Linking Cache Performance to User Behaviour.

Ian Marshall and Chris Roadknight
BT Research Laboratories, Martlesham Heath, Ipswich, Suffolk, UK. IP5 7RE
{marshall,roadknic}@drake.bt.co.uk

Abstract.

The performance of HTTP cache servers varies dramatically from server to server. Much of the variation is independent of cache size and network topology and thus appears to be related to differences in the user communities. Analysis of a range of user traces shows that, just like caches, individual users have highly variable hit rates, Zipf locality curves and show strong signs of long range dependency. In order to predict cache performance we propose a simple model which treats a cache as an aggregation of single users, and each user as a small cache.

Keywords. Traffic, Cache, Internet, Users

1.Introduction.

HTTP caching has been shown to deliver considerable benefits in network utilisation and user perceived performance [3, 1]. In order to optimise the benefits it is important to develop accurate models of cache behaviour. To our knowledge, accurate models have yet to be fully defined. The performance of HTTP cache servers varies dramatically from server to server and invariants have proved hard to identify. Much of the observed variability is independent of cache size [6] and network topology [2] and appears to be related to differences in the user communities of the caches. It therefore seems desirable to develop a good model of user behaviour. Unfortunately studies of the behaviour of individual users are not readily available. This is probably because it is not possible to analyse the behaviour of anything other than large (atypical) groups of users in the anonymised statistics published by some cache operators [e.g. ftp://ircache.nlanr.net/Traces/]. In order to build reliable statistics it is necessary to obtain long-term traces for small and medium users. If user request rates follow a single sided distribution, like many cache parameters, the smallest users will be by far the most numerous and will dominate results. However, if users request rates follow a Poisson distribution (as do telephony call rates [5]), it will be safe to ignore the smallest users.

In this paper we report the first results from a study of individual user behaviour that we have undertaken. In section 2 we illustrate the analysis that is possible using the published daily statistics from a large cache, and show that there is strong evidence that user request rates are Poisson. In section 3 we report some longer-term results for single users from several different caches, and show for the first time that large and moderate individual users have highly variable hit rates, Zipf locality curves and exhibit long range dependency. The smallest users do not generate large enough samples to enable meaningful analysis. In section 4 a simple model is proposed which models individual

users using only key cache performance parameters. Although this model is based on the behaviour of large and moderate users, it should generate reasonable predictions since the distribution of user request rates is not single sided and the smallest users can be safely ignored. In section 5 we illustrate how this simple model could be used to predict cache performance, and propose an empirical constant representing the fact that cache user communities are not random selections from the total user community.

2. Analysis of Daily statistics

The clearest link to user choices and behaviour in published cache statistics is the host popularity curve (or locality plot). This curve plots the number of requests for files from a given site, against the popularity ranking of the site. Previous studies [6] have shown that the curve usually approximates to Zipf's law. Zipf's First Law [11] states that if the frequencies of occurrence of each item are ranked by frequency then the frequency of the second most popular item will be half the frequency of the most common item. The frequency of the third most popular item will be proportional to a third of the frequency of the most popular, and so on. That is:

$$Frequency\ of\ Ranking_N = (Frequency\ of\ Ranking_1)/N$$

In order to get a clear picture of user activity however, it is necessary to know the target file for each request, rather than just the target host. This is because some users could access a large range of files on a small range of hosts, and some users could repeatedly access a small range of files on the same range of hosts. This would lead to identical popularity curves but very different results for metrics such as hit rate, and number of cached files. Unfortunately the statistics published by most caches do not provide any detail of the files accessed by individual users. However, we were able to obtain a log which reported the target URL for each request, from the operators of the Funet (Finnish University and Research Network) cache in Finland [http://www.funet.fi/funet/]. The Funet cache is a primary cache serving universities, polytechnics and some research organisations in Finland. The cache has about 2500 users, 500000 requests per day, a hit rate of ~30% and a popularity curve which is approximately Zipf. The cache appears, on these metrics, to be similar to many primary caches run by other universities or groups of academic institutions [e.g. http://www.swin.edu.au/proxy/usage/]. The analysis reported in this section is based on a log file for a single days activity, obtained from Finland. Analysis over a longer period was not possible as the logs are anonymised on a daily basis. In addition, many of the logged file requests are automatically generated by embedded objects such as .gif files, rather than by the user. As we are attempting to analyse user behaviour rather than the composition of web pages we have attempted to eliminate the auto-generated requests from the analysis. Therefore all requests which occurred within a second of the previous request and were made to the same host as the previous request have been filtered out of the data.

