International Journal of Information Management 33 (2013) 752–763

Contents lists available at ScienceDirect

International Journal of Information Management

journal homepage: www.elsevier.com/locate/ijinfomgt

Understanding Chinese online users and their visits to websites: Application of Zipf's law

Qiqi Jiang[a], Chuan-Hoo Tan[b,*], Chee Wei Phang[c], Juliana Sutanto[d], Kwok-Kee Wei[e]

[a] Department of Information Systems, City University of Hong Kong, To Yuen Building, 83 Tat Chee Avenue, Kowloon Tong, Hong Kong
[b] Department of Information Systems, City University of Hong Kong, P7917, 7/F., Academic Building, 83 Tat Chee Avenue, Kowloon Tong, Kowloon, Hong Kong
[c] Information Management and Information Systems, Fudan University, Room 707, Siyuan Building, 670 Guoshun Road, Shanghai 200433, China
[d] Management Information Systems, ETH Zürich, SEC D7, Scheuchzerstrasse 7, 8092 Zürich, Switzerland
[e] Department of Information Systems, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Tong, Hong Kong

ARTICLE INFO

Article history:
Available online 21 June 2013

Keywords:
Zipf's law
Internet usage pattern
User demographics
Chinese Internet market

ABSTRACT

Competition for consumers to visit company websites has intensified in recent years. An important indicator of website popularity (and consequent survival) is the extent to which the website can draw consumer visits vis-à-vis other websites. A majority of the current understanding on consumer visits is limited to a single website, and leaves little knowledge on the performance of one website compared with others. In tracking the Internet usage behavior of 200 individuals in Mainland China for 30 consecutive days, we applied Zipf's law to identify the divergence points separating popular websites from non-popular ones. Two measurements were used, namely, visit traffic (number) and visit engagement (time spent). We observed that 94.87% of the entire visit traffic is devoted to 15.08% of all visited websites, whereas 84.63% of engagements are on the top 6.16% visited websites. These findings suggest that few websites accounted for the bulk of online traffic and time. Further, we segmented the dataset based on two key proxy variables of user demographics, which are gender and occupation. The findings on visit traffic remained salient after considering user segments, but the findings on website engagement varied across different user segments. Our further analysis, which categorized the visited websites by their main service, revealed the type of Internet users attracted to popular websites.

1. Introduction

With websites seen as impartial components of a company (Dias, 2001), the competition among companies for consumers to visit their websites has intensified in recent years. While some studies have examined consumer responses to a website based on their decision to adopt or to use it (e.g., Hansen, Jensen, & Solgaard, 2004; Lin & Lu, 2000), and other studies have assumed the Internet as a whole (e.g., Elliot & Fowell, 2000), none of these studies have considered how websites fair among themselves. Thus, an important indicator of website popularity is the extent to which it can draw consumer visits compared with other websites (Chu, Shen, & Hsia, 2004), or the extent to which it can engage visitors more than other websites could. This indicator can be measured based on the time spent by visitors on a particular website (Lee, Ungson, & Russo, 2011; O'Brien & Toms, 2012). Such knowledge is important to the entire Internet industry, which includes website operators, Internet business consultants, and consumers. Website operators view both consumer visit traffic and website engagement as precursors of sales opportunities. Internet business consultants, having an overarching knowledge of the consumer visit traffic and engagement distributions, foster good understanding of the development and trends in the online market. To the consumers, knowledge of popular websites provides a gist into the potential service quality of the website operators. This knowledge could also assist them in choosing websites to visit wisely among the huge number of websites today.

Despite the importance of gaining a good understanding of website popularity, significant gaps exist in extant literature. Majority of the current understanding on consumer traffic or website engagement is limited to a single website (Bonniface & Green, 2007; Cho & Kim, 2004; Lee, Podlaseck, Schonberg, & Hoch, 2001; Lee et al., 2011; Lin & Lu, 2000), or a handful of related websites (Ford, Huerta, Schilhavy, & Menachemi, 2012; Hanna, Rohm, & Crittenden, 2011). These studies leave little knowledge on how one website performs compared with others. Büchner and Mulvenna (1998) studied the incoming traffic log of a specific website to understand online customers as the basis for proposing marketing intelligence campaigns. Several studies, such as Krashakov, Teslyuk, and Shchur (2006) and Saxena, Sharan, and Fahmy (2008),

* Corresponding author. Tel.: +852 3442 9720; fax: +852 3442 0370.
E-mail address: ch.tan@cityu.edu.hk (C.-H. Tan).

0268-4012/$ – see front matter.
http://dx.doi.org/10.1016/j.ijinfomgt.2013.05.006

proposed approaches to assess website popularity, but these studies employed limited statistics. Other studies have investigated the patterns of connectivity among web pages, and found an apparent gap between websites having a large number of links (popular) and websites that did have as many (non-popular) (Albert, Jeong, and Barabasi, 1999). Although these studies did not focus on a particular website (Albert et al., 1999; Saxena et al., 2008), they did not specify how popular websites could be differentiated from non-popular ones. Bonniface and Green (2007) conducted several rounds of interviews to measure website engagement by discovering its different levels, but their study lacked objectivity. In addition, Coleman, Lieber, Mendelson, and Kurpius (2008) argued that a well-designed website can enhance visitor engagement, but such findings are too intuitive. Collectively, most studies used perceptual measurement to evaluate user engagement in specific websites, but such studies seemingly underestimated objective indicators, such as average time spent on the website in the real context.

