American Journal of Computational Mathematics, 2016, 6, 17-22
Published Online March 2016 in SciRes. http://www.scirp.org/journal/ajcm
http://dx.doi.org/10.4236/ajcm.2016.61002

Self Similarity Analysis of Web Users Arrival Pattern at Selected Web Centers

Pushpalatha Sarla, Mallikarjuna Reddy Doodipala*, Manohar Dingari

Department of Engineering Mathematics, GITAM University Hyderabad Campus, Hyderabad, India
Email: *mallik.reddyd@gmail.com

Received 5 December 2015; accepted 1 March 2016; published 4 March 2016

Abstract

The paper focuses on measuring self-similarity using few techniques by an index called Hurst index which is a self-similarity parameter. It has been evident that Internet traffic exhibits self-similarity. Motivated by this fact, real time web users at various centers considered here as traffic and it has been examined by various methods to test the self-similarity. The results from the experiments carried out verify that the traffic examined in the present study is self similar using a new method based on some descriptive measures; for example percentiles have been applied to compute Hurst parameter which gives intensity of the self-similarity. Numerical results and analysis we discussed and presented here play a significant role to improve the services at web centers in the view of quality of service (QOS).

Keywords

Long-Range Dependence, Self-Similarity, Poisson Process, Percentiles, Hurst Parameter

1. Introduction

At present one of the major issues to know various traffic flows is in self similar nature to study and design some performance metric as that of Ethernet traffic etc. Until recently Poison approach has been used to model the road traffic irrespective of traffic intensity [1]. This was similar to the practice in the cases of Ethernet, LAN, WAN, and WWW traffic. But seminal studies [2] [3] reveal that IP packet traffic in supposed networks tends to be bursty in nature on many time scales. This burstiness of traffic can be characterized mathematically as self-similar or long-range dependence (LRD). It is clear from the work agreed [4] that Poisson process could not emulate the self-similar network traffic. Markovian arrival process (MAP) emulating self-similar traffic is fitted over desired time scales by equating second-order statistics of the counts [5]-[9].

*Corresponding author.

How to cite this paper: Sarla, P., Doodipala, M.R. and Dingari, M. (2016) Self Similarity Analysis of Web Users Arrival Pattern at Selected Web Centers. *American Journal of Computational Mathematics*, **6**, 17-22.
http://dx.doi.org/10.4236/ajcm.2016.61002

The idea of this paper is, we examine whether web users traffic data has the self similar property. This is to enhancement earlier results using a real time data [10]. This kind of research is useful for future studies to know the performance metrics and continuous improvement of web centers at various private and in public sector organizations. The rest of the paper has been organized as follows: Definition of self-similarity or long range dependence is given in Section II. Materials and methods are placed in Section III. In Section IV, Hurst parameter is computed using various methods. Finally, conclusions are given in Section V.

2. Long-Range Dependence and Self-Similarity

In this section we give a short description of the mathematical basis for second order self-similar processes (long-range dependence).

Exact Second-Order Self-Similar Process

The exact second-order self-similar process is defined as follows. Arrival instants are modeled as point process. Divide the time axis into disjoint intervals of unit length and let $X=\{X_t : t=1,2,\cdots\}$ be the number of points (arrival) in the t^{th} interval. Let X be a second order stationary process with variance σ^2 and the autocorrelation function $\gamma(k), k\geq 0$ is given by

$$\gamma(k)=\frac{Cov(X_t, X_{t+k})}{Var(X_t)} \tag{2.1}$$

For each $m=1,2,3,\cdots,$ let a new time series $X_t^{(m)}$ is obtained averaging the original time series X over non-overlapping blocks of size m. That is

$$X_t^{(m)}=\frac{1}{m}\sum_{i=1}^{m} X_{(t-1)m+i}, \quad t=1,2,\cdots \tag{2.2}$$

This new series $X_t^{(m)}$, for each m, is also a second order stationary process with autocorrelation function $\gamma^{(m)}(k)$.