Figure 1 shows the user community of the cache sorted by number of requests generated in a single day (20/1/98). The logarithmic distribution of request rates is approximately Poisson with a mean of 101 requests (where the mean is defined as the inverse log of the mean of the logarithmic distribution). This observation means we can build a model without requiring analysis of the smallest users. We will, however, need to build up a

detailed understanding of users generating around 100 requests per day.

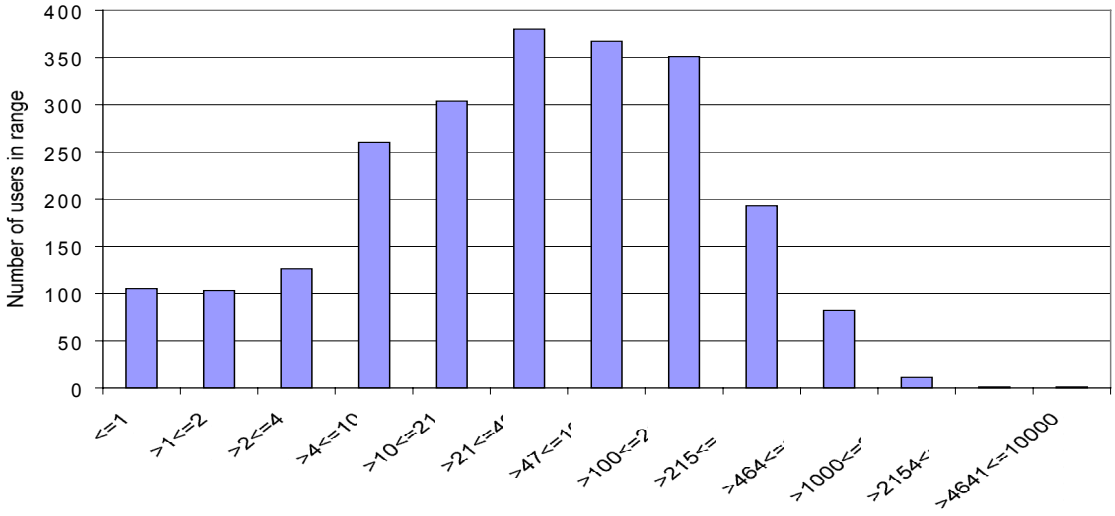

Figure 1. Distribution of user request rates on the Funet cache.

Figure 2 shows the host popularity data for the largest two users of the cache. A line with the gradient of Zipf's law is also shown as an aid to the eye. It is apparent that user 1 (no.of requests) visits fewer hosts than user 2 (no.of requests), and that whilst user 1 approximately follows Zipf's law user 2 does not. In fact user 2 shows weaker locality since his most popular sites are less popular than Zipf would predict. We chose to analyse large users at this stage because it was more practical and resulting statistics would be more sound.

Analysis of requests made over one day are not ideal, common sense would suggest that data from only one day will consistently underestimate the locality shown by users because daily visits to sites (e.g. news sites) will not be represented. Secondly smaller users generate samples which are too small after only one day, so the selection of users from a daily log can never be fully representative.

