Isola Ajiferuke
Faculty of Information and Media Studies, University of Western Ontario
London, Ontario

Dietmar Wolfram & Hong (Iris) Xie
School of Information Studies, University of Wisconsin-Milwaukee
Milwaukee, WI U.S.A.

Modelling Website Visitation and Resource Usage Characteristics by IP Address Data

Abstract: Two large data sets representing site visitation data based on IP addresses, and resource request frequency for EBSCOhost interactions accessed through a state digital library service were fitted to different mathematical models used in informetrics. Results reveal that a Zipf model provided the best fit for the site visitation data and a generalized logarithmic series model provided the best fit for the resource request data.

1. INTRODUCTION

Informetric studies of observed patterns of Internet document content and usage have become common in recent years with the wider availability of user data collected by websites. Regularities in the occurrence of text or usage of search systems in electronic environments have been found to exhibit similar regularities as for print media, where frequency distributions of data sets result in a large number of rare events (LNRE) with lengthy tails (Baayen, 2001). With Internet/Web-based phenomena, these regularities are frequently concluded to be Zipfian based on a visual inspection of a trend line of logarithmically transformed data. However, a simple Zipf distribution may not adequately model the observed behaviour for electronic environment characteristics (Nelson, 1989; Ajiferuke & Wolfram, 2004), particularly for highly skewed, larger data sets representing potentially millions of observations.

Research investigations of website characteristics can be divided into two broad categories: 1) system studies based on regularities in domain/document content and their implications for efficient storage and retrieval, and; 2) user and usage studies that focus on query and session analysis or resource utilization. Literature discussed in the present study focuses on the latter, where the processes studied are at the Internet Protocol (IP) address and document level.

Studies of Web-based service usage have appeared with increasingly regularity in a variety of information retrieval (IR) contexts. Log analysis has been identified as one of the effective and unique approaches for studying usage patterns on a large scale, and it has been conducted to characterize the use of Web-based online public access catalogs(OPACs) of library holdings (Cooper, 2001). The emergence of digital libraries led to more log analysis of not only the usage of digital libraries themselves (Jones, Cunningham, McNab, & Boddie, 2000) but also of Web-based online databases. These have included vendor-based comprehensive bibliographic and full text databases (Wolfram & Xie, 2000), one of the largest and most heavily used full text e-journal systems (Ke, Kwakkelaar, Tai, & Chen, 2002), as well as specialized domain databases

(Jantz, 2003). Log analysis has also been widely used in Web search engine studies to analyze query reformulations and resource utilization, such as AltaVista (Silverstein, Henzinger, Marais, & Moricz, 1999), Excite (Jansen, Spink, & Saracevic, . 2000; Spink, Wolfram, Jansen, & Saracevic, 2001), Fireball (Hoelscher, 1998) and Intranet search engines (Fitchter, 2003). In addition, longitudinal transaction log analysis has been conducted to detect Web information searching patterns(Cothey, 2002). Most of these studies focus on characterizing users' information seeking behaviour and making suggestions to improve the existing IR systems.

Because it is difficult to identify individuals using the Internet, proxies for individual users have taken the form of login information, cookies stored on a computer, or originating IP address data. IP address data to study resource usage represents an easily observed way track and tally usage behaviours with fewer concerns regarding privacy associated with cookies or direct login information. Davis and Solla (2003), for example, relied on IP address data to study usage statistics of electronic journals in chemistry. The authors found there was a strong relationship between the number of articles downloaded and the number of users, thereby making it possible to estimate the total user population based on the number of downloads.

The study of resource usage has become especially popular for market analysis of websites. Resource usage can also have implications for caching of frequently accessed documents for improved response time (Cunha, Bestavros, & Crovella, 1995). Early reports of regularities observed in user and usage behaviour have been largely descriptive. A number of studies have gone one step further by reporting that frequency distributions of observed data sets follow inverse power laws, or conform to Zipf's law (Broder, et al., 2000; Huberman, 2001; Jansen, Spink, & Saracevic, 2000; Nielsen, 1997a, 1997b; Spink, et al., 2001). For example, Nielsen (1997b) noted that the rank-frequency distribution of hits per month distribution for pages on the Sun Microsystems Web site was largely Zipfian, at least for the most frequently accessed pages. The lower than expected number of hits for low ranking pages was attributed to the paucity of accumulated pages of low-frequency interest. Crovella, Taqqu, and Bestravos (1998) also noted that the distribution of document accesses could be modelled by a power law, citing implications of this finding for caching.

