WHICEB 2018 Proceedings

Wuhan International Conference on e-Business

Summer 6-30-2018

Research on Zipf's Law of Hot Events in Search Engines

Yingfan Xu
School of Business, Sichuan Agricultural University, China

Mingliang Shi
School of Business, Sichuan Agricultural University, China

Research on Zipf's Law of Hot Events in Search Engines

XU Ying-fan[1*], *SHI Ming-liang*[1]

[1]School of Business, Sichuan Agricultural University, China

Abstract: This paper focuses on the amount of searching and browsing of hot events in China and finds that the searching index sequences of daily hot events and weekly hot events are in line with Zipf's law. Through continuous collection of large data samples of multiple dates , We find that the Zipf index of the searching index series for daily hot events fluctuates in a very small range.Through Zipf analysis, we find that only a few events maintain long-term heat. A few events will be the focus of most people, while a few will focus on some directional events. So Zipf distribution describes the balance of economic propensity of sender and receiver during the transmission of information. This research is of some reference to commercial activities that make use of hot events for e-commerce.

Keywords: Zipf’s law, hot events, searching sequence, Zipf distribution

1. INTRODUCTION

Right now, we have entered an era of rapid development of the Internet, and the Internet has become the main way for people to access a large amount of information. As a kind of creative technology, Internet technology innovation has brought great changes to all aspects of economic society [1]. According to "Internet Moore's Law", the amount of Internet information is growing at an extremely high rate, the factors of production tend to be virtualized, and a new factor of production is produced - big data [2].

Network hot events arise spontaneously, which are formed by social members in accordance with specific logic and value requirements [3]. It has the characteristics of fast propagation, strong diffusion, strong interactivity and strong linkage with the real society [4]. The Internet-based hot event is the relatively fixed part of the updated network information, which can present the important events, the focus and the direction of public opinion in the Internet [5]. At the same time, through the statistics of big data samples, it also accords with a phenomenon of herd behavior in the network hot events discovered by Cass R. Sunstein, which provides value guidance for our business activities in e-commerce [6].

So, as one of the most important web information retrieval tools, the search engine has developed rapidly. According to its characteristics of fast retrieval speed and high accuracy, Zipf analysis will be used to study the Zipf distribution of search index sequences of single-day and multi-day hot events and the reason of the fluctuation of Zipf index on different dates [7].

2. CONTENTS OF ZIPF’S LAW

Zipf's law was proposed by George K. Zipf, a professor of linguistics at Harvard University in 1948, who conducted a large number of statistics on the frequency of occurrences of words in English documents to test the quantitative formulas of predecessors [8,9]. Zipf's law is mainly used in natural language courses, its content is: If the frequency of occurrence of each word in a given article is counted, each word is arranged in decreasing order of frequency of occurrence as a sequence, and each word after arrangement is labeled with a natural number increasing from 1. With *r* that serial number, with *f(r)* that frequency, there is the following power law holds:

$$f(r) \sim r^{-\alpha} \qquad (2\text{-}1)$$

[*] Corresponding author. Email: xu850765216@163.con(XU Ying-fan)

In the formula, the index α is called the Zipf index. In the Zipf distribution, the index α is a positive constant, which depends on the distribution and has no relation with other parameters. In most countries' languages, the index $\alpha \approx 1$. It shows that in most countries, only a few words are frequently used, while most words are rarely used.

At present, Zipf’s law is not only used in the analysis of the phenomenon of natural language, but also used in the research of petroleum price, town scale, biological engineering, medicine and other fields. In recent years, Zipf's law has also started to play its role in the field of finance and e-commerce.For example, through Zipf's law, N. Vandewalle, M. Ausloos analyzed the stock price index [10], Y. Fujiwara analyzed bank bankruptcy and its causes [11].

3. ZIPF DISTRIBUTION OF HOT EVENTS IN SEARCH ENGINES

In view of the fact that the main way for people to access information is to search the Internet at the present time, and the retrieval of hot events needs the help of search engines, it is very important to study the general law of frequency of hot events in search engines. By using the ideas and methods of metrology, we regard all the information that can be retrieved by search engine as a whole, take the search index of hot events that are positively correlated with the popularity of hot events as the research object, and explore the Zipf distribution of the search index sequence of hot events.