In the next sections, we introduce Zipf's law, and apply it in our study through two ranking means, namely, ordinal and dense. The basic notion of Zipf's law is to understand the relationship between the frequency and the rank order (Zipf, 1949) by which the leading items within the overall population could be located, such as words in the language (Zipf, 1935, 1949) and population of metropolitan areas (Strogatz, 2009). In the context of Internet usage, we can utilize such paradigm to identify the leading websites from two dimensions, namely website traffic and website engagement, with two approaches, namely "ordinal rank" and "dense rank."

The approach to the analysis was empirically validated by tracking 30 days of Internet usage of 200 users in Mainland China, which were drawn from the major Internet user population. Based on our analysis, two findings were deduced regarding website visit traffic (number) and website visit engagement (time spent). In particular, 94.87% of the entire visit traffic was contributed by 15.08% of all visited websites, and 84.63% of the visiting time was spent on 6.16% of all the visited websites. Further, we segmented the data based on two key proxy variables of user demographics, which are gender and occupation. These two demographics are generally believed to influence online user behavior significantly (e.g., Bartel Sheehan, 1999; Teo, 1998; Wolin & Korgaonkar, 2003; Yiu, Grant, & Edgar, 2007). The results of segmenting the dataset according to the demographic variables showed that the discrimination between popular and unpopular websites remained effective for visit traffic but not for visit engagement. We also categorized all the visited websites according to their main service through several rounds of sorting to gain holistic understanding. The results on multi-dimensional arrangement of websites and their popularity revealed the type of Internet users attracted to popular websites.

This study has several important contributions. First, this research is the first to apply Zipf's law to understand the visit activities of Internet users. The application of the theory also extends it by incorporating ranking techniques to differentiate between popular and non-popular websites. Second, this research considers the entire Internet ecosphere through categorizing various websites into seven categories, and analyzing their relative competiveness in terms of visit traffic and visit engagement holistically. Such analysis provides a more complete understanding of consumer visits across the nature of websites. Third, by considering both visit traffic and visit engagement, this research suggests to companies how to identify websites that are mostly visited or stayed on, and thus, leads to more appropriate marketing strategies.

2. Zipf's law

Zipf's law builds on a fundamental premise that many types of data could follow a Zipf-like distribution (Zipf, 1949). Several studies demonstrated the significance of Zipf's law in explaining various phenomena in different areas. For instance, research has demonstrated Zipf-like distribution forcity size or population in urban planning (Gabaix, 1999a,b; Hill, 1974; Ye & Xie, 2012), biological taxonomies and cancer genes in medicine (Chiu, Hsieh, & Wang, 2012; Li & Yang, 2002), fracture process in geometry (Schroeder, 2009), English sentence composition in linguistics and in information science (Edwards & Collins, 2011), and firm size in organizational studies (Di Giovanni & Levchenko, 2012). With few exceptions (e.g., study of Hill, 1974), these studies primarily focused on demonstrating Zipf-like distribution in various areas of interest. Our study applies Zipf's law as a way to distinguish popular from non-popular websites.

Hill (1974) extended previous studies by analytically deriving Zipf's law from the Bose–Einstein allocation of categories. Specifically, Hill showed that if a country was divided into K regions with almost equal values of logarithmic population, and the largest city was taken from each region, then the ordered values of the population of the largest city should be subjected to a curve of Zipf's law. The Bose–Einstein allocation of people to cities occurred within each of these divided regions. Deducing that observation from other scientific fields, the regions can be seen as different industries by describing personal income distribution, or as different chapters of a book when investigating the frequency of word occurrence. In our research context (website visit traffic and visit engagement), the available categorized regions are not apparent. Thus, to adopt faithfully the proposed model of Hill (1974) in our work will be less plausible. Our study provides empirical validation (as opposed to the analytical deduction) of the application of Zipf's law.

Despite its limitations, Zipf's law has been recently applied to understand web-related behavior of consumers. For instance, Breslau, Cao, Fan, Phillips, & Shenker (1999) applied Zipf's law to understand webpage requests (a user issues a browsing request for a specific webpage), and observed a correlation between the asymptotic properties of the webpage request ratio and the request inter-arrival time. Taking a user perspective, Yamakami (2006) adopted Zipf's law to compare the Internet browsing activities between users of mobile and personal computer devices, and found no significant difference between the two groups. These studies, however, did not apply Zipf's law to understand the popularity of websites in terms of consumer visit and their visit engagement.

Zipf-like distribution, belonging to the family of discrete power law probability distributions, states that the frequency of any unit, symbol, or element is inversely related to its rank in the frequency table, which can be visualized by plotting the data on a graph with the y-axis represented by *ln(frequency)* and the *x*-axis represented by *ln(rank)* (Zipf, 1949). In the context of our study, the target data in the visited websites, *frequency*, would refer to the number of clicks per site (visit traffic) or average time spent per site (i.e., visit engagement), while *rank* would be denoted by the ranking of this site in the whole sample. To illustrate the distribution, we can imagine that the most frequently visited or the longest time stayed in a website, as reflected by website visit traffic and engagement, respectively, could attract traffic or visit duration that are twice the traffic of the second most popular website, and three times the traffic of the third most popular website, and so on.

Owing to the ranking of the frequency of occurrence, Zipf's law is also referred to as frequency distribution of the "ranking data." Equation (1) depicts Zipf's law:

$$F(r) = \frac{1}{r} \tag{1}$$

F denotes the frequency function of the occurrence of a symbol or element (e.g., the English word "apple" in linguistics), and *r* refers to the ranking. In log-log scales, Zipf's law gives a linear line with slope = −1. The Zipf-like distribution can be written as Eq. (2). In