daily on average. While we only focus on Japanese blogs in this paper, the share of Japanese blog sites is known to be largest, about 37%, followed by English and Chinese blog sites for the year 2007 according to the report of Technorati,[3] an internet search engine company for blogs (Technorati 2007). In this paper, we firstly focus on the temporal change of word frequency on blog entries. Specially, we count number of blog entries including a target keyword at least once. Namely, if one blog entry includes the target keyword more than two times, we regard the number of blog entry as one. We randomly choose words from a dictionary of Japanese morphological analysis,[4] which is widely used in the field of natural language processing.

3 Noise reductions

In this section, we introduce a basic procedure to evaluate the genuine word frequency in blogs independent of external factors.

3.1 Effect of spam blogs

While reading the collected blog entries, we easily find that there are blog entries that are obviously not generated by humans. For example, there are cases of blog entries' texts comprising of a meaningless sequence of words, copied articles from major internet news articles or simply repeated advertisement keywords. Further, some entries contain sexual or violent content that lead to a paid-membership site. Collectively, these examples are called *spam* blogs. Some spam blogs are created with the intention to enhance their ranking at sites such as PageRanks (Page et al. 1998). A large amount of spams is generated daily and it causes heavy fluctuations in word frequencies.

In the study of blog analysis, spam is attracting considerable interest, and thus, various methods for the detection of spams have been developed (MIC 2009; Narisawa et al. 2006; Sato et al. 2008). In the search engine of Dentsu Buzz Research, the following spam filters are installed:

- *word salad*: Blog contents are a mixture of seemingly meaningful words that together signify nothing.

[1] http://blogsearch.google.com/.

[2] http://www.dbuzz.jp/.

[3] http://technorati.com/.

[4] http://chasen.naist.jp/.

- *copy & paste*: Blog contents are automatically or manually excerpted from other sources.
- *template*: Blog entry comprises template sentences and fixed keywords.
- *multi post*: Identical blog entries are posted to different blog sites.
- *adult and gamble*: Blog entry contains adult or gambling contents.

In this paper, we use these spam filters to categorize spam and normal blogs. As a result of benchmark testing of the filter for 200 blog entries, the total detection accuracy was 83%, about 40% of the collected blog entries being categorized into spam.

3.2 Effect of system maintenance and population growth

Since blogs are supported by computer systems, there is a possibility that some blog servers suddenly stop working because of maintenance or hardware replacement, and thus, there may be a sudden decrease of word frequency for a period. Moreover, there is a tendency that the total number of blog sites increase almost monotonically (MIC 2009). Hence, the average number of appearances of any word tends to increase in a non-stationary way. Here, we introduce a procedure for adjusting these external or systemic non-stationary effects.

For a given time series of flux fluctuation, Menezes and Barabási introduced a method of separating external noise effect from internal contributions in an open system of complex network model (Menezes and Barabási 2004). Utilizing the fact that flux time series comprise independent small parts of fluxes, they computed the share of the small parts of fluxes in the entire range of collected fluxes. With regard to their mathematical models, the method works successfully by separating external noises from the time series. However, in this case, we cannot assume that each blogger acts independently. Therefore, we introduce a new revised method for the separation of internal and external fluctuations.

For words with low frequency, we assume that bloggers pay little attention to these words and the blog entry numbers may not be significantly affected by external factors. For words with high frequency, we conjecture that bloggers focus on appearance numbers, and thus, these words are significantly affected by external factors. Based on these assumptions, we discern that the contribution of external factors depends on the average value of word frequency. For any given keyword j, we calculate the average value of daily blog entries time series $x_j(t)$ and the correlation coefficient between $x_j(t)$ and the time series of whole collected blogs through spam filters, $X(t)$, as shown in Fig. 1,

$$C_j = \frac{\sum_{t=0}^{T} \left((X(t) - \langle X \rangle) \right) \left(x_j(t) - \langle x_j \rangle \right)}{\sqrt{\sum_{t=0}^{T} (X(t) - \langle X \rangle)^2} \sqrt{\sum_{t=0}^{T} \left(x_j(t) - \langle x_j \rangle \right)^2}}, \tag{1}$$

where $\langle X \rangle \equiv \frac{1}{T+1} \sum_{t=0}^{T} X(t)$ and $\langle x_j \rangle \equiv \frac{1}{T+1} \sum_{t=0}^{T} x_j(t)$. As shown in Fig. 2, we confirm that a clear positive correlation exists between $\langle x_j \rangle$ and C_j compared with

