Strong Regularities in World Wide Web Surfing

Bernardo A. Huberman, Peter L. T. Pirolli, James E. Pitkow and Rajan M. Lukose
Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304

Abstract

One of the most common modes of accessing information in the World Wide Web (WWW) is surfing from one document to another along hyperlinks. Several large empirical studies have revealed common patterns of surfing behavior. A model which assumes that users make a sequence of decisions to proceed to another page, continuing as long as the value of the current page exceeds some threshold, yields the probability distribution for the number of pages, or depth, that a user visits within a Web site. This model was verified by comparing its predictions with detailed measurements of surfing patterns. It also explains the observed Zipf-like distributions in page hits observed at WWW sites.

The exponential growth of World Wide Web (WWW) is making it the standard information system for an increasing segment of the world's population. From electronic commerce and information resource to entertainment, the Web allows inexpensive and fast access to unique and novel services provided by individuals and institutions scattered throughout the world (1).

In spite of the advantages of this new medium, there are a number of ways in which the Internet still fails to serve the needs of the user community. Surveys of WWW users find that slow access and inability to find relevant information are the two most frequently reported problems (2). The slow access has to do at least partly with congestion problems (3) whereas the difficulty in finding useful information is related to the balkanization of the Web structure (4). Since it is hard to solve this fragmentation problem by designing an effective and efficient classification scheme, an alternative approach is to seek regularities in user patterns that can then be used to develop technologies for increasing the density of relevant data for users.

A common way of finding information on the WWW is through query-based search engines, which allow for quick access to information that is often not the most relevant. This lack of relevance is partly due to the impossibility of cataloguing an exponentially growing amount of information in ways that anticipate users' needs. But since the WWW is structured as a hypermedia system, in which documents are linked to one another by authors, it also supports an alternative and effective mode of use, one in which users surf from one document to another along hypermedia links that appear relevant to their interests.

In what follows we describe several strong regularities of WWW user surfing patterns discovered through extensive empirical studies using different user communities. These regularities can be described by a law of surfing, derived below, which determines the probability distribution of the number of pages a user visits within a Web site. In conjunction with a spreading activation algorithm, the law can be used to simulate the surfing patterns of users on a given Web topology. This leads to accurate predictions of page hits. Moreover, it explains the observed Zipf-like distributions of page hits to WWW sites (5).

We start by deriving the probability *P(L)* of the number of links *L* that a user follows in a Web site. This can be done by considering that there is value in each page a user visits, and that clicking on the next page assumes that it will be valuable as well. Since the value of the next page is not certain, one can assume that it is stochastically related to the previous one. In other words, the value of the current page is the value of the previous one plus or minus a random term. Thus, the page values can be written as

$$V_L = V_{L-1} + \varepsilon_L \qquad (1)$$

where the values ε_L are independent and identically distributed Gaussian random variables. Notice that a particular sequence of page valuations is a realization of a random process and so is different for each user. Within this formulation, an individual will continue to surf until the expected cost of continuing is perceived to be larger than the discounted expected value of the information to be found in the future. This can be thought of as a real option in financial economics, for which it is well known that there is a threshold value for exercising the option to continue (6,7). Note that even if the value of the current page is negative, it may be worthwhile to proceed, since a collection of high value pages may still be found. If the value is sufficiently negative, however, then it is no longer worth the risk to continue. That is, when V_L falls below some threshold value, it is optimal to stop.

The number of links a user follows before the page value first reaches the stopping threshold is a random variable *L*. For the random walk of Eq. 1 the probability distribution of first passage times to a threshold is given asymptotically by the two parameter inverse Gaussian distribution (8)

$$P(L) = \sqrt{\frac{\lambda}{2\pi L^3}} \exp\left(\frac{-\lambda (L-\mu)^2}{2\mu^2 L} \right) \qquad (2)$$

with mean $E[L] = \mu$ and variance $Var[L] = \mu^3 / \lambda$.

This distribution has two characteristics worth stressing in the context of user surfing patterns. First, it has a very long tail, which extends much further than that of a normal distribution with comparable mean and variance. This implies a finite probability for events that would be unlikely if described by a normal distribution. Consequently, large deviations from the average number of user clicks computed at a site will be observed. Second, because of the asymmetry of the distribution function, the typical behavior of users will not be the same as their average behavior. Thus, since the mode is lower than the mean, care has to be exercised with available data on the average number of clicks, as it overestimates the typical depth being surfed.