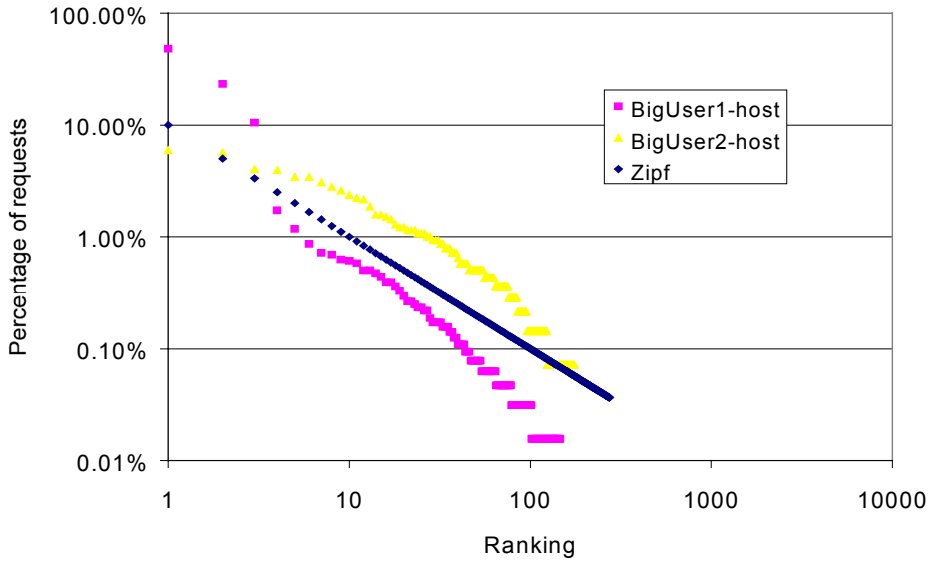

Figure 2. Popularity curve for hosts accessed by two users of the Finland cache.

In order to illustrate the need for data on file requests rather than just the usual data on host requests, the file request popularity curve for the same two users is shown in Figure 3, again with the Zipf gradient shown as an aid. It is clear that user 1 is visiting more pages than user 2, even though he visited fewer hosts, illustrating the importance of the file data. It is also clear that page request popularity does not follow Zipf's law, since both curves now show less locality than Zipf would predict. The reason for this is likely to be that data has only been analysed for one day and thus the popularity of the most frequently visited sites has been underestimated as discussed above. It also important to note that the request gradient for hosts will always be steeper than files because the former is usually an aggregate of the latter. This means that there is always more files requested than hosts visited and the most popular host is always more popular than the most popular file.

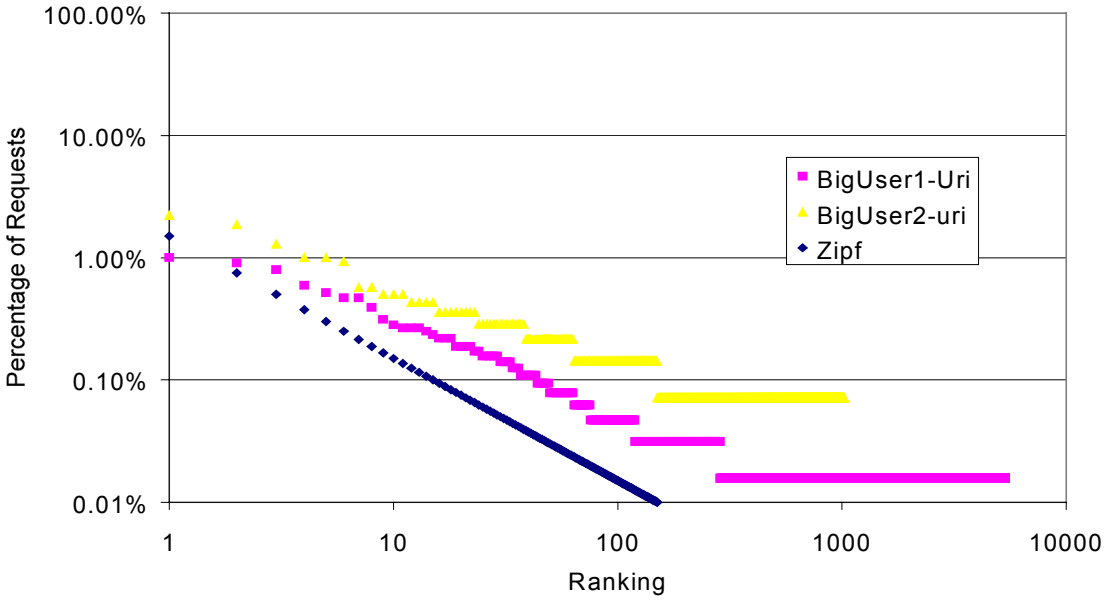

Figure 3. Popularity curves for files accessed by two users of the Finland cache.

Even with access to the file request data from the Funet cache we are unable to draw any strong conclusions. The reasons can be summarised as follows:

1) Only large users generate enough requests to be worthy of analysis and they are not likely to be representative of smaller users.

2) A user identity cannot be traced for long time periods since the data is freshly anonymised on a daily basis. This is particularly important as there is likely to be significant long range dependency in user activity based on the users memory which will be lost if only the activity of a single day is analysed.