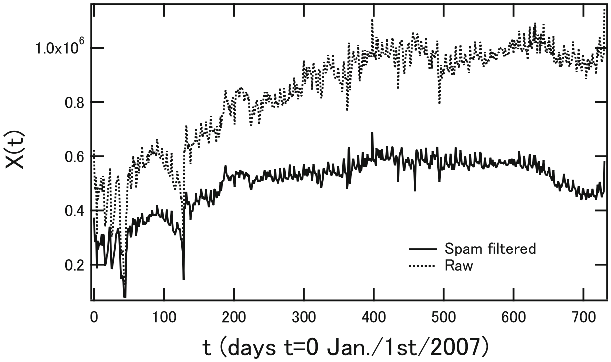

Fig. 1 Time series of whole collected blog entries with spam filtered $X(t)$ and time series without filtered (*dashed line*)

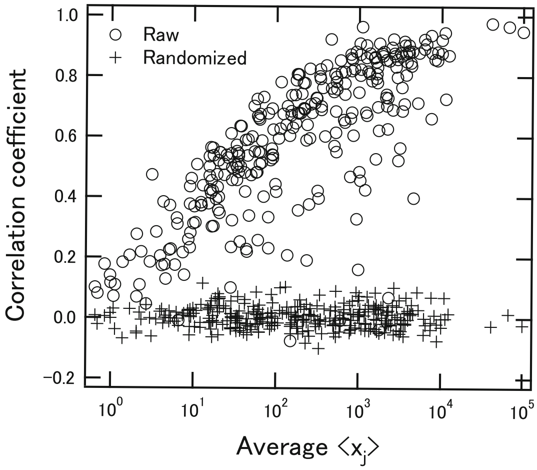

Fig. 2 Comparison of correlation between the average value of frequency $\langle x_j \rangle$ and correlation C_j (*circle*) and C'_j (*cross*)

the case that the blog timestamps are randomly shuffled C'_j. Once we obtain C_j from $\langle x_j \rangle$, we define a new normalization of the time series as follows.

$$F_j(t) = C_j \left(\frac{x_j(t)}{X(t)} \langle X \rangle \right) + (1 - C_j) x_j(t) \tag{2}$$

As shown in Fig. 2, for a small value of $\langle x_j \rangle$, the value of C_j is also small. In this case, the first term of Eq. (2) can be neglected and we have $F_j(t) \approx x_j(t)$. On the other hand, for a large value of $\langle x_j \rangle$, the second term in the right hand side of Eq. (2) becomes negligible and we have $F_j(t) \approx \frac{x_j(t)}{X(t)} \langle X \rangle$. An example demonstrating the effect of this normalization is shown in Fig. 3. We confirm that systematic fluctuations are reduced from the time series (Fig. 3)

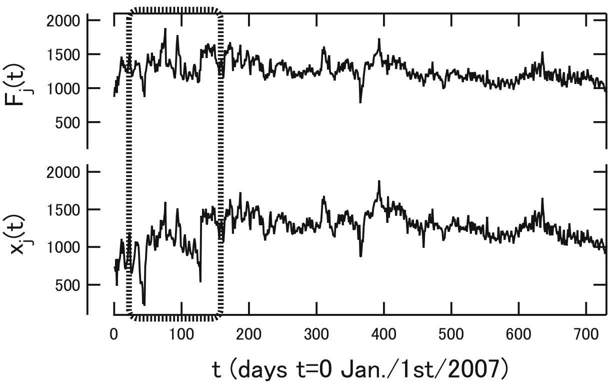

Fig. 3 An example of a word with high frequency (keyword: "if") before and after the normalization ($\langle F_j \rangle = 1241.3$)

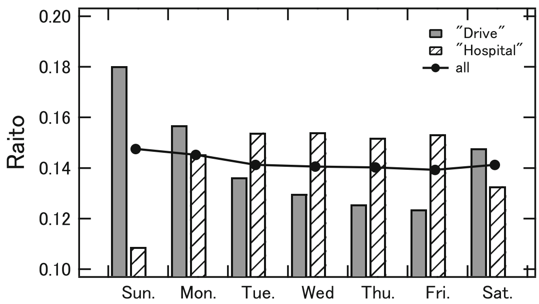

Fig. 4 Ratio for the day of the week; keywords are "hospital" and "drive"

3.3 Effect of weekly period

There are words that exhibit clear periodic behaviors depending on the day of week. For example, "hospital", "office", or "school" are typical words that appear more frequently on weekdays (Fig. 4). For the purpose of flattening such a weekly period, we sum up the number of appearances of words for each day of the week, $N(k)$, $k = 0, 1, \ldots, 6$, where $k = 0$ means Sunday, $k = 1$ means Monday, etc. Then, the time series of word frequency, $F_j(t)$ is normalized by the following way:

$$\overline{F}_j(t) \equiv \frac{F_j(t)}{N_j\ (t \bmod 7)} \frac{N_j}{7}, \tag{3}$$

where $N_j \equiv \sum_{k=0}^{6} N_j(k)$ is the total number of blog entries. Note that for most words, such week dependence normalization is not necessary.