In order to test the validity of Eq. 2, we performed an analysis of data collected from a representative sample of America Online (AOL) WWW users. For each day of November 29, 30, and December 1, 3, and 5, 1997, the entire activity of one of AOL's caching-proxies was instrumented to record an anonymous but unique user identifier, the time of each URL request, and the requested URL. For each day in the AOL sample, there were between 3,247,054 and 9,120,199 requests for Web pages. To compare with the predicted distribution, a user that starts surfing at a particular site, such as http://www.sciencemag.org/, is said to have stopped surfing after *L* links as soon as she requests a page from a different Web site. For this analysis, if the user later returned to that site a new length count *L* was started. Requests for embedded media such as images were not counted.

On December 5, 1997, the 23,692 AOL users in our sample made 3,247,054 page requests from 1,090,168 Web sites. Fig. 1 shows the measured Cumulative Distribution Function (CDF) of the click length *L* for that day. Superimposed is the predicted one from the inverse Gaussian distribution fitted by the method of moments (8). To test the quality of the fit, a Quantile-Quantile was analyzed against the fitted distribution. Both techniques, along with a study of the regression residuals confirmed the strong fit of the empirical data to the theoretical distribution. The fit was significant at the $p < 0.001$ level and accounted for 99% of the variance. Notice that while the average number of pages

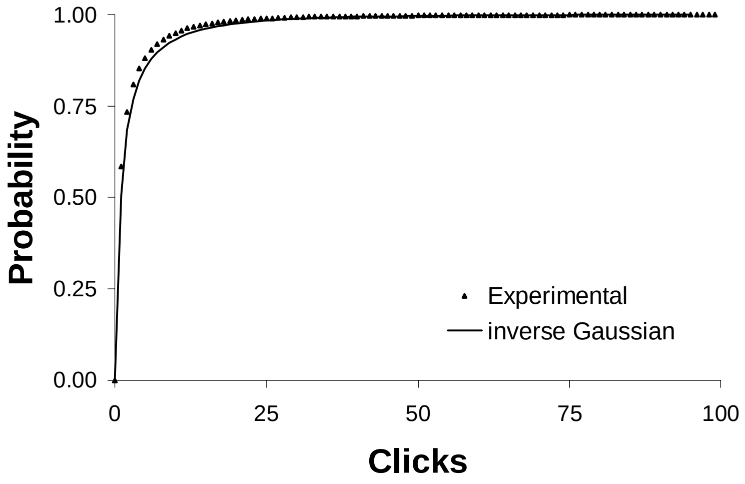

Figure 1. The Cumulative Distribution Function of AOL users as a function of the number of clicks surfing. The observed data were collected on December 5, 1997 from a representative sample of 23,692 AOL users who made 3,247,054 clicks. The fitted inverse Gaussian distribution has a mean of $\mu = 2.98$ and $\lambda = 6.24$.

surfed at a site is almost three, typically users only requests one page. Other AOL data from different dates showed the same strength of fit to the inverse Gaussian with nearly the same parameters.

For further confirmation of the model, we considered the simplest alternative hypothesis, in which a user at each page simply conducts an independent Bernoulli trial to make a stopping decision. This leads to a geometric distribution of click lengths, which was found to be a poor fit to the data.

We also examined the navigational patterns of the Web user population at Georgia Institute of Technology for a period of three weeks starting August 3, 1994. The data were collected from an instrumented version of NCSA's Xmosaic that was deployed across the student, faculty, and staff of the College of Computing (9). One hundred and seven users (67% of those) invited chose to participate in the experiment. The instrumentation of Xmosaic recorded all user interface events. Seventy three percent of all collected events were navigational, resulting in 31,134 pages requests. As with the previous experiment, the surfing depth of users was calculated across all visits to each site for the duration of the study. The combined data has a mean number of clicks of 8.32 and a variance of 2.77. Comparison of the Quantile-Quantile, CDF and a regression analysis of the observed data against an inverse Gaussian distribution of same mean and variance confirmed the ability of the law of surfing to fit the data (R^2 of 0.95, $p < 0.001$). It is important to note that the model is able to fit surfing behavior using data sets from diverse user communities, at dramatically different time periods, using different browsers, and connection speeds.

An interesting implication of the law of surfing can be obtained by taking logarithms on both sides of Eq. 2. One obtains

$$l\log P(L) = -\frac{3}{2}l\log L - \frac{\lambda(L-\mu)^2}{2\mu^2 L} + l\log\left(\sqrt{\frac{\lambda}{2\pi}}\right) \qquad (3)$$

That is, on log-log plot one observes a straight line whose slope approximates 3/2 for small values of *L* and large values of the variance. As *L* gets larger, the second term provides a downward correction. Thus Eq. 3 implies that, up to a constant given by the third term, the probability of finding a group surfing at a given level scales inversely in proportion to its depth, $P(L)\alpha L^{-3/2}$. This Pareto scaling relation was verified by plotting the available data on a logarithmic scale. Fig. 2 shows that for a range of click lengths the inverse proportionality holds well.

