Available online at www.sciencedirect.com
ScienceDirect
Information Processing Letters 102 (2007) 187–190
Information Processing Letters
www.elsevier.com/locate/ipl

Nonstationary Poisson modeling of web browsing session arrivals

Edward Chlebus *, Jordy Brazier

Network Modeling and Teletraffic Analysis Laboratory, Department of Computer Science, Illinois Institute of Technology, 10 W. 31st Street, Chicago, IL 60616, USA

Received 30 September 2005

Available online 8 March 2007

Communicated by F. Dehne

Abstract

In this paper we show that daily web session arrivals are nonstationary Poissonian with a rate surprisingly stable, i.e., changing only once per 24 hours in all the examined cases.

Keywords: Internet; Web traffic; Performance evaluation; Nonstationary Poisson process

1. Introduction

A user visiting a certain web site makes a sequence of requests for pages or files that he would like to download. There are three classes of web files [1]:

(i) base files, i.e., HTML files that make web pages and contain embedded files;
(ii) embedded files, e.g., images or graphics referenced by base files;
(iii) single files, e.g., *.pdf, *.doc, *.ppt, *.ps, *.zip, *.txt files, etc. that do not fall into the two above categories.

Web visitors make intentional requests only for the base or single files, the embedded ones are downloaded automatically along with the base files referencing them.

We assume that each IP address originating a web request uniquely identifies a user. This provides a reasonable approximation of the number of distinct visitors [6].

Each user's activity is traced by analyzing time stamps of his requests for the base or single files recorded in an access log, i.e. $t_1, t_2, t_3, \ldots, t_n$. If the interarrival time $t_{i+1} - t_i$ of two subsequent requests is greater than a certain user inactivity threshold T, we assume that t_i is a final request of a browsing session and t_{i+1} initiates a new one. Based on this criterion we can group user's requests into separate sessions, e.g., $(t_1, t_2, t_3, \ldots, t_i)$, $(t_{i+1}, t_{i+2}, t_{i+3}, \ldots, t_j)$, $(t_{j+1}, t_{j+2}, t_{j+3}, \ldots, t_k), \ldots, (t_{m+1}, t_{m+2}, t_{m+3}, \ldots, t_n)$. No request interarrival time within a session can exceed the inactivity threshold T. The consecutive time stamps $t_1, t_{i+1}, t_{j+1}, \ldots, t_{m+1}$ determine an arrival stream of browsing sessions originated by a single user. In this paper we investigate superposition of all such streams generated by the entire population of visitors, i.e., all the sessions incoming to a web site.

* Corresponding author.
E-mail address: chlebus@iit.edu (E. Chlebus).
URL: http://www.cs.iit.edu/~nemtal (E. Chlebus).

0020-0190/$ – see front matter
doi:10.1016/j.ipl.2006.12.015

2. Measurement data

The presented analysis is based on an access log for a web site of Computer Science Department of Illinois Institute of Technology (www.cs.iit.edu). The log was collected over an 8-day period from midnight (0:00), Wednesday, December 1st to midnight (24:00), Wednesday, December 8th 2004. Basic traffic characteristics corresponding to the $T = 60$ min session inactivity period are given in Table 1.

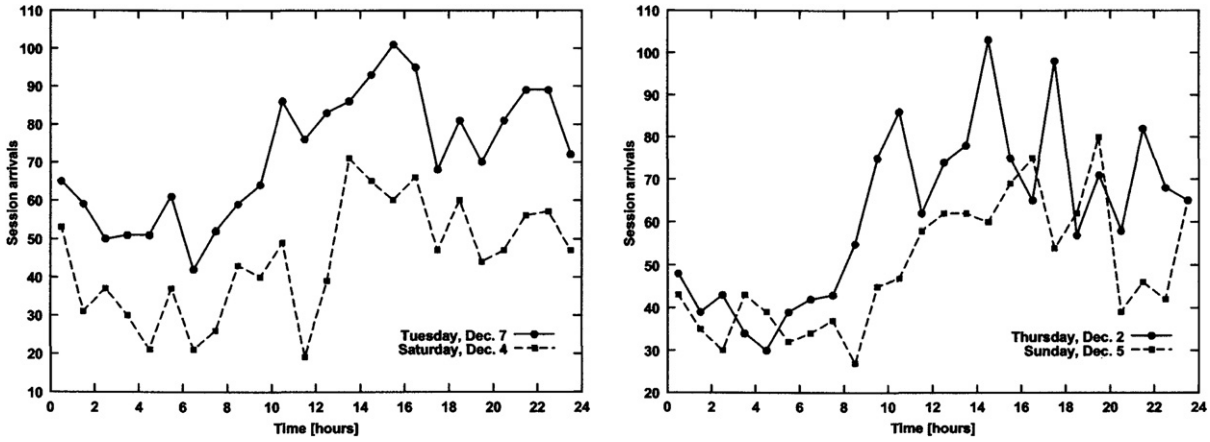
Table 1
Basic web traffic characteristics for the entire data collection period (session idle timeout $T = 60$ minutes)

