

Presenting author:

Dr. George Bilchev
British Telecom Research Laboratories
Admin 2, pp. 5,
Martlesham Heath
Ipswich IP5 3RE
UK
Email: george.bilchev@bt-sys.bt.co.uk

Other Authors:

Ian Marshall
BT Labs
B-54
Ipswich IP5 3RE
UK
Email: marshall@drake.bt.co.uk

Dr.. Sverrir Olafsson
BT labs
Admin 2, pp. 5
Ipswich IP3 RE
UK
Email: sverrir.olafsson@bt-sys.bt.co.uk

Chris Roadknight
BT Labs
B-54
Ipswich IP5 3RE
UK
Email: roadknic@drake.bt.co.uk

Modelling Http Traffic Generated by Community of Users

George Bilchev, Ian Marshall, Sverrir Olafsson, Chris Roadknight

BT Laboratories, Martlesham Heath, Ipswich IP5 3RE, UK
<http://http://www.labs.bt.com/>

Abstract. A model of the http traffic generated by a community of users connected to the Internet via a proxy cache is described. The model reproduces Internet traffic realistically and is used as input to the Internet cache simulation models developed by British Telecom research laboratories.

1. Introduction

Single users generate file requests in a certain manner consistent with their Internet browsing behaviour. When all the requests from the individual users are aggregated at the proxy, they form the incoming file request pattern that is experienced by the caching algorithm. In previous research [1, 2] this pattern has been used in order to extract a relevant single user browsing behaviour model. Results have shown that users typically follow an avalanche of hyper-link clicks until they find the information they have been looking for. Taking each avalanche of hyper-link clicks as a single *browsing session*, and plotting the distribution of session lengths (a browsing session ends if the client is inactive for more than a predefined amount of time, say 5 minutes) we find that the resulting distribution has a long tail (fig. 1). This suggests that there will be a significant degree of auto-correlation in the aggregated requests of all the users in a community. Fig. 2a shows measured http traffic (number of requests) aggregated over a 5 minute time window. We suggest that two main characteristics, illustrated in fig. 2b can be identified from the graph. The first relates to the underlying trend, which reflects the (daily) request pattern of the clients. It can be extracted by calculating a moving average. The second is a stochastic component, which is “superimposed” on the trend. This stochastic component is clearly not random noise and exhibits a certain degree of auto-correlation. Figures 2c and 2d show the prediction of the model for a similar community. It is clear that the main characteristics are reproduced and the predicted traffic is thus realistic.

2. The Model

To model the underlying trend we suggest using a superposition of periodic functions:

$$\begin{aligned} y_i^{\text{trend}}(t) &= \max \left\{ a_i + b_i \sin(2\pi c_i \frac{t}{T} + d_i), 0 \right\} \\ y^{\text{trend}}(t) &= \max_i \{ y_i^{\text{trend}}(t) \} \end{aligned} \quad (1)$$

where a_i is an amplitude shift, b_i is the amplitude, c_i is the frequency, d_i is the phase and T is the period during which cyclic patterns are observed. The values of the parameters can be tuned by curve fitting the trend model (1) to the approximated trend.

Once the trend has been approximated the stochastic component can be modelled as a Brownian motion:

$$y^{\text{BM}}(t) = y^{\text{BM}}(t-1) + \eta \quad (2)$$

Two points are worth mentioning. First, since the number of requested files is always non-negative we have to truncate a negative value of $y^{\text{BM}}(t)$ to zero. Second, bursts in positive direction are higher than bursts in negative direction. To accommodate for this we define η as:

$$\eta = \begin{cases} \eta' & \text{if } \eta' > 0 \\ \frac{\eta'}{\lambda} & \text{otherwise} \end{cases} \quad (3)$$

where $\eta' \in \text{Norm}(0, \sigma)$ and λ is a parameter determining the ratio between the heights of the positive and negative bursts. The second modification also has the effect of reducing the number of times the series has to be truncated due to negative values.

Since the auto-correlated stochastic component (2) must be superimposed on the trend (1), a way of “guiding” the random walk of the Brownian motion towards the trend without destroying the desired properties is needed. We suggest using a sequence of non-overlapping random walks each starting from around the trend:

$$y(k\Delta t) = y^{\text{trend}}(k\Delta t) + \text{Norm}(0, \sigma^{\text{trend}}) \quad (4)$$

i.e., at each time step $k\Delta t, k = 0, 1, 2, \dots$, a Brownian motion process begins for Δt steps:

$$y^{\text{BM}}(k\Delta t + m) = y^{\text{BM}}(k\Delta t + m - 1) + \eta \quad (5)$$

where, $m = 1, 2, \dots, \Delta t - 1$. Then it stops and a new process begins. This completes our model of the intensity of the http requests. But before we can use it in our simulations of Internet caches we also need to define the popularity distribution of the requests. There is significant evidence in the literature [3, 4, 5] suggesting that the popularity distribution follows a Zipf’s-like law [6], where the relative popularity of the i^{th} most popular file is given by:

$$p_i^{\text{relative}} = \frac{1}{i^\alpha} \quad (6)$$

Therefore, we also need a random number generator that produces Zipf’s distributed numbers. We define it in the following way. First the total domain size N and the exponent α must be specified. Then the probability of selecting file i is given by:

$$p_i = \frac{i^{-\alpha}}{\sum_{j=1}^N j^{-\alpha}} \quad (7)$$

A uniform random number n is generated in the range between 0 and 1 (most programming languages have already defined uniform random number generators) and an index k is found such that the following inequalities hold:

$$\begin{aligned} n &\leq \sum_{j=1}^k p_j \\ n &> \sum_{j=1}^{k+1} p_j \end{aligned} \quad (8)$$

The index k is the desired random number coming from the specified Zipf's-like distribution.