Definition 1: The process "X" is said to be exactly second order self-similar with Hurst parameter $H=1-\frac{\beta}{2}$ and variance σ^2 if

$$\gamma(k)=\frac{\sigma^2}{2}\left[(k+1)^{2H}-2k^H+(k+1)^{2H}\right], \forall k\geq 1 \tag{2.3}$$

Definition 2: The process "X" is said to be asymptotically second order self-similar with Hurst parameter $H=1-\frac{\beta}{2}$ and variance σ^2 if

$$\sum_{m\to\infty}\gamma^{(m)}(k)=\frac{\sigma^2}{2}\left[(k+1)^{2H}-2k^H+(k+1)^{2H}\right], \forall k\geq 1 \tag{2.4}$$

In terms of variance, self-similar process is defined as follows:

Definition 3: The process "X" is said to be exactly second order self-similar with Hurst parameter $H=1-\frac{\beta}{2}$ and variance σ^2 if

$$Var\left(X^{(m)}\right)=\sigma^2 m^{-\beta}, \forall m\geq 1 \tag{2.5}$$

Now we shall differentiate long range dependence (LRD) and short range dependence (SRD) processes. For $H\neq 0.5,$ from the Equation (2.3), we can see that $\gamma(k)=H(2H-1)k^{2H-2}$ as $k\to\infty$, and we have

$$\sum_k \gamma(k)\sim c\sum_k k^{-\beta}, \quad c=H(2H-1). \tag{2.6}$$

The series $c\sum_k k^{-\beta}$ is divergent if $0.5<H<1$ or $0<\beta<1$ otherwise they are convergent, being a posi-

tive term series. Accordingly the left hand series $\sum_k \gamma(k)$ is divergent if $0.5 < H < 1$ or $0 < \beta < 1$, otherwise they are convergent. That is, for $0.5 < H < 1$, the autocorrelation functions decays slowly, that is hyperbolically. In this case, the process *X* is called Long Range Dependent (LRD). The process *X* is Short Range Dependent (SRD) if $0 < H < 0.5$ and the autocorrelation function is summable (finite).

3. Materials and Methods

As discussed in the introduction, we are primarily interested collecting data from various sources. Real time web users data has been considered. The sample number of users logged on to an Internet server each minute over 100-minutes (see Appendix). In the study web users data can be treated as traffic and verify it is self-similar or not.

4. Methods for Estimating Hurst Parameter of Self-Similar Process

The intensity of self-similarity is given by Hurst parameter, *H*. The parameter *H* was named after the hydrologist H.E. Hurst who spent many years to investigate the problem of water storage and also to determine the level patterns of the Nile river. The parameter *H* has range $0.5 \le H \le 1$. Estimation of *H* is a difficult task. Several methods are available to estimate degree of self-similarity in a time-series. We also present the three basic methods to calculate the Hurst parameter: Periodogram analysis, Correlogram method, R/S analysis, Variance-time analysis etc. Here is a method based on percentiles is applied and validated with the said methods.

4.1. Periodogram Analysis

In the frequency domain, analysis of time series is merely the analysis of a stationary process by means of its spectral representation. The periodogram [11] is given by

$$I_N(\lambda) = \frac{1}{2\pi N}\left|\sum_{k=0}^{N-1} X_k e^{k\lambda}\right|^2 \tag{2.7}$$

where λ is the Fourier frequency, *N* is the number of terms in the time series and X_j is the data of the given series. To estimate *H*, first, one has to calculate this periodogram. Since $I_N(\lambda)$ is an good sample estimator of the spectral density, a series with long-range dependence should have a periodogram, which is proportional to $|\lambda|^{1-2H}$ close to the origin. Then a regression of the logarithm of the periodogram on the logarithm of the frequency λ should give a coefficient of $1-2H$. The slope of the fitted straight line is the estimate of $1-2H$. Using this method the *H* value is computed for the data given in section III. The obtained value of *H* in this case is 0.763.

4.2. Correlogram Method

In time series analysis [12], plot of ACF (autocorrelation function) is known as correlogram where the estimated correlation can be given in terms of auto-covariance function $\gamma(k)$

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} \tag{2.8}$$

It has already been observed that slow decay of correlation, which is proportional to K^{2H-2} for $\frac{1}{2} < H < 1$ indicates the long-memory process. Therefore, the plot of the sample autocorrelation should exhibit this property. A much better plot for the handling of long-range dependence is the plot of ACF in logarithmic scale. If the asymptotic decay of the correlation is hyperbolic, then the points in the plot should be approximately scattered around a straight line with a negative slope of $2H-2$ for the long memory processes but for short memory, the points should tend to diverse to minus infinity at an exponential rate. If the time series is long enough or if the series has strong long-range dependence, then this log-log correlogram is useful. Correlogram is useful as a preliminary heuristic approach to the data. Some pitfalls of sample correlation which are less known can be found in Mandelbrot [13]-[15]. Even though it is neither widely used nor attractive method for estimation, still *H*,

the self-similarity parameter, can be estimated by this method deriving an equation of the form

$$\rho(k) = \hat{H}\left(2\hat{H} - 1\right)K^{2\hat{H}-2} \tag{2.9}$$

Using this method, the obtained value of H in this case is 0.79.