Total number of sessions	12 098
Total number of requests for base or single files	42 951
Total transferred load [MB]	2851.77
Mean session arrival rate [1/hour]	63.01
Mean number of requests for base or single files per session	3.55
Mean transferred load per session [MB]	0.24

3. Does session arrival rate vary with time?

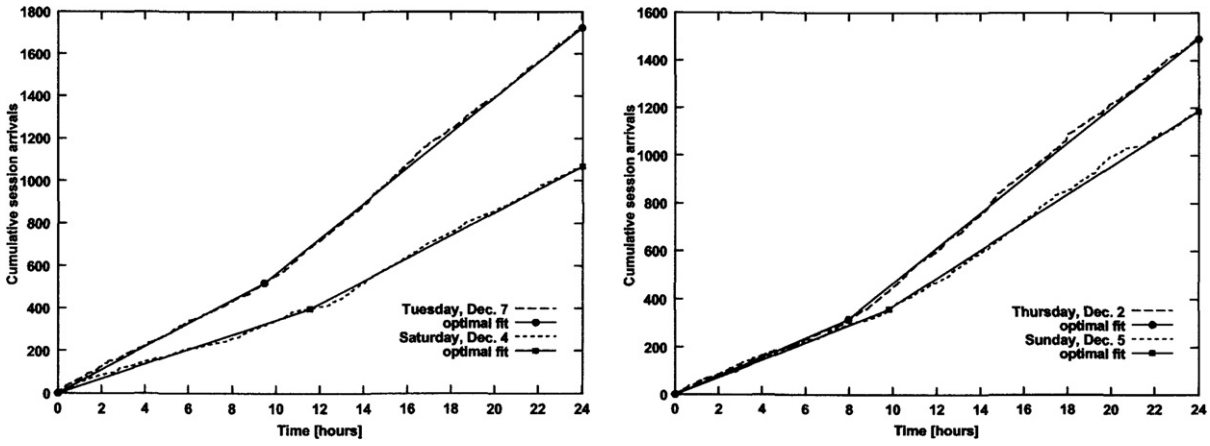
A daily traffic profile is the common way of presenting intensity of arrivals incoming to a system under study. This presentation method, applied to the traffic data extracted from the log, results in Fig. 1. Each point (x, y) represents a session arrival count y over a 1-hour interval $[x - 30$ min, $x + 30$ min], e.g., (15:30, 101) means that web site visitors started 101 browsing sessions from 3 to 4 pm. Judging by Fig. 1 the session arrival rate varies with time significantly which supports an intuitively positive answer to the question initially asked in the title of this section. Let us see whether this answer is correct.

We have reprocessed the data depicted in Fig. 1 and presented them in a different form in Fig. 2. The curves



(a) (b)

Fig. 1. Daily traffic profile. Session arrival counts for 1-hour intervals.



(a) (b)

Fig. 2. Cumulative session arrivals as a function of time.

illustrating cumulative session arrivals $a(t)$ as a function of time t reveal a piecewise linear trend. There are two linear segments with different slopes. The first segment corresponds to the period from midnight (0:00) to the trend break point $t = b$ that occurs in the morning, the second segment covers the interval from the break point $t = b$ to midnight (24:00). Judging by the gradient of the trend, we can conclude from Fig. 2 that the session arrival rate is constant over a surprisingly long period of time (of the order of hours) and changes only once per day at the break point $t = b$.