Because Zipf's law is universally applicable to the distribution of word frequency in natural languages, we adopt the strategy of changing data statistics or adjust the statistical objects to verify the statistical results of Zipf distribution, so that we can study Zipf's law in more depth and have more meaningful expansion.

Then, if the search index of each hot event in search engines is counted, the search index of each hot event is arranged in decreasing order of frequency of occurrence as a sequence, and each hot event after arrangement is labeled with a natural number increasing from 1. We use *r* to denote the sequence number of the search index sequence, *f(r)* the search index, and β as the Zipf index for this distribution, which should be satisfied:

$$f(r) \sim r^{-\beta} \tag{3-1}$$

In order to make the relationship more intuitive, we simply transform the above formula. After taking the logarithm of the above formula, we obtain:

$$\lg f(r) = -\beta \lg r + C \tag{3-2}$$

In the above formula, C is a constant. It is easy to see that the slope of a straight line in logarithmic coordinates is the Zipf index.

This paper selects the hot events and their search indexes in search engines on December 29, 2017 as the data set and arranges them according to decreasing order of search indexes.

Table 1. Relevant data on hot events in search engines on December 29, 2017

Hot events	S/N *r*	Index *f(r)*	*f(r) * r*	lg *r*	lg *f(r)*
Twenty-eight vegetarian dishes on the wedding	1	343999	343999	0.000000	5.536557
Digging out Eight-Diagram tactics when refurbishing	2	333818	667636	0.301030	5.523510
Shared boyfriends appeared in Haikou	3	295163	885489	0.477121	5.470062
Cashing out by Ant-Check-Later was sentenced	4	227910	911640	0.602060	5.357763
A rich second generation hurt people with the death reprieve	5	224724	1123620	0.698970	5.351649
Netizens bumped into Jay Chou	6	212413	1274478	0.778151	5.327181
Ran Yingying was exposed speculation	7	149435	1046045	0.845098	5.174452

The crowd ticketed the police car	8	125267	1002136	0.903090	5.097837
The employee was fired because of sick dozing	9	109429	984861	0.954243	5.039132
Qianbao-Net Zhang Xiaolei surrendered	10	103238	1032380	1.000000	5.013840
……	……	……	……	……	……
Treasury reverse repurchase	47	9296	436912	1.672098	3.968296
Pan Yueming take selfies in the background	48	9179	440592	1.681241	3.962795
Netizens bumped into Yue Yunpeng	49	8751	428799	1.690196	3.942058
Six people were sentenced to death in Bahrain	50	8394	419700	1.698970	3.923969

In the statistical analysis, I treat each set of data as a separate observation. This approach not only reduces the risk of contingency or asymmetry in large data studies, but also allows me to explore the considerable within-study variation in Zipf estimates [12]. More on this below.

Looking at the term $f (r) * r$ in Table 1, we find that the data in a certain section of the middle is stable within a certain range of values, that is, there is a section with stability and the section gradually decreases to both sides. These two phenomena show that the data we collected and studied basically satisfy Zipf's law [13].

Next, we describe the collected data and its logical relationships in figures, which are more intuitive. In the following figure, we use lg*r* as the abscissa and lg*f(r)* as the ordinate. Based on the data shown in Table 1, we can predict that, except for the beginning part, the points in the figure should approximate a straight line [14].