In an attempt to overcome these issues we have assembled long term data sets from a range of caches. The initial analysis of some of the data is reported in section 3 below.

3.Analysis of long term statistics

The cache which is the source of the majority of the data presented in this section has been used by 5 individuals over the course of a year. The cache is small since only users who are members of our research team, and are happy for their logs to be analysed in non-anonymised form have been encouraged to use it. Over a 26-week period (25/8/97 – 22/2/98) the cache generated 14000 requests and sustained a hit rate of around 20%. The cache is sited at the subnet router for the team's network and is thus caching requests to sites in other buildings on the local Intranet. These local requests do not appear in the data for larger caches and should possibly be eliminated. In this paper these requests have not been eliminated as it is difficult to know where to stop eliminating sites from a global Intranet such as that operated by BT. We do not feel that the local requests

significantly alter the principal findings of the work.

	Client 1
requests	6545
hosts	389
uri	3535
hit rate	24.04%
Requests on Busiest Day	345
Number of Distinct Hosts on Busiest Day	27
Number of distinct URL's on Busiest Day	240

Table 1. Users statistics for biggest user of small scale cache.

Table 1 shows the usage statistics of the most consistent user of the cache over the 26-week period. This user was using the WWW through the cache whilst in the office, primarily as a research tool.

On days when the WWW was used the user spent between one and three hours browsing and generated up to 345 page requests in a session. The typical number of initial page requests (auto generated requests for embedded files have been filtered as above) on a working day was 100. This places the user in the most probable part of the usage distribution observed on the Finland cache. One cannot on this basis alone claim that the user is representative, but at least we can be sure we are analysing a user with significant differences to the users analysed in section 2.

On the basis of the amount of time this user spent in WWW based activity it seems unlikely that the requests attributed to the heavy users on the Finland cache were entirely generated by single human end users. 14% of the requests generated by our user were actually automated requests. The user was using Internet explorer and has experimented with the subscription mechanisms this browser provides. In the 26 week period the user had 6 subscriptions for a period of about one month. One of the subscriptions was checking the currency of bookmarked pages. Had these subscriptions been extended over the entire 26 weeks 50% of the traffic would have been "robot" generated. It is conceivable that an enthusiastic subscription user would have a much higher proportion of robot traffic. Table 2 shows the ten most popular sites for all three users.

Finland - User 1	Requests	Finland - User 2	Requests	BT - Client 1	Requests
www.msnbc.com	3070	freespace.virgin.net	83	www.sunday-times.co.uk	1057
www.microsoft.com	1487	photogallery.simplenet.com	79	local project management	245
ads.msn.com	671	website.yle.fi	56	www.zdnet.co.uk	234
investor.msn.com	110	us.imdb.com	55	ad.doubleclick.net	212
www.mungopark.com	75	www.unitedmedia.com	48	www.altavista.digital.com	206
travel.state.gov	55	altavista.telia.com	48	www.windows95.com	163
www.superbowl.com	46	icons.imdb.com	43	local experimental site	157
ad.preferences.com	44	www.geocities.com	39	local directory	139
www.zdnet.com	40	www.bubblegumtv.com	36	rc5stats.distributed.net	133
msnbc.com	39	free.prohosting.com	33	www.microsoft.com	109

Table 2. The 10 most popular hosts for 3 users.

As expected from previous reports [8] the favourite sites for our moderate user include a

local directory, some intranet pages, some news pages, some experimental research activity, and developer sites of large software companies. On the other hand the popular sites list for Finland user 1 is heavily biased towards the sites which are default subscriptions in internet explorer. This tends to support a supposition that some high request rates are robot assisted. It also leads us to question the benefits of the subscription mechanism which is potentially increasing web traffic by up to an order of magnitude. It is possible that Finland user 2 is a small subsidiary cache.