4. Fitting an optimal trend to the data

The piecewise linear trend $f(t)$ fitted to the data in Fig. 2 has been defined as follows. The first segment of $f(t)$ links a pair of points $[t = 0, a(t = 0)]$ and $[t = b, a(t = b)]$, the second one links the points $[t = b, a(t = b)]$ and $[t = t_s, a(t = t_s)]$, where $b \in A$, $A = \{0, t_1, t_2, \ldots, t_s\}$, t_k is the kth session arrival time and s denotes the total number of session arrivals over the daily observation period, i.e., from 0:00 to 24:00. Note, that the gradient of such defined trend has the meaning of an instantaneous session arrival rate and approximates $da(t)/dt$. To indicate that the fit $f(t)$ is uniquely determined by the choice of the break point b we introduce the notation $f(t, b)$.

In order to find the optimal break point $[b_{\min}, a(b_{\min})]$ we have successfully run the following optimization algorithm:

Step 1. Assume the initial break point $[b = 0, a(b = 0)]$.
Step 2. Find the fit $f(t, b)$ corresponding to the break point $[b = 0, a(b = 0)]$.
Step 3. Determine the goodness of the fit for a given b as $g(b) = \sum_{k=1}^{s} [a(t_k) - f(t_k, b)]^2$.
Step 4. Minimize $g(b)$ searching for $b = b_{\min}$ over the entire domain [0:00, 24:00], i.e.,

$$g(b_{\min}) = \min_{b \in A} g(b).$$

Note, that one can identify an approximate location of the optimal break point by visual examination of the curve $a(t)$, which may narrow the region searched for $b_{\min}$ in Step 4 above. However this is not necessary in practice since the proposed algorithm is computationally very efficient. We have run it for the entire time domain [0:00, 24:00] for the sake of optimization accuracy.

The optimal break points $b_{\min}$ and corresponding session arrival rates are given in Table 2. As we can see

Table 2
Optimal fit related session arrival traffic characteristics

Periods determined by $b_{\min}$	Arrival rate [1/hour]	Kolmogorov–Smirnov test		Ljung–Box test	
		Test statistic	Acceptance level	Test statistic	Acceptance level
Wed, Dec. 1					
0:00–5:45	48.40	1.4147	< 1%	55.3505	5%
5:45–24:00	70.98	0.9022	15%	198.9553	25%
Thu, Dec. 2					
0:00–7:57	39.71	0.6769	15%	67.2617	1%
7:57–24:00	73.15	1.3000	1%	176.8082	25%
Fri, Dec. 3					
0:00–4:46	46.76	0.6626	15%	39.7492	10%
4:46–24:00	56.41	1.2184	1%	158.0384	50%
Sat, Dec. 4					
0:00–11:33	34.16	1.1624	2.5%	79.6501	1%
11:33–24:00	53.86	1.1285	2.5%	114.8100	10%
Sun, Dec. 5					
0:00–9:49	36.54	0.7792	15%	56.2402	25%
9:49–24:00	58.41	1.6766	< 1%	146.8633	5%
Mon, Dec. 6					
0:00–6:51	54.57	0.5972	15%	56.1969	25%
6:51–24:00	90.59	2.5609	< 1%	273.0724	1%
Tue, Dec. 7					
0:00–9:28	54.78	1.3586	< 1%	78.0151	25%
9:28–24:00	82.92	1.2831	1%	164.4426	50%
Wed, Dec. 8					
0:00–6:36	56.94	0.7907	15%	56.2613	25%
6:36–24:00	83.31	1.5334	< 1%	232.0684	10%

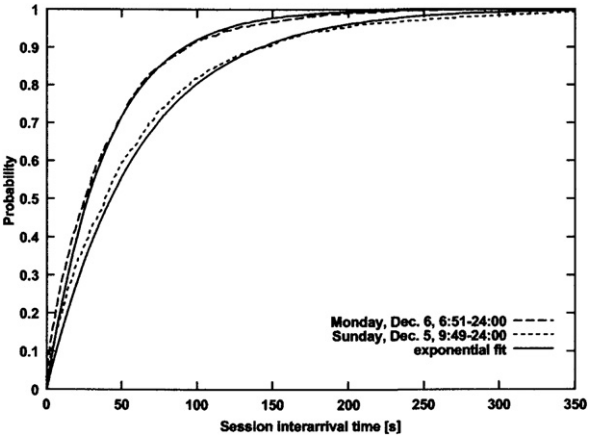


Fig. 3. Empirical CDF versus the exponential fit.

from Fig. 2 the goodness of the sample optimal fits is excellent.