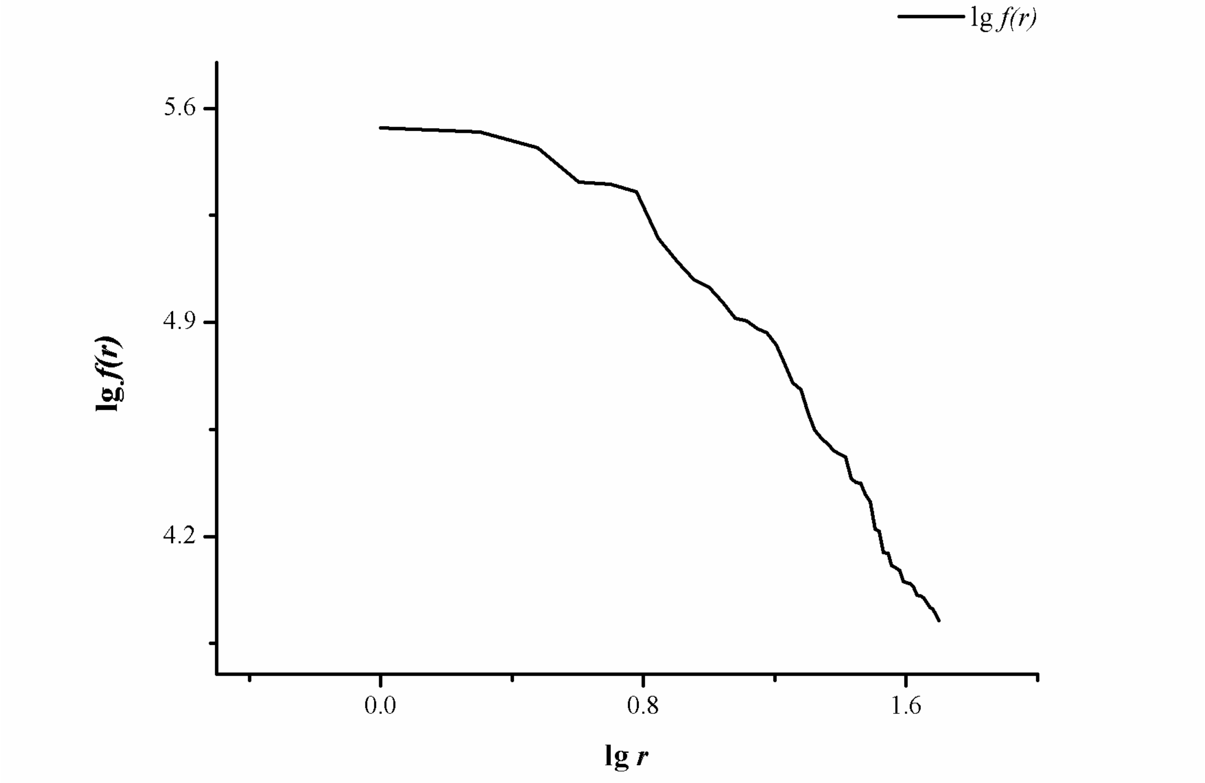

Figure 1. Relevant data on hot events in search engines on December 29, 2017

At the same time, we make a power-law fit to the points in Figure 1 to find the Zipf index. According to the power-law fitting line in Figure 2 below, we can get the Zipf index $\beta = 1.21419$.

From these two figures, we can easily see that the index points except for the first few index points converge to a straight line, which is in line with the prediction and is in line with Zipf's law. For the first few index points, let's explain below.