3. Discussion and Future Work

To further understand Internet traffic, BT research laboratories are developing models of the caching mechanisms. Current work involves building dynamical models of Internet proxy caches. These models include both knowledge of the inner working of the cache management algorithms and popularity statistics models as observed from real data.

References

1. R. M. Lukose and B. A. Huberman, Surfing as a Real Option., paper presented at the Computational Economics Symposium, Cambridge, England, June 1998
2. S. Olafsson, A Stochastic Model for Internet Browsing, submitted to IEEE/ACM Transaction on Networks
3. Lee Breslau, Pei Cao, Li Fan, Graham Phillips and Scott Shenker, Web Caching and Zipf-like Distributions: Evidence and Implications., To appear in Proceedings of Infocom'99
4. Margo Seltzer, The World Wide Web: Issues and Challenges", presented at IBM Almaden, July 1996
5. C. A. Cunha, A. Bestavros, and M. E. Crovella, Characteristics of WWW Client-based Traces, Technical Report TR-95-010, Boston University Department of Computer Science, April 1995
6. G. K. Zipf, Human Behavior and the Principle of Least Effort, Addison-Wesley, Cambridge, MA, 1949

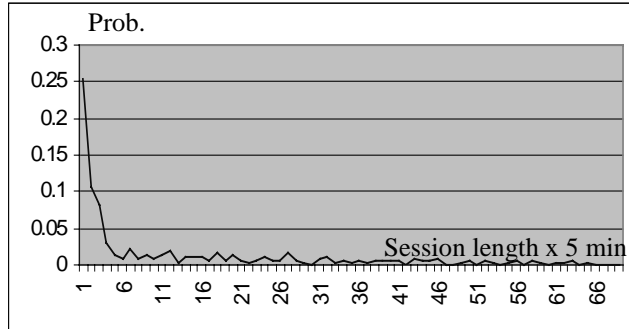


Fig. 1. Distribution of browsing session lengths measured over a period of twenty four hours at the university of Pisa, Italy (acknowledgements to Luigi Rizzo for providing the data). The graph comprises data from 114 clients.

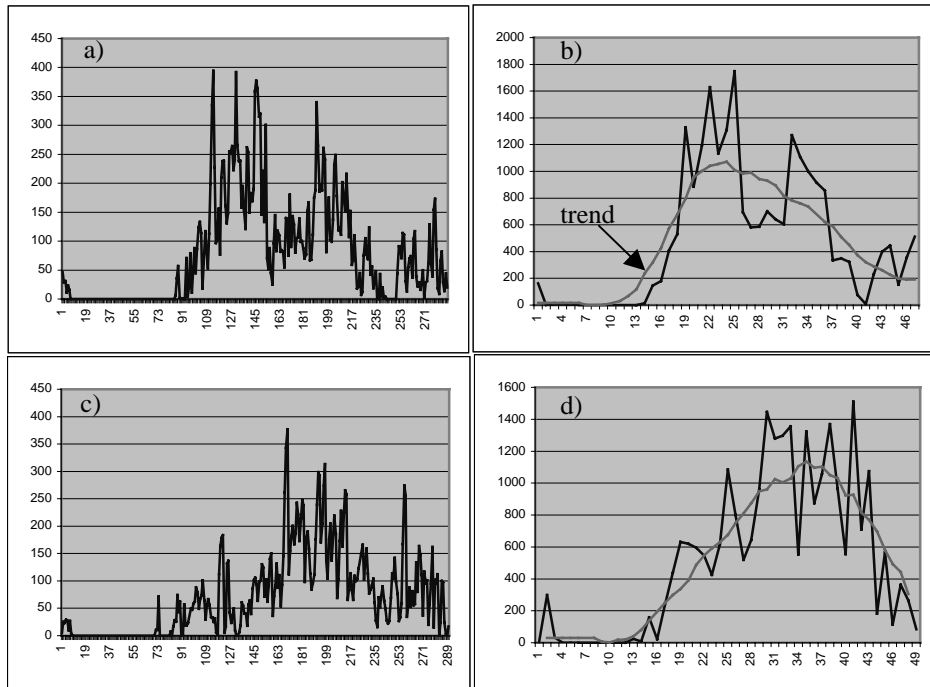


Fig. 2. Intensity of file requests aggregated over two non-overlapping time windows of 5(a) and 30(b) minutes. 2b shows an approximated trend using a moving average. The graphs use the same data as in fig. 1. 2c and 2d show the prediction of the model for a similar community.