4.3. Percentile Method

In statistical methodologies, a percentile (or centile) is the value of a variable below which a certain percent of observations fall, like partition values of a process such as quartiles and deciles. There is no exact definition of percentile [16], however all definitions yield similar results when the number of observations is very large. One definition of percentile, often given in texts, is that the P^{th} percentile $(1 \le P \le 100)$ of N ordered values is obtained by first calculating the rank.

$$n = \frac{P * N}{100} + \frac{1}{2} \tag{2.10}$$

Given data set or time series $(t, Z_t)(t \ge 0)$. First we can find the percentiles $(P_i,\ i = 1, 2, \cdots, 100)$ for a given time series using

$$P_i = \frac{i * N}{100} + \frac{1}{2};\quad i = 1, 2, \cdots, 100. \tag{2.11}$$

$P_i = i^{th}$ percentile, this a special type of average such as partition values in descriptive statistics like quartiles (Q_1, Q_2, Q_3). Draw a scattered Plot percentile number against percentiles on log scales. A linear equation $Z_t = \beta t + c$ (say) is obtained with the slope (β). The Hurst parameter (H) is then computed by

$$H = 1 - \frac{\beta}{2}. \tag{2.12}$$

Using this method, the H value is computed for the data. The pertaining scattered data and trend line with the slope $(\beta = m = 0.476)$ are depicted in **Figure 1**. The obtained value of H in this case is 0.762. One relevant paper [17], which explained how the 95-percentile depends on the aggregation window size, and how this phenomenon justifies the mathematical definition of self similarity. The advantages of this method are: This method is matter of a simple empirical formula, unlike other two methods. Data however large it may be is divided into 100 parts (partition values) and the plotting involves only 100 points (percentile versus percentile number).

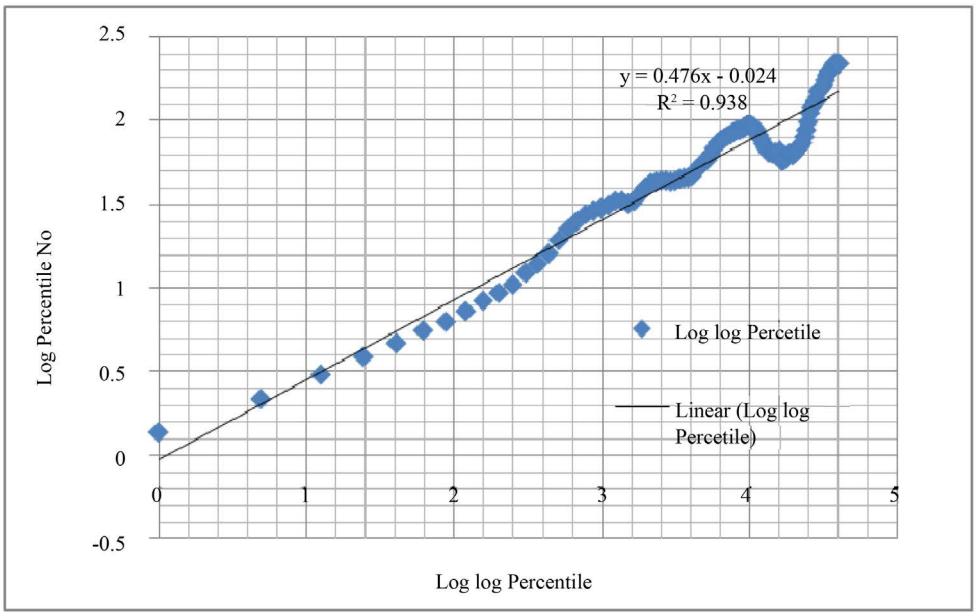

Figure 1. Percentile versus percentile number.

4.4. Math Lab Implementation Code is Linked with the Percentile Method

```
function H = percentile(userspermin,isplot)
if nargin == 1
   isplot = 0;
end
pps_fname = 'userspermin.txt';
[sequence] = textread(pps_fname, '%u');
N = length(sequence);
for i=1:100
   p(i)=i*N/100+0.5;
   disp(p(i));
   i=i+1;
end
```

5. Some Conclusions

In this paper, real time web user's data has been considered as traffic from various web centers and it has been proved to be self-similar. Various methods to test the self-similarity have been used. The obtained values of Hurst parameter *H* are reasonably close to each other. This kind of research is useful for future studies to know the performance metrics at web centers.

References

[1] Perati, M.R., Raghavendra, K., Koppula, H.K.R., Doodipala, M.R. and Dasari, R. (2012) Self-Similar Behavior of Highway Road Traffic and Performance Analysis at Toll Plazas. *Journal of Transportation Engineering*, **138**, 1233-1238.