In most cases power law models have been cited as fitting observed data sets based on the conformance of log/log plots of data sets against straight lines with no determination of the goodness-of-fit beyond visual inspection. However, as noted above, observed data sets may deviate noticeably from a straight line, particularly at the low end of the distribution. Because the majority of an observed curve appears to follow a straight line, it is concluded that the data set as a whole follows an inverse power law. Studies investigating the fitting of observed electronic data characteristics to other types of theoretical distributions used in informetrics, along with more rigourous goodness-of-fit testing to provide more accurate models of the observed characteristics, are needed. The present research explores whether a Zipf model adequately fits large observed data sets. More broadly, this study investigates the ability of a number of theoretical distributions used in informetric research to model the observed frequency distributions for site visitation and resource requests associated with a high traffic public website using data collected from transaction logs.

2. METHOD

Transaction log data were collected from Wisconsin's BadgerLink service. BadgerLink is a state-funded Web-based information service, providing access to a range of electronic information resources to residents of Wisconsin, including library catalog information, access to full-text and bibliographic databases through service providers such as EBSCO (EBSCOhost) and ProQuest, and publicly available Web-based resources (Wolfram & Xie, 2000). Two large data sets were extracted from BadgerLink's EBSCOhost data tracking feature for the time period January through June 2001. One data set consisted of site visitations based on known IP addresses, representing over 800,000 visits by 78,000 distinct IP addresses. The second data set consisted of more than 3.6 million resource requests for over 7,100 titles indexed by EBSCO. These data sets represent different points of reference for the same processes (Figure 1). Raw data were then tabulated into frequency distributions for site visitation and resource requests (Figures 2 and 3 respectively).