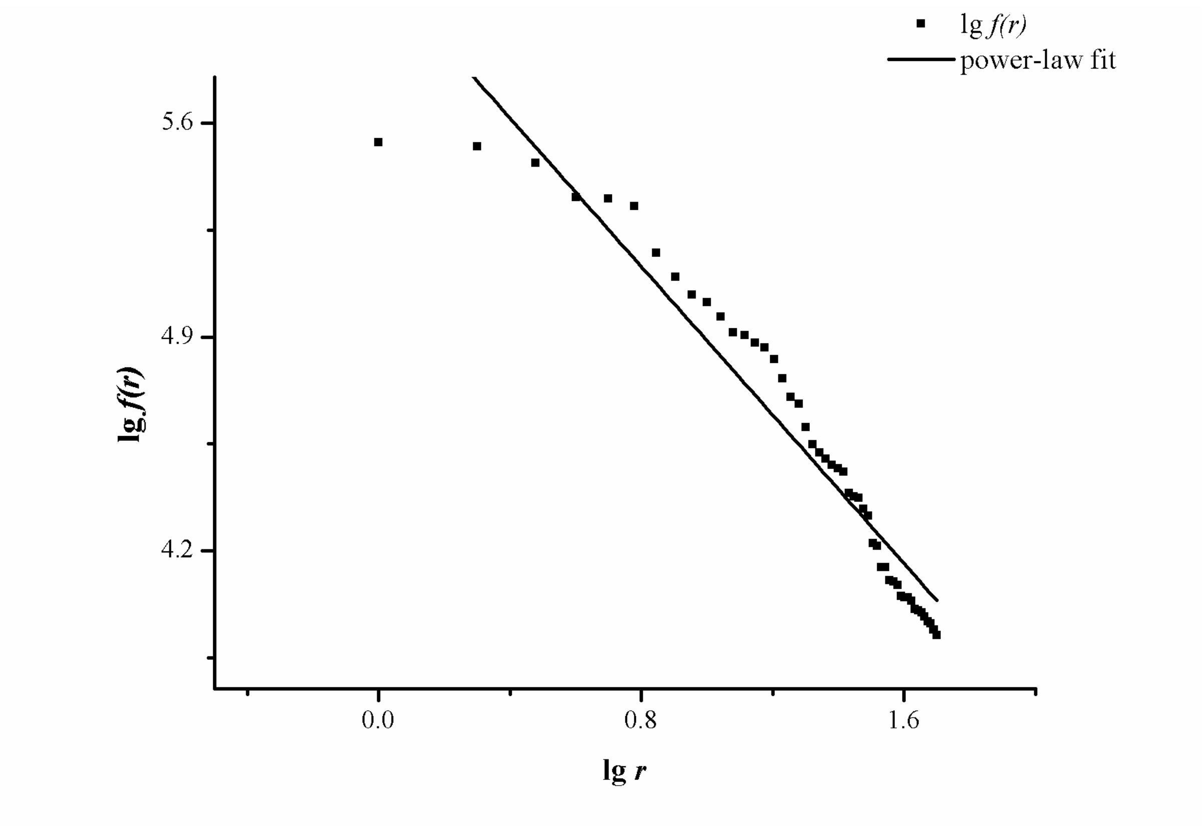

Figure 2. Relevant data on hot events in search engines on December 29, 2017

The distribution of the first few index points is not very regular, because the search engine will recommend the hottest hot events in a variety of ways, which directly affects the search index of these hot events, so this is an inevitable accidental phenomenon without objectivity.

4. ZIPF DISTRIBUTION OF WEEKLY HOT EVENTS IN SEARCH ENGINES

In order to eliminate the risk brought by the asymmetric situation, this paper also takes a week as a unit to collect the comprehensive search indexes of hot events in search engines. Similarly, the search index of each hot event is arranged in decreasing order as a sequence, and each hot event after arrangement is labeled with a natural number increasing from 1. We study this data set in Zipf estimates.

This paper selects the hot events and their search indexes in the 53rd week of 2017 as a data set and sets up the following table according to decreasing order of search indexes.

Table 2. Relevant data on weekly hot events in search engines in the 53rd week of 2017

Hot events	S/N *r*	Index *f(r)*	*f(r)* * *r*	lg *r*	lg *f(r)*
Didi-Motorcycle was stopped	1	764087	764087	0.000000	5.883143
Ma Rong questioned Wang Baoqiang	2	738114	1476228	0.301030	5.868123
Shared boyfriends appeared in Haikou	3	552193	1656579	0.477121	5.742091
Huaxi Village debt 38.9 billion	4	514508	2058032	0.602060	5.711392
Workers removed bones from chicken claws through their mouths	5	490205	2451025	0.698970	5.690378
Ma Su ' s playing Yang Yuhuan is amazing	6	469353	2816118	0.778151	5.671500
Tencent computer housekeeper apologized	7	447991	3135937	0.845098	5.651269
Hu Ge knelt on the ground to sign	8	431208	3449664	0.903090	5.634687

The event of subway rolling people in Shenzhen	9	424482	3820338	0.954243	5.627859
Zhang Han broke up with Nazha G.	10	423504	4235040	1.000000	5.626858
……	……	……	……	……	……
Father accidentally crushed the son while reversing	47	146380	6879860	1.672098	5.165482
The Imperial Palace wall was blown down by the wind	48	134405	6451440	1.681241	5.128415
Two-year-old boy hit the wall and died	49	132501	6492549	1.690196	5.122219
A three-no mp3 exploded suddenly	50	131790	6589500	1.698970	5.119882

Similarly, we use the lg*r* as the abscissa, lg*f(r)* as the ordinate, and the collected data and its logical relations are described by a figure to construct a Zipf distribution model with a week as a unit.