[2] Leland, W.E., Taqqu, M.S., Willinger, W. and Wilson, D.V. (1994) On the Self-Similar Nature of Ethernet Traffic (Extended Version). *IEEE/ACM Transactions on Networking*, **2**, 1-15. http://dx.doi.org/10.1109/90.282603

[3] Crovella, M. and Bestavros, A. (1997) Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes. *IEEE/ACM Transactions on Networking*, **5**, 835-846. http://dx.doi.org/10.1109/90.650143

[4] Paxson, V. and Floyd, S. (1995) Wide Area Traffic: The Failure of Poisson Modeling. *IEEE/ACM Transactions on Networking*, **3**, 226-244. http://dx.doi.org/10.1109/90.392383

[5] Anderson, A.T. and Nielsen, B.F. (1998) A Markovian Approach for Modeling Packet Traffic with Long Range Dependence. *IEEE Journal on Selected Areas in Communications*, **16**, 719-732. http://dx.doi.org/10.1109/49.700908

[6] Yoshihara, T., Kasahara, S. and Takahashi, Y. (2001) Practical Time-Scale Fitting of Self-Similar Traffic with Markov-Modulated Poisson Process. *Telecommunication Systems*, **17**, 185-211. http://dx.doi.org/10.1023/A:1016616406118

[7] Shao, S.K., Perati, M.R., Tsai, M.G., Tsao, H.W. and Wu, J. (2005) Generalized Variance-Based Markovian Fitting for Self-Similar Traffic Modeling. *IEICE Transactions on Communications*, **E88-B**, 4659-4663.

[8] Doodipala, M.R. (2013) Second Order Statistics of Time Series of Various Real Time Problems in Conjunction with Periodogram Technique. *International Journal of Latest Trends in Engineering and Technology* (*IJLTET*), **3**.

[9] Nagatani, T. (2005) Self-Similar Behavior of a Single Vehicle through Periodic Traffic Lights. *Physica A*, **347**, 673-682. http://dx.doi.org/10.1016/j.physa.2004.08.007

[10] Markidakis, S., Wheelwright, S.C. and Hydman, R.J. (1998) Forcating Methods and Applications. 3rd Edition, John Wiley & Sons. Inc., Hoboken.

[11] Meng, Q. and Khoo, H.L. (2009) Self-Similar Characteristics of Vehicle Arrival Pattern on Highways. *Journal of Transportation Engineering*, **135**, 864-872.

[12] Brockwell, P.J. and Davis, R.A. (1996) An Introduction to Time Series and Fore-Casting. Springer - Verlag, New York. http://dx.doi.org/10.1007/978-1-4757-2526-1

[13] Sarker, M.M.A. (2007) Estimation of the Self-Similarity Parameter in Long Memory Processes. *Journal of Mechanical Engineering*, **ME38**, 32-37. Transaction of the Mech. Eng. Div., The Institution of Engineers, Bangladesh.

[14] Beran, J., Taqqu, M.S. and Willinger, W. (2002) Long-Range Dependence in Variable Bit Rate Traffic. *IEEE Transactions on Communications*, **43**, 1566-1579.

[15] Beran, J. (1994) Statistics for Long-Memory Processes. Chapman and Hall.

[16] Lane, D. Percentiles. http://cnx.org/content/m10805/latest

[17] Web Hosting Talk Forum: 95th Percentile Billing Polling Interval. http://www.webhostingtalk.com/showthread.php?t=579003(2008)

Appendix

The number of users logged on to an Internet server each minute over 100-minutes.

Minute	Web Users	Minute	Web Users	Minute	Web Users	Minute	Web Users
1	88	26	139	51	172	76	91
2	84	27	147	52	172	77	91
3	85	28	150	53	174	78	94
4	85	29	148	54	174	79	101
5	84	30	145	55	169	80	110
6	85	31	140	56	165	81	121
7	83	32	134	57	156	82	135
8	85	33	131	58	142	83	145
9	88	34	131	59	131	84	149
10	89	35	129	60	121	85	156
11	91	36	126	61	112	86	155
12	99	37	126	62	104	87	171
13	104	38	132	63	102	88	175
14	112	39	137	64	99	89	177
15	126	40	140	65	99	90	182
16	138	41	142	66	95	91	193
17	146	42	150	67	98	92	204
18	151	43	159	68	84	93	208
19	150	44	167	69	84	94	210
20	148	45	170	70	87	95	215
21	147	46	171	71	89	96	222
22	149	47	172	72	88	97	228
23	143	48	172	73	85	98	226
24	132	49	174	74	86	99	222
25	131	50	175	75	89	100	220