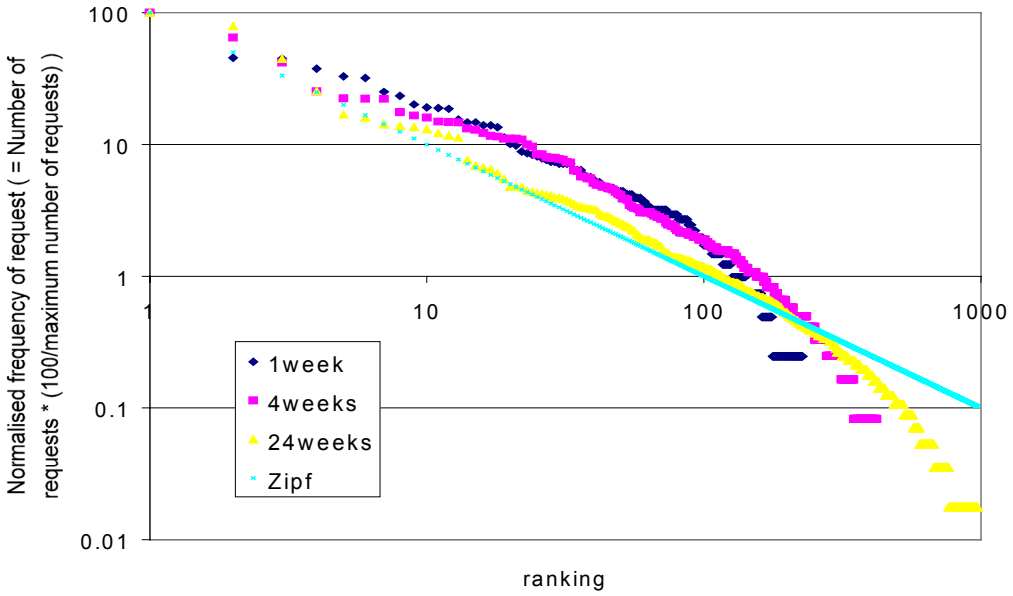

Figure 4. Comparison of popularity curves over different timespans.

Figure 4 shows the host popularity data for our cache analysed over progressively increasing periods of 1 week, one month and 6 months. As the sample size and time period increase the results tend towards following Zipf's law. It is not clear whether this is an artifact of small sample sizes or whether it indicates a very long-range dependency of more than 6 months. It is to be expected that there will be significant self-similarity at long timescales as the users' memory and interests are quite persistent, but it is also clear that for less popular sites the statistics will be very poor.

Figure 5 compares the page popularity curves for the two Finland users and for our local user (with auto generated requests filtered as before). It is clear from figure 6 that our local, moderate user has a popularity curve with a similar gradient to that of Finland user 2, but Finland user 1 has a curve with a steeper gradient. Despite the results being taken over a longer time period, our user still exhibits less locality than would be predicted by Zipf's law. We suspect that this may be because the lifetime of pages is significantly shorter than that of sites so the usage of the most popular pages cannot build up over long periods as it does for the most popular sites. It is equally possible that it is simply because our user is different. To resolve this question we analysed some long term user traces obtained from other sites.