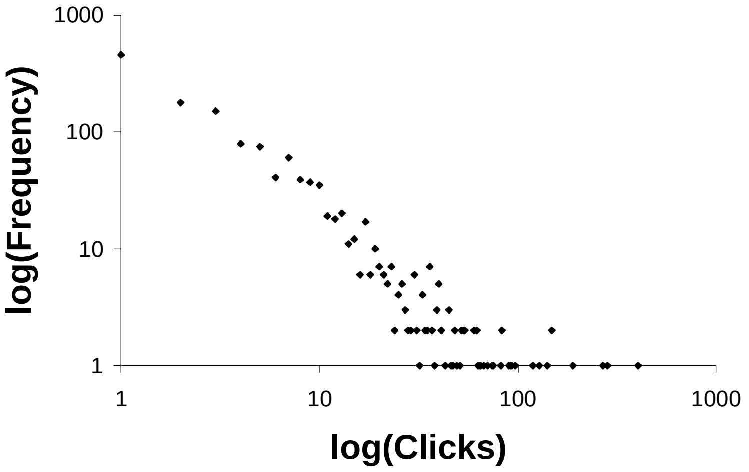

Figure 2. Figure 2. The frequency distribution of surfing clicks on log-log scales. Data collected from the Georgia Institute of Technology, August 1994.

The previous data validated the law of surfing for a population of users who had no constraints on the Web sites they visited. We also considered the case of surfing within a single large Web site, which is important from the point of view of site design. The site used was the Xerox Corporation's external WWW site (http://www.xerox.com). During the period of August 23 through August 30, 1997, the Xerox site consisted of 8,432 HTML documents and received an average of 165,922 requests per day. The paths of individual users were reconstructed by a set of heuristics that used unique identifiers, i.e., cookies, when present, or otherwise used the topology of the site along with other information to disambiguate users behind proxies. Automatic programs that request the entire contents of the site, a.k.a. spiders, were removed from the analysis. Additionally, a stack-based history mechanism was used to infer pages cached either by the client or by intermediary caches. This resulted in a data set consisting in the full path of users and the number of clicks performed at the Xerox Web site.

Fig. 3 shows the Cumulative Distribution Function plot of the Xerox WWW site for August 26, 1997 against the fitted inverse Gaussian defined by Eq. 2. The mean number of clicks was 3.86, with a variance of 6.08 and a maximum of 95 clicks. As with the client path distributions, both the Quantile-Quantile and the CDF plots of the site data showed a strong fit to Eq. 2. Moreover, these results were very consistent across all the days in the study.

Having shown that Eq. 2 is a good description of user surfing patterns, we show how, in conjunction with a spreading activation algorithm, it can predict the number of hits for each page in a Web site, a quantity of interest in electronic commerce. Spreading activation refers to a class of algorithms that propagate numerical values (or activation levels) among the connected nodes of a graph (10). Their application ranges from models of human memory(11) and semantics (12), to information retrieval (13). In the context of the Web, the nodes correspond to pages and the arcs to the hyperlinks among them, so that spreading activation simulates the flow of users through a WWW site.

Consider a collection of n Web pages, each indexed by i = 1, 2,..., n, and connected by hyperlink edges to form a graph. One can simulate the surfing activity of users by assigning a weight, $S_{j,i}$, between the i^{th} and j^{th} node. This weight can be interpreted as the fraction of continuing users at node i who proceed to node j if $\sum_{j} S_{j,i} = 1$, where the sum

$$f_L = \frac{1 - F(L, \mu, \lambda)}{1 - F(L - 1, \mu, \lambda)} \tag{5}$$

With this definition, Eq. 4 can be iterated from initial conditions $N_{i,1}$. After most of the surfers have stopped, the predicted aggregate number of hits at each page is simply the sum over all iterations for each page.