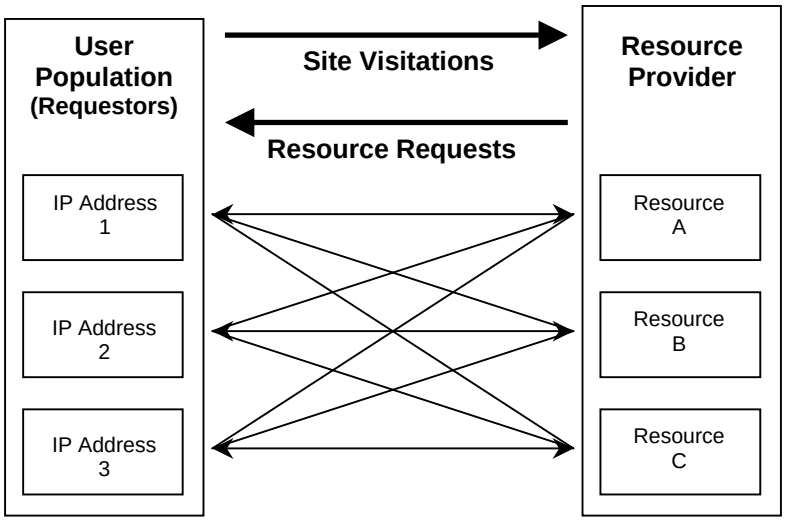

Figure 1. Data Relationships

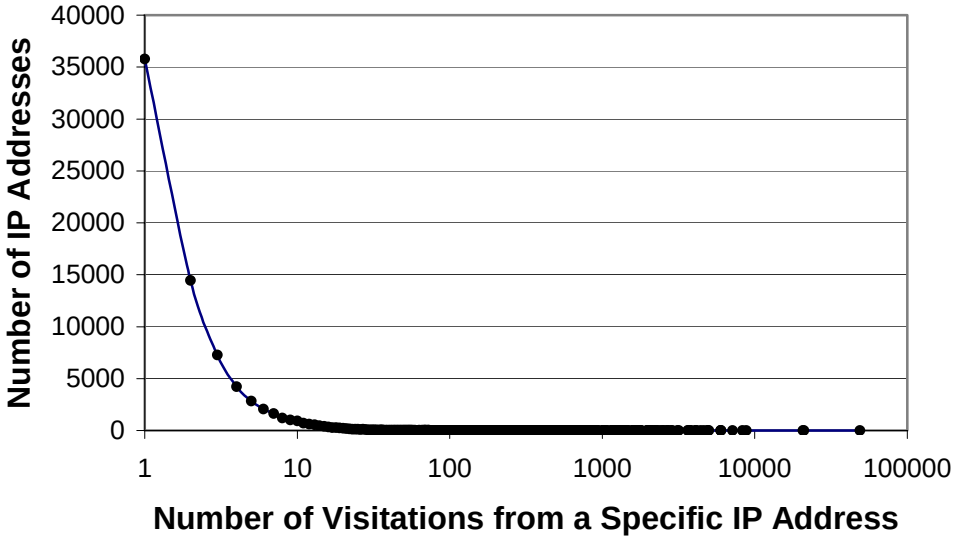

Figure 2. Distribution of site visitation frequency by identifiable IP addresses

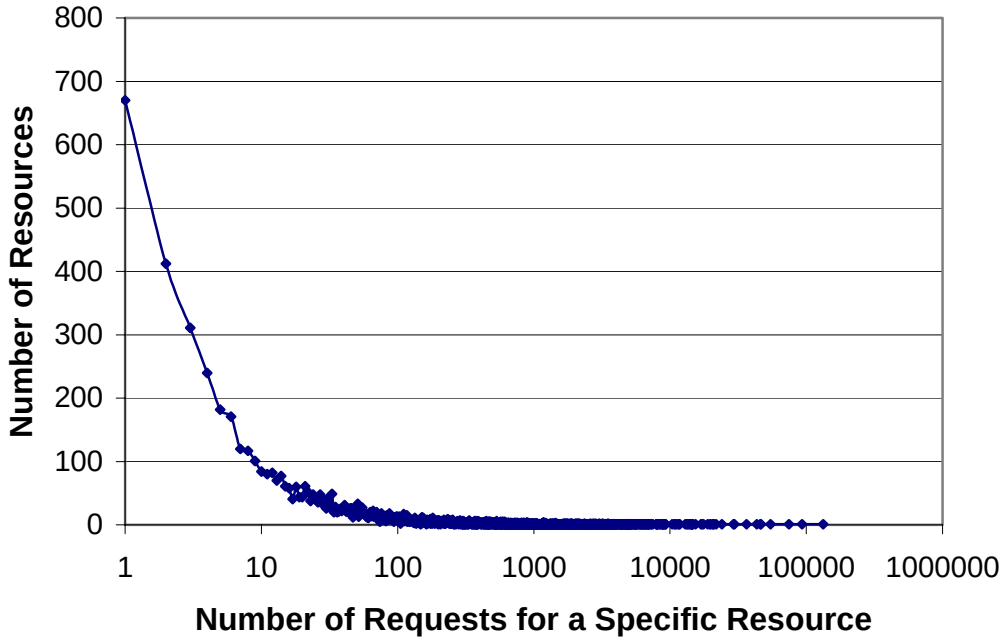

Figure 3. Distribution of resource request frequencies

The resulting frequency distributions were fitted to a number of theoretical models used in informetric research. Models tested included: Zipf, Mandelbrot-Zipf, Yule-Simon, geometric, Borel-Tanner, beta binomial, Geeta, Consul, generalized logarithmic series, generalized Waring, generalized Poisson, generalized negative binomial, and generalized inverse Gaussian-Poisson. Some of these models have a natural origin of one (e.g., Zipf) whereas others have an origin of zero (e.g., generalized negative binomial). For each of the models of the latter type, the zero-truncated version was tested, where each outcome of f(x) for x greater than zero is divided by 1- f(0). The Geeta, Consul, and generalized logarithmic series distributions have not been widely used in informetric study, but have properties similar to inverse power models such as Zipf, with a starting value of one and an ability to model long tails.