3.4 Result of noise reduction

With the noise reduction procedures introduced in Sects. 3.1–3.3, we confirm that systematic noises are removed and that the time series appears more stationary. For

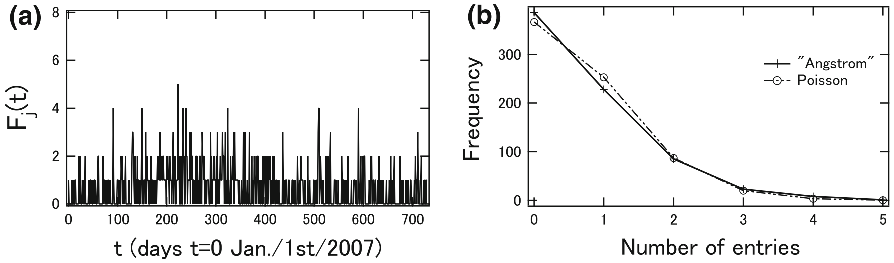

Fig. 5 An example of time series (**a**) and frequency distribution (**b**) for less frequency word ($\langle F_j \rangle = 0.7$); keyword is "angstrom"

words with low frequency, specifically, words that appear less than 1 times per day on average, we confirm that autocorrelation is 0 in 95% significance level. and the distribution of intervals of appearance is checked to pass the statistical tests of Poisson distribution as demonstrated in Fig. 5. The result of χ square test shows that it is not rejected by 2.5% significance level. As a result of randomly selected 300 words from Japanese morphological analysis dictionary, only one word, "Angstrorm", passed the χ square test while remaining words appeared more than 1 times per day on average.

In a pioneering study of the basic statistics of blogs, while Lambiotte et al. (2007) reported that the Poissonian hypothesis is always rejected, they did not apply systematic noise reduction procedures. For words with high frequency, we generally find a clear deviation from the simple Poisson process as already presented (Lambiotte et al. 2007), even these noise reduction processes are fully applied. This implies that there is potential interaction among bloggers.

4 Bloggers' individual properties

So far, we have discussed about word frequency in blogs from a macroscopic point of view. In this section, we focus on the individual properties of bloggers, which are expected to form the base for the development of agent-based modeling of blogs.

4.1 Intervals of posting

In this section, we focus on the bloggers' behaviors of posting blogs. We analyze bloggers' data in which individual bloggers' entries are recorded with the time stamp of precision in second from November 1st 2006 to March 31st 2009. If a blogger's behavior is approximated by an independent Poisson process, the distribution of intervals is approximated by an exponential function and the autocorrelation is almost 0. From this theoretical viewpoint, we categorize the bloggers into two cases:

– *Case 1*: Poissonian bloggers
– *Case 2*: Non-Poissonian bloggers

In Fig. 6, we show two typical examples belonging to these two cases. In Case 1, the observed time series of postings (the top figure of Fig. 6a) is characterized by a

