

Statistical properties of number fluctuations observed in Internet blog keywords

Yukie Sano, Kenta Yamada, Kota Watanabe,
Takayuki Mizuno¹, Hideki Takayasu², Misako Takayasu

Tokyo Institute of Technology, 4259 Nagatuta-cho, Yokohama 226-8502, Japan

¹Institute of Economic Research Hitotsubashi Univ., 2-1 Naka, Kunitachi, Tokyo 186-8603, Japan

²Sony CSL, 3-14-13, Higashi-Gotanda, Shinagawa, Tokyo 141-0022, Japan

Abstract

Human activity of word-of-mouth may be very important for our societies, however, it was impossible to observe its historical record quantitatively. The Internet has changed the situation drastically. Instead of vocal information exchange, people use textual information in blogs. By using the search-engine technology we can observe appearance of any given keyword in blogs automatically with detail time stamps. It is a new scientific activity to explore empirical laws in the number fluctuation of blog keywords and to clarify its impact to the society.

In order to establish empirical statistical laws from time sequential data in general, it is required that the data is stationary. However, in the case of blog keywords there are a few inevitable non-stationary factors which make the analysis difficult. For example, the number of blog sites tends to increase nearly monotonically, so the average number of keywords may grow. Or some blog servers suddenly stop working due to maintenance or hardware replacement, which may cause sudden decrease of word frequency for a while. Moreover, there is always a calendar effect such that keyword numbers increase on holidays. It is important to introduce a procedure of normalization which can evaluate the keyword frequency independent of such non-stationary factors.

To this end we calculate daily summation of frequencies for randomly chosen N sample adverbial words such as "more". Then, by dividing the number of keyword frequency by this summation we get a time sequence of normalized word frequency. It is confirmed that the normalized time sequence successfully removes the above non-stationary factors. Applying this method we find that any resulting normalized time sequence does not follow an independent Poisson process, instead the keyword frequency shows a long autocorrelation characterized by so-called the $1/f$ noise for those keywords which appear frequently everyday, such as "TOYOTA".

There are keywords which clearly show sharp boom such as "Christmas" which apparently tends to diverge on December 25th. For such a case we can approximate the functional form of increase before the critical day by a power law in terms of the difference of the observing day and the critical day. Also the decay form after the critical day can also be modeled nicely by a power law.

E-mail: sano@smp.dis.titech.ac.jp