Due to the large value of either the highest number of visits or the highest number of uses, it was difficult to obtain the maximum likelihood or minimum chi-square estimates of the parameters for a number of the models, whereas the moment estimates led to poor fitting due to the large variability in the data. Hence, only the results for the size-frequency form of the Zipf (the rank-frequency form was not tested), generalized inverse Gaussian-Poisson and generalized logarithmic series that provided some reasonable fits to the data sets are reported here. Their functional forms are:

(i) Size-frequency Zipf (ZIPF)

$$p(x) = \alpha x^{-\beta}$$

where $0 < \alpha < 1$, $\beta > 0$ for $x = 1, 2, \ldots$

(ii) Zero–truncated Generalized Inverse Gaussian-Poisson (GIGP) (Sichel, 1992)

$$p(x) = \frac{(1-\theta)^{\gamma/2}}{K_\gamma\{\alpha\,(1-\theta)^{1/2}\}} \frac{(\tfrac{1}{2}\alpha\theta)^x}{x!} K_{x+\gamma}(\alpha) * \left[1 - \frac{(1-\theta)^{\gamma/2}}{K_\gamma\{\alpha\,(1-\theta)^{1/2}\}} K_\gamma(\alpha)\right]^{-1}$$

where $-\infty < \gamma < \infty$, $0 \le \theta \le 1$, $\alpha \ge 0$ for x = 1, 2,…, and $K_\nu\{z\}$ is the modified Bessel function of the second kind of order ν and argument z

(iii) Generalized Logarithmic Series (GLS) (Famoye, 1997)

$$p(x) = \frac{\Gamma(\beta x+1)\ \alpha^x\ (1-\alpha)^{\beta x-x}}{x!\ (\beta x)\ \Gamma(\beta x-x+1)\ [-\log(1-\alpha)]}$$

where $0 < \alpha < 1$, $1 < \beta < 1/\alpha$ for x = 1, 2,… and $\Gamma(x)$ represents the gamma function

After the parameters were estimated for each theoretical model, goodness-of-fit assessment between the observed and fitted theoretical distributions was carried out. Significant departures between observed and predicted values may be determined by applying a chi-square test. Another test used to measure goodness-of-fit, the Kolmogorov-Smirnov (K-S) test, is also sometimes employed in informetric modelling studies. However, because the K-S test is intended for use with continuous data in cumulative form, it is inappropriate for the present data sets, which represent discrete data in non-cumulative form. Also, the K-S test should be used only when the hypothesized distribution is completely specified. In situations where parameters are to be estimated from the data, the test has to be modified. Different critical values must be obtained for each hypothesized distribution using Monte Carlo simulation techniques (NIST/SEMATECH e-Handbook of Statistical Methods, Available at:, http://www.itl.nist.gov/div898/handbook/eda/section3/ eda35g.htm).

3. FINDINGS

The model fitting outcomes (expected frequencies) for the theoretical distributions tested appear in Table 1 for the site visitation data and Table 2 for the resource request data. In the case of the distribution of IP address visits, the Zipf model appears to have provided the best fit and, while it provided fairly good fits to the top part of the distribution (i.e. lower x-axis values representing the least frequently occurring contributors), the fits for the middle and tail parts are very good (see Table 1). For the distribution of resources usage, the GLS and Zipf models provided very good fits, especially to the middle and tail parts (see Table 2). The GLS model, however, results in a better fit. The fits in all cases resulted in significant outcomes, where chi-square outcomes exceeded critical values at the .01 level.