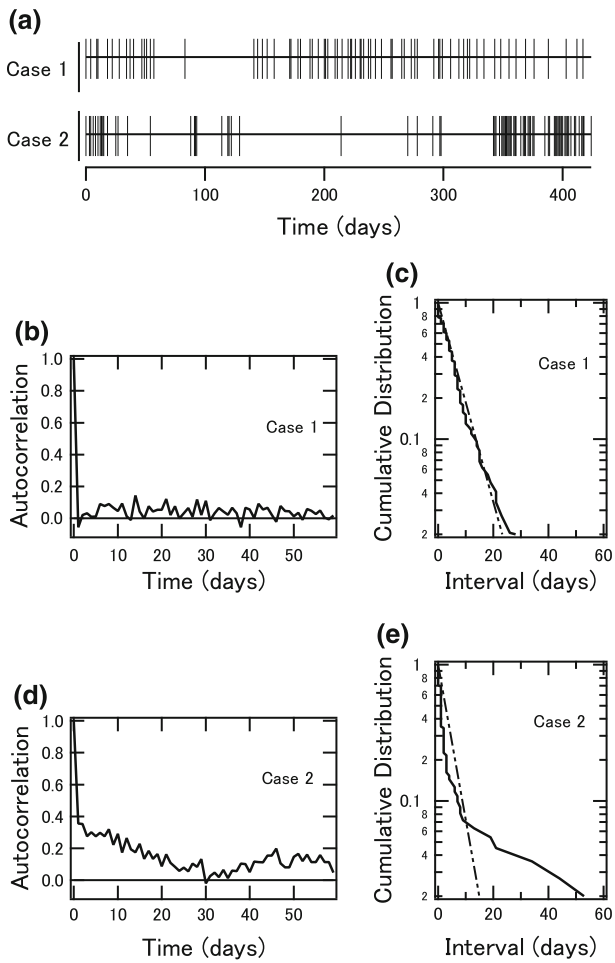

Fig. 6 Comparison of two cases of posting blog entries (**a**); Case 1 is well approximated by Poisson process and Case 2 reveals non-trivial correlation; Autocorrelation function (**b**), (**d**); and cumulative distribution of posting interval (**c**), (**e**)

quick decay of the autocorrelation function (Fig. 6b) and the distribution of intervals is well approximated by an exponential function (Fig. 6c). Therefore, an independent Poisson process can form the base of the behavior for such bloggers. On the other hand, in Case 2, the occurrence time series clearly shows clustering (the bottom figure of Fig. 6a), and the autocorrelation decays slowly (Fig. 6d). Furthermore, the interval distribution has a fat-tail that is approximated by a power law (Fig. 6e). From this example, we find that a non-Poissonian blogger possesses strong memory in that once he (or she) posts an entry, he (or she) tends to continue posting entries. In this analysis, we classify 110 bloggers into these two cases. There are 10 Poissonian bloggers and the remaining 100 bloggers are categorized into the non-Poissonian case by Kolmogorov-Smirnov test applied to the sequence of posting time intervals. We also

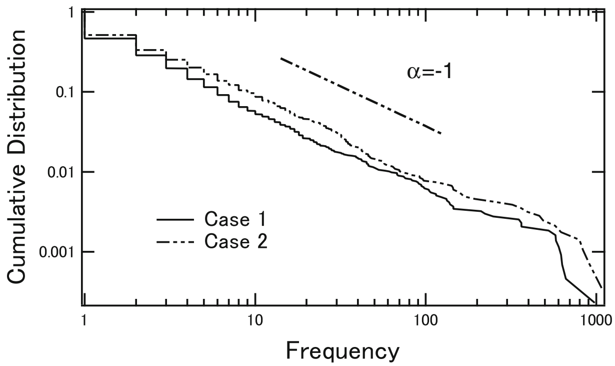

Fig. 7 Word frequency distribution of different bloggers. The guideline follows a power law with an exponent of -1

confirm that autocorrelation functions of Poissonian bloggers are always within 95% confidence bands. Additionally, we also checked 9753 bloggers during the observation period from November 1st 2006 to July 10th 2009 by the Kolmogorov-Smirnov test, and we confirmed that 1089 (about 11%) are categorized into the Poissonian bloggers. In contrast to the keyword appearance described in Sect. 3.4, the behaviors of bloggers can not be simply categorized by the average number of entries, that is, there are both kinds of bloggers in any posting rate groups.

4.2 Individual word frequency

In this section, we investigate the frequency of words in individual blog entries. The study of the frequency of words started in the 1930s by the linguist Zipf who counted the number of appearance of words in various documents. He determined the rank by sorting the words with respect to the frequency and found an empirical law that the frequency of a word is approximately proportional to the inverse of its rank (Zipf 1949). This old law, generally called "Zipf's law," still attracts considerable interest among scientists of various fields because it is applicable not only to linguistic problems but also to a wide variety of phenomena such as the incomes of companies (Okumura et al. 1999) and the abundances of expressed genes distributions (Furusawa and Kaneko 2003).

In Fig. 7, word frequency distributions are plotted for both Poissonian and non-Poissonian bloggers, and we can find that Zipf's law holds in both cases. High frequency words are mainly postpositional particles and auxiliary verbs that commonly appear in all blog sites. Further, some topical keywords also appear frequently in each blogger's entries. In the case of words with low frequency, there are no common words and the words depend on each blogger's characteristics. It is noteworthy that even under such non-uniformity, Zipf's empirical law holds for each blogger's entries.