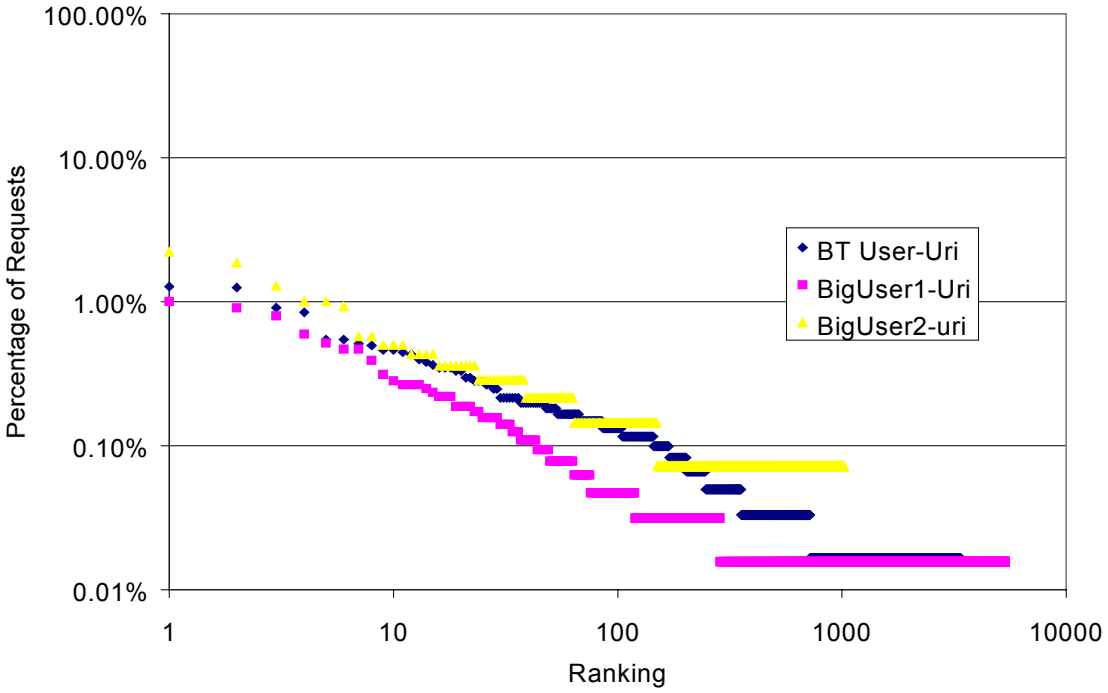

Figure 5. Popularity curves of three users at two sites.

Figure 6 shows the variance of the "auto hit rate" generated by a range of users plotted against sample size. We are confident that all the users plotted in this figure are genuine individual users. One user is the BT user analysed above. One additional user is derived from traces collected in 1995 at Boston University [6]. The remaining users are drawn from early analysis of logs provided by Research Machines plc (RMPLC). RMPLC [http://www.rmplc.net] is an internet service and content provider to schools and colleges in the UK, they also have a small number of modem users who are assigned a static IP address. We analysed the hit rate statistics of two of the modem users, during a period of 6 weeks in early 1998. All of the users are generating around 100 requests per day, and are thus near the peak of the request rate distribution found in the Finland data.

The "auto hit rate" is the rate of request for files previously requested by the user and successfully cached. As observed elsewhere [10], for the aggregate hit rate on much larger caches, the variance does not decay rapidly with sample size. The curves are clearly linear and the Hurst parameters can be estimated from the data and fall in the range 0.55 – 0.7 (with estimation errors of 0.1 due to the use of small samples). All the users we have analysed thus exhibit the statistical anomalies that have been interpreted elsewhere as self similarity [6]. The data presented does not prove self similarity in a strict mathematical sense as the range of the curves is small, due to the small sample sizes. However, the data is sufficient to demonstrate, for the first time, that individual users exhibit the same anomalies as caches. The anomalies can be safely interpreted as a type of long range dependency, which we postulate originates in the memory of the users.

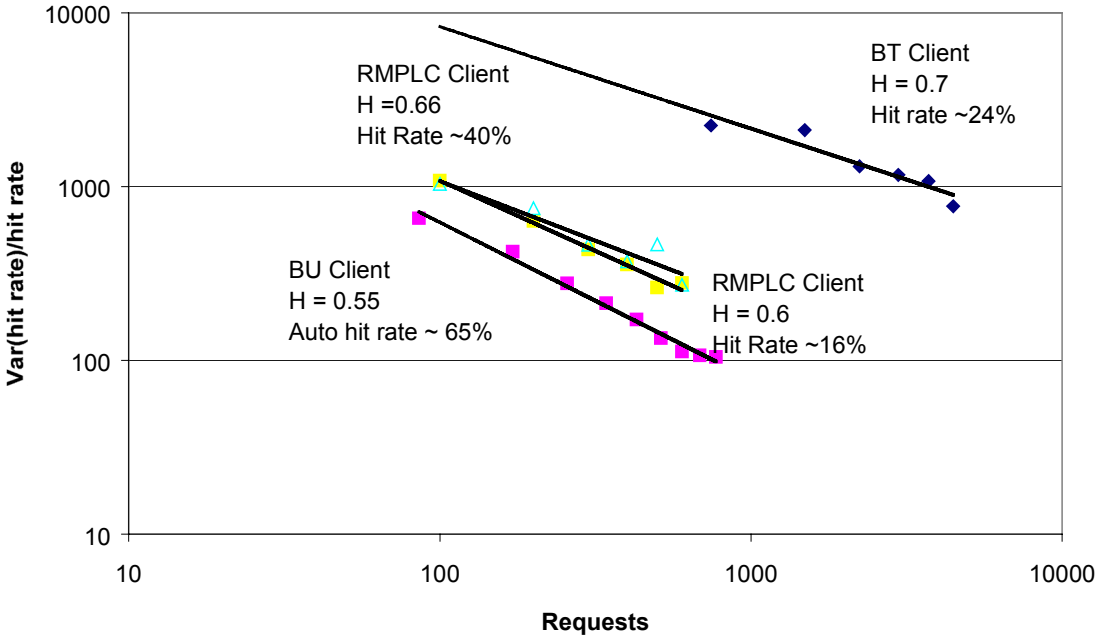

Figure 6. Decline in hit rate variance with increasing sample size.