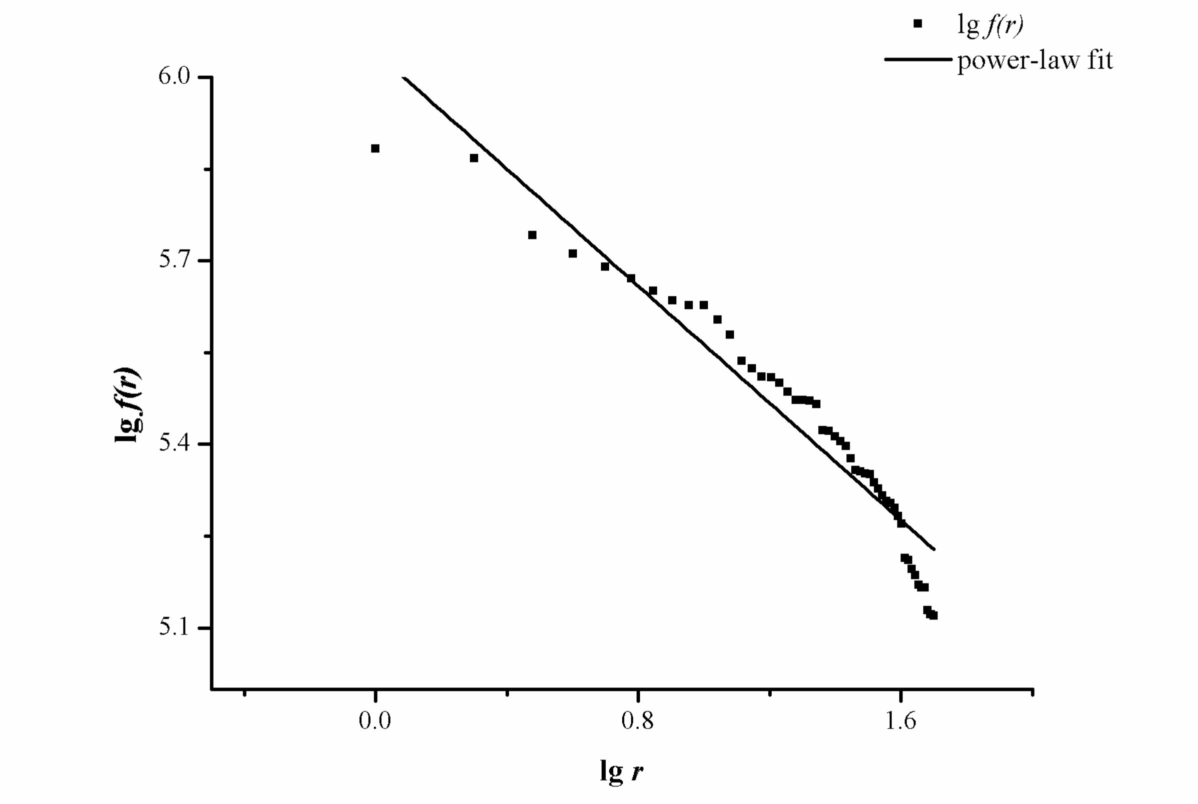

Figure 3. Relevant data on weekly hot events in search engines in the 53rd week of 2017

According to the power-law fitting line in Figure 3, we can get the Zipf index *β* = 0.47826. We find that this figure structure approximates the Zipf distribution of the search index sequence of daily hot events. Even better, the first few index points of the figure are closer to power-law fitting straight line because the amount of data collected in a week as a unit to describe the comprehensive search index is much larger, avoiding the asymmetry of the data distribution.

5. PRINCIPLE OF LEAST EFFORT

Zipf discovered Zipf's law in a number of unrelated phenomena and proposed the Principle of Least Effort to explain the causes of this regularity [15]. Zipf considered that the economy of words need to be discussed from perspectives of both the speaker and the listener. From a speaker's point of view, it is economical to express various meanings in a single word. On the contrary, a listener wants the exact correspondence between the forms and meanings of words [16]. These two principles are contradictory.