5. Are the session arrivals Poissonian?

Statistical characteristics of session arrivals within each period determined by the optimal trend break point still remain unknown. The Poisson distribution $p_k = (\lambda\Delta t)^k/k! \cdot \mathrm{e}^{-\lambda\Delta t}$, which gives probability of k arrivals to a system fed with an arrival rate λ over a period of length Δt, is a straightforward candidate for modeling session inflow to a web site. The Poisson distribution has been successfully used numerous times in performance analysis of systems with very large population of subscribers, e.g., a telephone central office, a node of a packet-switched network and recently also an email server [4].

To verify the hypothesis that web session arrivals within each period [0:00, $b_{\min}$] or [$b_{\min}$, 24:00] are Poissonian we have tested the session interarrival times for exponentiality and independence (see Appendix A of [5] for discussion on testing methodology). The Kolmogorov–Smirnov (KS) test with the mean estimated [2] was used for the former and the Ljung–Box (LB) test [3] for the latter. The numerical values of the test statistics are given in Table 2.

All the data have passed the LB test at the significance level of at least 1% but five out of sixteen tested data sets have not passed the KS test at this level. We have examined two worst cases with the highest KS test statistics, i.e., 2.5609 and 1.6766. The corresponding cumulative distribution functions (CDF) of session interarrival times collected on Monday, Dec. 6, from 6:51 to 24:00 and Sunday, Dec. 5, from 9:49 to 24:00 are compared with the exponential fits in Fig. 3. As we can see both fits are very good.

Thus we can conclude that all the tested interarrival times, including those that have not passed the KS test, can be recognized as exponential for engineering purposes (actually the values 2.5609 and 1.6766 of the KS test statistics are very close to the 1.308 acceptance threshold at the 1% significance level [2]). At this point we are in a position to state that the Poisson distribution is an acceptable session arrival model for every examined period [0:00, $b_{\min}$] or [$b_{\min}$, 24:00].

6. Conclusions

Daily web browsing session arrivals can be modeled by applying the nonstationary Poisson process. In all the analyzed cases a distinct break of traffic intensity trend splits the entire 24-hour period into two regions of different user activity approximately corresponding to night and day. Traffic is Poissonian with constant arrival rate in each of these regions. Stationarity of traffic for several hours is surprising.

Nightly traffic examined for the entire data collection period, i.e., Dec. 1–8 ranges from mid 30s to mid 50s session arrivals per hour, daily traffic varies from mid 50s to low 90s and peaks on Monday. On weekend traffic intensity is lower than on weekdays and the break point of the piecewise linear trend occurs later.

References

[1] P. Barford, M. Crovella, A performance evaluation of Hyper Text Transfer Protocols, in: Proc. 1999 ACM SIGMETRICS Int. Conf. on Measurement and Modeling of Computer Systems, Atlanta, GA, May 1999, pp. 188–197.

[2] A.M. Law, W.D. Kelton, Simulation Modeling and Analysis, third ed., McGraw-Hill, 2000.

[3] G.M. Ljung, G.E.P. Box, On a measure of lack of fit in time series models, Biometrika 65 (2) (1978) 297–303.

[4] R. Ohri, E. Chlebus, Measurement-based e-mail traffic characterization, in: Proc. 2005 Int. Symp. on Performance Evaluation of Computer and Telecommunication Systems, SPECTS'05, Philadelphia, PA, July 24–28, 2005.

[5] V. Paxson, S. Floyd, Wide area traffic: The failure of Poisson modeling, IEEE/ACM Trans. on Networking 3 (3) (1995) 226–244.

[6] M. Rosenstein, What is actually taking place on web sites: E-commerce lessons from web server logs, in: Proc. 2nd ACM Conf. on Electronic Commerce, Minneapolis, MN, 2000, pp. 38–43.