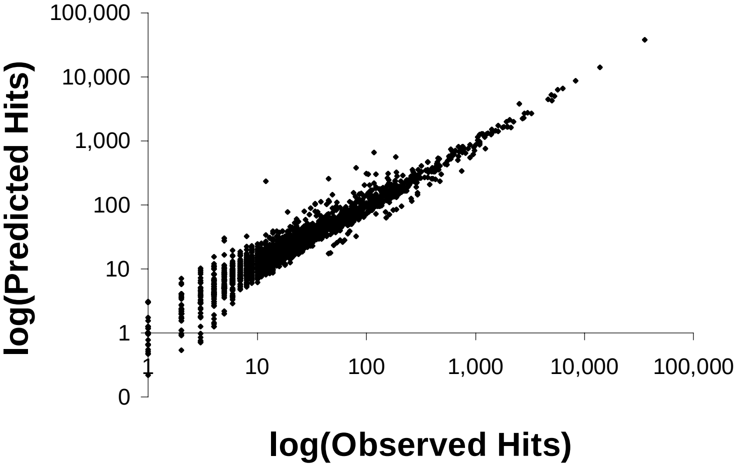

Figure 4. Histogram of the predicted number of visits per page (hits) to the Xerox Web site versus the observed number of visits generated by spreading activation simulations on log-log scale.

Fig. 4 shows the observed and predicted daily average number of hits per page for the Xerox corporate WWW site data described above. Eq. 4 was initialized with estimates from the data of the number of users who started surfing at each page *i*, $N_{i,1}$, and the proportion of users who surf from each page *i* to connected pages *j*, $S_{i,j}$. We used the inverse Gaussian estimated in Fig. 3 for the Xerox site to compute f_L and iterated Eq. 4 for L = 15 levels of surfing. Fig. 4 shows that the hits predicted by spreading activation are highly correlated with the observed hits, with r = .995.

This algorithm can also be used for a number of interesting Web applications. For example, if a Web site were to be reorganized, spreading activation plus the law of surfing could give an indication of the expected usage. Alternatively, one may be able to automatically reorganize the Web site structure in order to obtain a desired hit pattern.

We also used a spreading activation model to address another universal finding in studies of WWW activity, that of a Zipf's like distribution in the number of hits per page. We ran spreading activation simulations on random graphs of 100 nodes each, with an average of five links per node, using various initial conditions. The resulting probability distribution of the number of hits received over the collection of pages followed a Zipf's law, in agreement with observed data (5).

These results show that surfing patterns on the World Wide Web display strong statistical regularities that can be described by a universal law. In addition, the success of the model points to the existence of utility maximizing behavior underlying surfing. Because of the Web's digital nature and great use, it is relatively easy to obtain online data that could reveal more novel patterns of information foraging. For example, one could extend these studies to determine the relationship between the characteristics of different user communities and the law of surfing parameters.

As the world becomes increasingly connected by the Internet, the discovery of new patterns in the use of the WWW can throw a timely light on the growth and development of this new medium. This is particularly important since the sheer reach and structural complexity of the Web makes it an ecology of knowledge, with relationships, information "food" chains, and dynamic interactions that could soon become as rich, if not richer, than many natural ecosystems.

Acknowledgements

Parts of this research were supported by National Science Foundation grant IRI-961511 to B. A. Huberman, the Office of Naval Research grant N00014-96-C-0097 to P. Pirolli and S. Card, and a Citibank fellowship at Stanford University for R. M. Lukose. We also thank Mark Crovella for his comments on a draft of this paper and Dan Aronson for providing the AOL data.

References

1. Special issue on the Internet, *Scientific American* **276** (1997).
2. J. E. Pitkow, C. M. Kehoe, *Online Publication: http://www.gvu.gatech.edu/user_surveys* (1997).
3. B. A. Huberman, R. M. Lukose, *Science* **277**, 535 (1997).
4. M. Van Alstyne, E. Brynjolfsson, *Science* **274**, 1479 (1996).
5. S. Glassman, *Computer Networks and ISDN Systems* **27**, 165-173 (1994).
6. A. K. Dixit and R. S. Pindyck, Investment under Uncertainty, Princeton University Press (1994).
7. R. M. Lukose and B. A. Huberman, Proceedings of the Fourth International Conference on Computational Economics (1998).
8. V. Seshardri, The Inverse Gaussian Distribution, Clarendon Press (1993).
9. L. Catledge, J. Pitkow, *Computer Networks and ISDN Systems* **27** (1995).
10. J. Shrager, T. Hogg, B. A. Huberman, *Science 236*, 1092-1094 (1987).
11. J. R. Anderson, P. L. Pirolli, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **10**, 791-798 (1984).
12. M. R. Quillan, *Semantic memory* (Bolt, Beranek, and Newman, Cambridge, MA, 1966).
13. P. Pirolli, J. Pitkow, R. Rao, Silk from a sow's ear: Extracting usable structures from the web, Conference on Human Factors in Computing Systems, CHI '96, Vancouver, Canada (1996).
14. G. K. Zipf, *Human behavior and the principle of least effort* (Addison-Wesley, Cambridge, MA, 1949).