The Principle of Least Effort applied to the field of e-commerce, the sender in the transmission of information and the receiver in the acquisition of information both have propensities for the economy. Zipf distribution describes the balance of economic propensity of sender and receiver during the transmission of information, and this is quite useful for improving the efficiency and effectiveness of commercial activities in e-commerce.

6. ANALYSIS OF THE FLUCTUATION OF THE ZIPF INDEX

Based on the big data sample of twenty consecutive days, we make the following fluctuation figure of the Zipf index and find that the search index sequences of daily hot events in the observation period all conform to the Zipf distribution. At the same time, their Zipf indexes fluctuate between a small range, mainly between 1.00 to 1.22. However, one of the Zipf indexes is particularly high, which we will analyze later. The average Zipf index in the observation period is calculated as 1.12636.

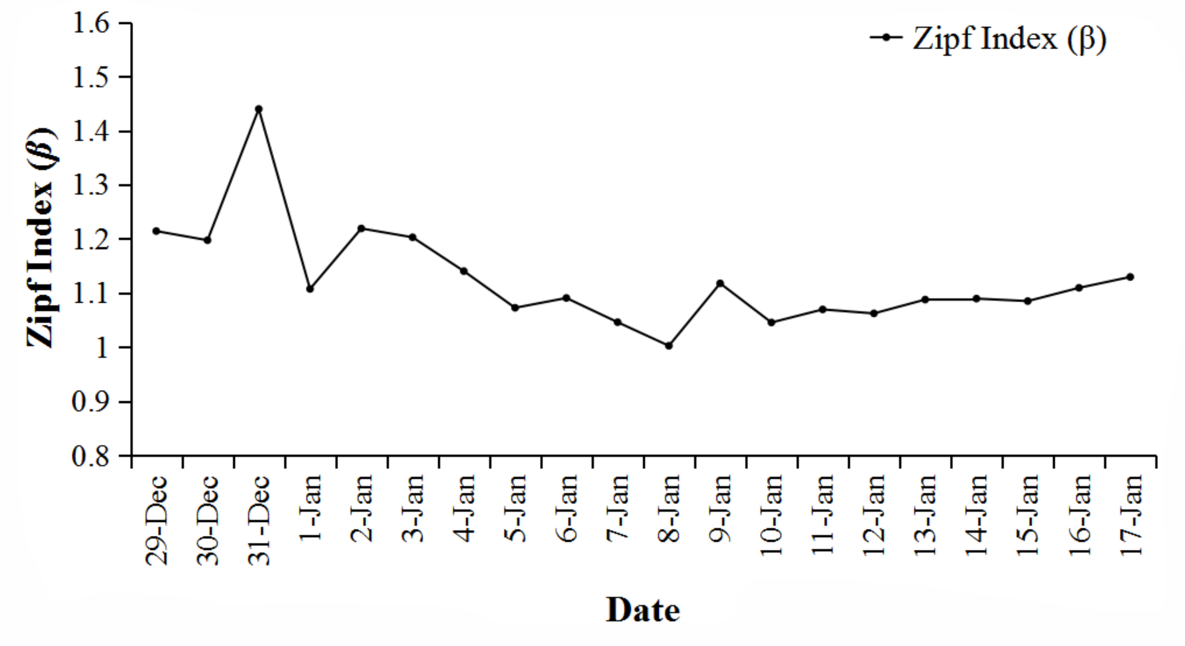

Figure 4. Fluctuation of the Zipf index of daily hot events in search engines in the observation period

As shown in Figure 4, we can visually see the fluctuation of the Zipf index.

According to the change of data in the observation period, we can not judge whether this fluctuation is cyclical or not. However, we suspect that the formation of fluctuation is affected by the non-network information medium, which is equivalent to being forced to vibrate by an irregular driving force.

Next, we focus on the peak of the fluctuation of the Zipf index, which is the Zipf index of the search index sequence of hot events on December 31, 2017. Obviously, the hot event on the top of the sequence on that day is different from that on previous few days, and the search index of this hot event is much higher than the search index of the top hot event on other dates.