5 An application of blogs

Finally, we introduce an example showing that blogs can efficiently capture certain kinds of social phenomena. Special keywords such as "flu" and "pollen allergy" appear

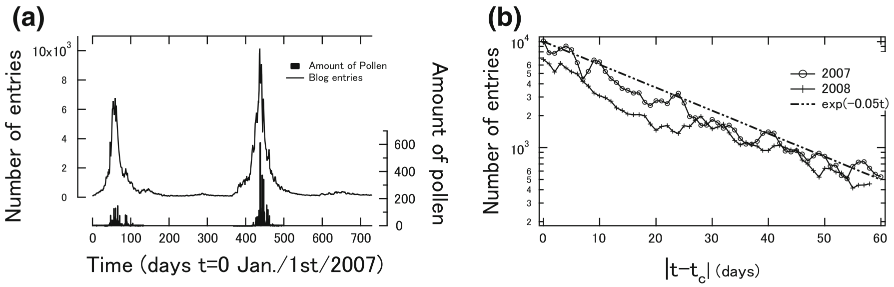

Fig. 8 Time series including word "pollen" and the amount of airborne pollen (**a**). Time series of after peak t_c in semi-log scale (**b**)

periodically every year with sharp peaks as shown in Fig. 8a. We find that the number of blog entries including the keyword "pollen" is closely associated with the amount of airborne pollen in Tokyo.[5] Furthermore, this sharp rise and decay is well approximated by exponential functions as plotted in Fig. 8b. A recent paper by Ginsberg et al. (2009) introduced a method of detecting influenza epidemics by using large numbers of Google search queries to track influenza-like illness. There is a possibility that blogs can also be used as an observation tool for the development of epidemics or allergy.

6 Discussion and conclusion

In this paper, we proposed a basic preprocess for the separation of external systematic noises from the time series of blog entries. After this normalization procedure, we confirmed that as theoretically expected, the appearance of low frequency words clearly follows a Poisson process. With regard to words with high frequency, the Poissonian assumption does not hold in any case, implying that existence of a strong non-trivial correlation among those words. We focused on each blogger's behavior of posting blog entries and found that bloggers can be categorized into two cases: Poissonian and non-Poissonian bloggers. About 20% of bloggers belong to the Poissonian case in which basic behaviors can be modeled by an independent Poisson process. The rest of the bloggers tend to behave in an intermittent manner with strong memory effect. In any case, the word frequency follows Zipf's law for each individual.

We expect that these basic results will play an important role in the construction of a microscopic agent-based model of bloggers in the near future. One of the targets of such a microscopic model will be the explanation of the macroscopic behaviors related to the exponential rise and decay of commonly used keywords such as "pollen" shown in Sect. 5.

Acknowledgments The authors are grateful to the corporations of Dentsu Inc. and Hottolink Inc. for providing blog data.

5 http://kafun.jaanet.org/.

References

Furusawa C, Kaneko K (2003) Zipf's law in gene expression. Phys Rev Lett 90:088102
Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L (2009) Detecting influenza epidemics using search engine query data. Nature 457:1012–1014
Lambiotte R, Ausloos M, Thelwall M (2007) Word statistics in blogs and rss feeds: towards empirical universal evidence. J Infometr 1:277–286
Menezes MA, Barabási A-L (2004) Separating internal and external dynamics of complex systems. Phys Rev Lett 93:068702
Ministry of Internal Affairs and Communications (2009) Report of institute for information and communication policy (Japanese)
Narisawa K, Yamada Y, Ikeda D, Takeda M (2006) Detecting blog spams using the vocabulary size of all substrings in their copies. In: Proceedings of the 3rd annual workshop on weblogging ecosystem
Okuyama k, Takayasu H, Takayasu M (1999) Zipf's law in income distribution of companies. Physica A 269:125–131
Page L, Brin S, Motwani R, Winograd T (1998) The pageRank citation ranking. Bringing from the standford digital library technologies project
Sato Y, Utsuro T, Fukuhara T, Kawada Y, Murakami Y, Nakagawa H, Kando N (2008) Analysing features of japanese splogs and characteristics of keywords. In: Proceedings of the 4th international workshop on adversarial information retrieval on the web, pp 33–40
The State of the Live Web, April 2007. In: Sifry's Alerts Available via http://www.sifry.com/alerts/archives/000493.html. Accessed April 30th, 2009
Zipf GK (1949) Human behavior and the principle of least effort. Addison-Wesley, Cambridge