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.reddy@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.

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 \( \sigma^2 \) and the autocorrelation function \( \gamma(k), k \geq 0 \) is given by

\[
\gamma(k) = \frac{\text{Cov}(X_t, X_{t+k})}{\text{Var}(X_t)}
\]  

(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

\[
X_t^{(m)} = \frac{1}{m} \sum_{i=1}^{m} X_{(t-1)+mi}, \quad t = 1, 2, \cdots
\]  

(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 \( \sigma^2 \) if

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

(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 \( \sigma^2 \) if

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

(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 \( \sigma^2 \) if

\[
\text{Var}(X^{(m)}) = \sigma^2 m^{-\beta}, \forall m \geq 1
\]  

(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 \( k \rightarrow \infty \), and we have

\[
\sum_{k} \gamma(k) \sim c \sum_{k} k^{-\beta}, \quad c = H(2H-1).
\]  

(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.


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  \( 0.5 \leq H \leq 1 \). 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

\[
I_N(\lambda) = \frac{1}{2\pi N} \left| \sum_{k=0}^{N-1} X_k e^{i\lambda k} \right|^2
\]  

(2.7)

where \( \lambda \) is the Fourier frequency, \( N \) is the number of terms in the time series and \( X_k \) is the data of the given series. To estimate \( H \), first, one has to calculate this periodogram. Since \( I_N(\lambda) \) is a good sample estimator of the spectral density, a series with long-range dependence should have a periodogram, which is proportional to \( |\lambda|^{-2H} \) close to the origin. Then a regression of the logarithm of the periodogram on the logarithm of the frequency \( \lambda \) 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  \( \gamma(k) \)

\[
\rho(k) = \frac{\gamma(k)}{\gamma(0)}
\]  

(2.8)

It has already been observed that slow decay of correlation, which is proportional to \( K^{-2H} \) 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

$$\rho(k) = \hat{H}(2\hat{H} - 1)K^{2\hat{H} - 2}$$  \hspace{1cm} (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 ($1 \leq P \leq 100$) of $N$ ordered values is obtained by first calculating the rank.

$$n = \frac{P \times N}{100} + \frac{1}{2}$$  \hspace{1cm} (2.10)

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

$$P_i = \frac{i \times N}{100} + \frac{1}{2}; \hspace{0.5cm} i=1,2,\cdots,100.$$  \hspace{1cm} (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 $Z_i = \beta t + c$ (say) is obtained with the slope $(\beta)$. The Hurst parameter $(H)$ is then computed by

$$H = 1 - \frac{\beta}{2}.$$  \hspace{1cm} (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.](image-url)
4.4. Math Lab Implementation Code is Linked with the Percentile Method

```matlab
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


Appendix

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

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