The Zipf index of the data set for that day is much higher than the average Zipf index of the entire observation period, however, when we remove the top-ranked hot event, the Zipf index of the search index sequence consisting of the remaining forty-nine hot events is very close to the average Zipf index over the entire observation period.

Because that day is the last day of 2017, the hot event on the top of the sequence that day is related to the

New Year, which is of special significance and receives a great deal of attention. This indicates that hot events with unusually high search index have a direct effect on the Zipf index of the sequence.

7. CONCLUSION

This paper takes the hot events in search engines as the research object, and verifies that the search index sequences of daily hot events and weekly hot events accord with Zipf's law. A few hot events are the objects of most people's attention at the same time, and most of the hot events are those of a few people, which shows that there are similarities and differences in the people's attention. Based on the statistics of big data samples lasting twenty dates, we find that the search index sequences of daily hot events in the observation period all conform to the Zipf distribution, and their Zipf indexes mainly fluctuate between 1.10 and 1.26. Only a small number of events can maintain long-term heat, and most of them are short-term hot events.

This paper also takes a week as a unit to collect the comprehensive search index of hot events in search engines so as to eliminate the interference of the asymmetric situation on the research. At the same time, this paper also collects big data samples for 20 consecutive days to eliminate the interference of contingency.

REFERENCES

[1] Manyika J, Chui M, Bughin J, et al. Disruptive Technologies: Advances That will Transform Life, Business, and the Global Economy[R]. McKinsey Global Institute, May, 2013.

[2] Hu Beibei, Wang Shengguang. The New Production Function of the Internet Age[J]. Studies in Science of Science, 2017, 35(9): 1308-1312. (in Chinese)

[3] Feng Shiqi, Zhu Dedong. On the Dual Impacts of Internet Hot Events on Cybercitizens' Social Mentality[J]. Journal of Chongqing Institute of Technology, 2016, 30(8): 67-71. (in Chinese)

[4] Wang Shiyong. Understanding Network Culture[M]. Chongqing: Chongqing Publishing Group, 2011: 47-50. (in Chinese)

[5] Zeng Yiling, Xu Hongbo. Research on Internet Hotspot Information Detection[J].Journal on Communications, 2007, 28(12): 141-146. (in Chinese)

[6] Cass R S. Going to Extremes: How Like Minds Unite and Divide[M]. New York: Oxford University Press, 2009.

[7] Fang Xi, Li Na, Ge Yuefeng. Evaluation System for Chinese Search Engines Based on the AHP-TOPSIS[J]. Science & Technology Review, 2012, 30(14): 49-54. (in Chinese)

[8] Zipf G K. Human Behavior and the Principle of Least Effort[M]. Cambridge:Addison-Wesley Press, 1949.

[9] Zipf G K. The Psycho-Biology of Language: An Introduction to Dynamic Psychology[M]. Cambridge: Addison-Wesley Press, 1968.

[10] Vandewalle N, Ausloos M. The n-Zipf Analysis of Financial Data Series and Biased Data Series[J]. Phys Rev E, 1995, 52(1): 446-452.

[11] Yoshi Fujiwara. Zipf Law in Firms Bankruptcy[J]. Physica A: Statistical Mechanics and its Applications, 2004, 337(1): 219-230.

[12] Nitsch V. Zipf Zipped[J]. Journal of Urban Economics, 2005, 57(1): 86-100.

[13] Zipf G K. Human Behavior and the Principle of Least Effort[M]. Cambridge: Addison-Wesley Press, 1949.

[14] Li Yujian, Xiao Chuangbai. Zipf’s Law Probably Existing in Protein Sequences[J]. Journal of Beijing University of Technology, 2005, 31(4): 366-368. (in Chinese)

[15] Zipf G K. Human Behavior and the Principle of Least Effort[M]. Cambridge: Addison-Wesley Press, 1949.

[16] Jian Wangqi. Zipf and the Principle of Least Effort[J]. Tongji University Journal(Social Science Section）, 2005, 16(1): 88-95. (in Chinese)