Number of Visits	Observed Frequency	Exp. Frequency (ZIPF)	Exp. Frequency (GIGP)	Exp. Frequency (GLS)
1	35777	39826.29	35777.06	31876.50
2	14458	13132.66	11805.79	11672.87
3	7280	6523.28	6021.11	6410.77
4	4215	3896.90	3767.32	4172.66
5	2825	2589.86	2631.39	2983.75
6	2067	1845.60	1967.39	2265.18
7	1618	1381.71	1540.45	1792.46
8	1192	1073.11	1247.10	1462.22
9	1012	857.48	1035.44	1220.94
10	912	700.89	876.88	1039.40
11	694	583.59	754.51	896.41
12	607	493.46	657.77	783.44
13	548	422.70	579.74	691.83
14	475	366.15	515.73	616.34
15	389	320.23	462.47	553.29
16	342	282.43	417.59	499.50
17	279	250.96	379.37	454.48
18	267	224.47	346.51	415.24
19	243	201.96	318.00	381.14
20	218	182.67	293.10	351.29
21	179	166.02	271.20	324.99
22	156	151.55	251.81	301.68
23	136	138.89	234.56	280.91
24	126	127.76	219.12	262.31
25	137	117.91	205.25	245.58
26	107	109.15	192.73	230.46
27	108	101.34	181.38	216.75
28	82	94.34	171.07	204.28
29	76	88.04	161.65	192.89
30	64	82.35	153.03	182.46
31-50	797	968.01	1960.28	2315.99
51 – 100	634	741.75	1779.53	2020.85
101 - 200	308	377.85	1004.73	1030.03
201 - 48936	479	385.64	625.94	458.62
Mean	10.51			
St. dev.	227.71			
α	-	0.505365	2712.193	0.064482
β	-	1.6006	-	14.082178
θ	-	-	0.999634	-
γ	-	-	-0.507359	-
χ^2	-	1115.867	3388.835	4997.032
d.f.	-	31	30	31
$P(\chi^2)$	-	1.0E-214	0.000000	0.000000

Table 1: Models fit to the IP Address Site Visitation data

Number of Visits	Observed Frequency	Exp. Frequency (ZIPF)	Exp. Frequency (GIGP)	Exp. Frequency (GLS)
1	670	1111.10	670.00	924.56
2	412	551.47	630.26	461.62
3	311	365.17	528.37	307.33
4	240	272.18	430.67	230.21
5	182	216.47	352.08	183.94
6	171	179.39	291.61	153.09
7	120	152.95	245.24	131.07
8	117	133.15	209.29	114.55
9	101	117.77	180.99	101.71
10	84	105.49	158.34	91.44
11	80	95.45	139.94	83.03
12	82	87.10	124.78	76.03
13	70	80.05	112.13	70.11
14	77	74.00	101.45	65.03
15	61	68.78	92.35	60.63
16	58	64.22	84.52	56.79
17	41	60.20	77.73	53.39
18	60	56.63	71.80	50.37
19	44	53.44	66.58	47.68
20	44	50.57	61.96	45.25
21	61	47.98	57.84	43.05
22	52	45.63	54.17	41.05
23	38	43.49	50.86	39.23
24	48	41.52	47.88	37.56
25	41	39.72	45.17	36.02
26	36	38.06	42.72	34.61
27	48	36.52	40.47	33.29
28	39	35.10	38.42	32.07
29	30	33.78	36.53	30.94
30	26	32.54	34.79	29.88
31-50	494	477.80	467.59	448.17
51 – 100	686	598.70	484.91	595.72
101 - 200	589	552.47	350.64	568.76
201 - 133045	1968	1262.11	798.92	1902.79
Mean	504.24			
St. dev.	2819.98			
α	-	0.154728	32.88761	0.999571
β	-	1.0115	-	1.000174
θ	-	-	0.999931	-
γ	-	-	-0.495576	-
χ^2	-	680.995	2636.977	131.373
d.f.	-	31	30	31
$P(\chi^2)$	-	2.2E-123	0.000000	2.56E-14

Table 2: Models fit to the Resource Requests data

4. DISCUSSION

The significant chi-square outcomes for each fitted model should not be seen as a failure of the theoretical models, but rather a product of the sizes of the data sets used. Chi-square goodness-of-fit cell values can become large due to the calculation method involving the squaring of the difference between observed and expected values divided by the expected value. The resulting chi-square value per cell becomes disproportionately higher for cells with large values. Also, with a large number of observation classes, the cumulative differences between observed and expected values eventually result in a significant outcome even with a higher number of degrees of freedom and a higher critical chi-square value. The resulting chi-square and probability values become more useful as comparative values.