It is clear from fig. 6 that the BT user is not obviously atypical. It is also clear that the hit rates of the users are just as variable as those of caches [10] and fall in a remarkably similar range (16% - 65%). Use of long term traces has thus enabled us to find that (at least large and moderate) users have hit rates, locality curves and long range dependency, all of which are indistinguishable from those of caches.

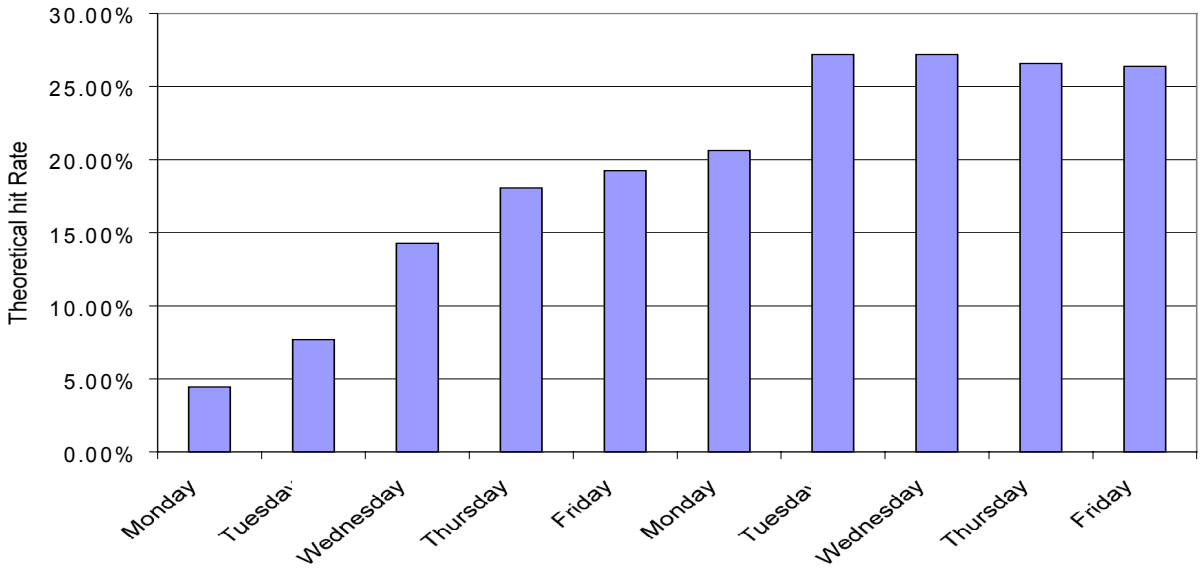

Figure 7. Change in theoretical hit rate over 10 working days.

In figure 7 we show the build up of the auto hit rate of the BT user with time. The rate plateaus after just over 1 working week, indicating that the characteristic time of the dominant long-range dependency should be around 1 week. The data also indicates that the ideal length of user trace is at least 1 week even for large users. For moderate users with ~100 requests per day a 25 day sample would be ideal to overcome stochastic effects. Users with 20 requests per day would ideally require a 6 month long trace. Users with less than 10 requests/day will be hard to analyse accurately as the sample would need to be 250 days long, i.e. it would exceed the typical lifetime of web pages. However since these users represent the tail of the distribution and have minimal impact on cache performance they can almost certainly be safely ignored.

