Statistical properties of number fluctuations observed in Internet blog keywords

Yukie Sano

Department of Computational Intelligence & Systems Science,

Interdisciplinary Graduate School of Science & Engineering,

Tokyo Institute of Technology, 4259-G3-52 Nagatsuta-cho, Midori-ku, Yokohama 226-8502

Abstract

Human activity of word-of-mouth may be very important for our society, 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", and show that the time sequential pattern of summation nearly converges for N larger than 20. 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".