A previous investigation of large data sets of user query and browsing patterns for a public search engine (Ajiferuke & Wolfram, 2004) revealed that a Zipf model did not provide the best fits when compared to other more sophisticated models such as the GIGP and generalized negative binomial. The present investigation demonstrates through a more rigourous assessment of goodness-of-fit beyond visual inspection of logarithmically transformed trend lines that a Zipf model may indeed be appropriate for some types of Internet usage behaviour data, even for large data sets. Other distributions that have been proposed for modelling informetric behaviour in the past few decades such as the GIGP and generalized negative binomial distributions have three parameters and are generally less tractable than the two-parameter Zipf distribution. Given a choice, it is preferable to select the most parsimonious model that adequately describes the observed data. For the current data, a Zipf distribution would be the model of choice for the IP address visits distribution and the GLS for the resource usage distribution.

Potential limitations of any study modelling Internet usage data logs include the model fitting techniques used and the reliability of the data. The use of other minimization or maximization algorithms for parameter estimation may reveal different outcomes for the models tested, so that the most sophisticated model with the greatest flexibility does not always result in the best fit (Baayen, 2001). The noted intractability of the GIGP and need for zero-truncation for fitting many informetric data sets makes its use challenging and presents more limited options for feasible parameter estimation. The more parsimonious two-parameter GLS model may provide a useful alternative. Second, in the case of the IP address data, a confounding factor inherent in any IP address data log is the potential for dynamic allocation of IP addresses by requestors' Internet service providers. Dynamic allocation of IP addresses could impact the frequency with which a given IP address (equated to a user site or machine) actually interacts with the investigated Web site. This is beyond the control of investigators unless 'cookies' are stored on each user machine, which may not be possible on all user machines.

5. CONCLUSIONS

The proliferation of networked database resources has made it possible to collect large data sets of resource usage based on IP addresses. By relying on informetric modelling techniques incorporating theoretical models that more closely emulate observed frequency distributions of resource usage and site visits, one may be able to more accurately measure the patterns of site visitation and resource requests. This has

implications for estimating requests for the range of resources available and the visitation patterns of users. The fitted models may also be used for the design of simulation programs to determine website data traffic and requests under different circumstances.

The present study represents an initial foray into a largely unexplored area of informetric modelling by applying a broader set of theoretical models beyond a simple Zipf model. The comparatively good fit provided by the GLS distribution for the resource request data merits its further investigation for other LNRE-type data sets. Additional data sets from other systems should be collected and fitted to similar distributions to determine the general utility of different theoretical distributions for modelling Internet use. As data sets have become larger, the fitting of observed data to theoretical distributions has become more challenging. Additional approaches for determining the goodness-of-fit of different theoretical models are needed to counter the sensitivity or insensitivity of existing methods such as chi-square values and visual inspection of logarithmically transformed plots, respectively. The authors are also currently investigating the influence of data set size on the applicability of different theoretical models, where some theoretical distributions may not be able to continue to adequately model observed patterns as the data sets grow and become more skewed.

ACKNOWLEDGEMENT:

The authors would like to thank the Wisconsin Department of Public Instruction, Division for Libraries, Technology & Community Learning for providing access to the data sets used and for partial funding of this research.

REFERENCES:

Ajiferuke, I., & Wolfram, D. 2004. Informetric modelling of Internet search and browsing characteristics. Canadian Journal of Information and Library Science 28(1): 3-17.

Baayen, R. H. 2001. Word frequency distributions. Boston: Kluwer.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Staa, R., Tomlins, A., & Wiener, J. 2000. Graph structure in the Web. Computer Networks and ISDN Systems 30:209-320.

Cooper, M. D. 2001. Usage patterns of a Web-based library catalog. Journal of the American Society for Information Science and Technology 52(2):137-148.

Cothey, Vivian. 2002. A longitudinal study of World Wide Web users' information-searching behavior. Journal of the American Society for Information Science and Technology 53(2):67-78.

Crovella, M. E., Taqqu, M. S., & Bestavros, A. 1998. Heavy-tailed probability distributions in the World Wide Web. In A practical guide to heavy tails: statistical techniques and applications edited by R. J. Adler, R. E. Feldman, & M. S. Taqqu, 3-25. Boston: Birkhäuser.

Cunha, C. R., Bestavros, A., & Crovella, M. E. 1995. Characteristics of WWW client-based traces [online]. [Cited 6 March 2004]. Available from World Wide Web: (http://cs-pub.bu.edu/faculty/crovella/paper-archive/TR-95-010/paper.html).

Davis, P. M. & Solla, L. R. 2003. An IP-level analysis of usage statistics for electronic journals in Chemistry: Making inferences about user behavior. Journal of the American Society for Information Science and Technology 54(11): 1062-1068.

Famoye, F. 1997. Sampling from the generalized logarithmic series distribution**.** Computing, 58(4): 365-376.

Fichter, D. 2003. Exploiting intranet search engines for data discovery. Online 27(6):47-49, 55.

Hoelscher, C. 1998. How Internet experts search for information on the Web. In Proceedings of WebNet98 – World Conference of the WWW, Internet & Intranet, edited by H. Maurer & R.G. Olson. Charlottesville, VA.

Huberman, B. A. 2001. The laws of the Web: Patterns in the ecology of information. Cambridge, MA: The MIT Press.

Jansen, B. J., Spink, A., & Saracevic, T. 2000. Real life, real users, and real needs: A study of user queries on the Web. Information Processing & Management*, 36*(2), 207-227.

Jantz, Ronald, 2003. Information retrieval in domain-specific databases: an analysis to improve the user interface of the Alcohol Studies Database. College & Research Libraries 64(5):229-239.

Jones, Steve, Cunningham, Sally Jo, McNab, Rodger, & Boddie, Stefan, 2000. A transaction log analysis of a digital library. International Journal on Digital Libraries 3(2):152-169.

Ke, Hao-Ren, Kwakkelaar, Rolf, Tai, Yu-Min, & Chen, Li-Chun. 2002. Exploring behavior of E-Journal users in science and technology: transaction log analysis of Elsevier's ScienceDirect Onsite in Taiwan. Library and Information Science Research 24: 265-291.

Nelson, M. J. 1989. Stochastic models for the distribution of index terms. Journal of Documentation, 45(3): 227-237.

Nielsen, J. 1997a. Do websites have increasing returns? [online] [cited 6 March 2004] Available from World Wide Web: (http://www.useit.com/alertbox/9704b.html).

Nielsen, J. 1997b. Zipf curves and website popularity. [online] [cited 6 March 2004] Available from World Wide Web: (http://www.useit.com/alertbox/zipf.html).

NIST/SEMATECH e-Handbook of statistical methods. [online] [cited 6 March 2004] Available from the World Wide Web: (http://www.itl.nist.gov/div898/handbook/eda/section3/ eda35g.htm).

Sichel, H. S. 1992. Anatomy of the generalized inverse Gaussian-Poisson distribution with special applications to bibliometric studies. Information Processing & Management 28(1): 5-17.

Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a very large web search engine query log. SIGIR Forum, *33*(1). [online] [cited 6 March 2004] Available from World Wide Web: Available: (http://www.acm.org/sigir/forum/F99/Silverstein.pdf).

Spink, A., Wolfram, D., Jansen, B. J., & Saracevic, T. 2001. Searching the Web: The public and their queries. Journal of the American Society for Information Science and Technology 52(3): 226-234.

Wolfram, D., & Xie, H. 2000. End user database searching over the Internet: An analysis of the state of Wisconsin's BadgerLink service. In Proceedings of the 20th National Online Meeting edited by M.E. Williams, 503-512. Medford, NJ: Information Today.