Based on the results reported in sections 2 and 3 we now move on to propose a simple model of users

4. User model

The simplest conceivable model of a web user has one assumption and one parameter. The assumption is that page popularity will follow Zipf's law as do other popularity curves for human beings [4]. The parameter is usage level. The results we have reported indicate that this is not sufficient to explain the variability of user behaviour – since in practice the users do not follow Zipf's law (although they do appear to follow a heavy tailed Pareto curve), and they have highly variable hit rates. In addition it is clear that users are a significant source of long range dependency. This leads us to the assertion that at least 3 parameters are required to model the observed behaviour

1) usage

2) Auto hit rate

3) Hurst parameter derived from auto hit rate

Of course for modelling situations where the burstiness is not important the third parameter can be neglected and we are left with a simple two-parameter model

Auto hit rate is chosen as the second parameter because it is the most obvious single number metric for the locality of users requests, and it has no dependencies on other users or the properties of the WWW. In addition it is plausible to argue that it will show a distribution with a well defined maximum likelihood, as does the usage parameter. In the next section we illustrate how a user model based on the above parameters can be used to predict cache performance.

5. Community model

Based on a two parameter user model as proposed in section 4 and additional input from the observations we can propose a method for building user community models. This method is intended as an illustration of our objectives rather than as a definitive proposal due to the small amount of data available.

Assumptions

1) the distribution of request rates is Poisson with a mean of 101 as observed in the Finland data

2) the distribution of auto hit rates is Poisson with a mean of 0.2 (the observed auto hit rates of our user) and independent of request rate

If the community is mostly human the above assumptions are likely to be true. However if there is a substantial proportion of robot traffic there are likely to be at least two independent distributions with different means. The impact of robots will be the subject of a further study.

Since we have assumed Poisson distributions it is sufficient to have studied users enough to characterise the means of the above distributions accurately. The community model is not dependent on individual user properties. However, we also need a characterising parameter – the degree of overlap between users in a given community. This parameter is required because communities are not random samples from the overall user population, but are subject to strong locational bias. Examples of bias include factors like language, dominant local industries and demographics (small towns with universities are dominated by single people in their early twenties). We propose an overlap parameter, ϖ, which can be derived empirically using

$$\varpi = 1 - \frac{observedno.ofcachedpages \div workingdays}{no.ofusers \times (meanrequestrate - meanrequestrate \times meanautohitrate)} \quad (1)$$

We anticipate that caches with similar user communities will have very similar values for ϖ, and that ϖ will have a useful probability distribution, which can be treated as the sum of the distributions for ϖ for a small set of community types (e.g. researchers, home users, business users, robots, mixed). Each type represents a different mixture of user objectives, and should thus represent a defined market sector for WWW products. We are currently attempting to calculate ϖ for a wide range of caches with published data in order to validate this hypothesis.

ϖ can be used to make predictions, for example it can be shown that:

$$No.ofhits = meanrequestrate \times meanautohitrate \times no.ofusers + \varpi \times (no.ofusers - 1) \times (meanrequestrate - meanrequestrate \times autohitrate) \quad (2)$$

Using $meanrequestrate = 101$

$meanautohitrate = 0.2$

and taking observations from the Finland cache for *observedno.of cachedpages = 161628, workingdays* = 1, and *no.ofusers = 2284,* we derive $\varpi = 0.124$

Applying this value of ϖ to the Finland cache we get *no.ofhits* = 69010

The true value of *no.of hits* was 70999 so our model has predicted a reasonable answer.

6. Conclusions

Based on a preliminary analysis of user behaviour in daily cache logs and in long term traces we have shown that individual users, with moderate daily request rates, exhibit hit rates, request locality and long range dependencies similar to those observed in caches. We have also shown that user request rates follow a Poisson type distribution, so it is not necessary to understand the behaviour of the small number of low level users as their requests make up an insignificant proportion of overall requests to a cache. Based on the observations a simple model of users has been proposed, and we have illustrated how this model could be used. A parameter was proposed to account for the systematic biases in the membership of real cache user communities. This parameter may be sufficient to characterise the observed differences between real caches.

7. Acknowledgements.

We would like to thank Pekka Järveläinen for supplying us with anonymised logs for the Funet proxy cache, and Simon Rainey of RM plc for supplying us with the logs of their caches in the UK.

8. References

[1] M. Abrams, C.R. Standridge, G. Abdulla, S. Williams and E.A. Fox. Caching Proxies: Limitations and Potentials. Proc. 4th Inter. World-Wide Web Conference, Boston, MA, Dec. 1995.

[2] M. Arlitt and C. Williamson. Internet web servers: Workload Characterization and performance implications. IEEE/ACM Transactions on Networking. Vol. 5, No. 5. October 1997.

[3] M. Baentsch, L. Baum, G. Molter, S. Rothkugel and P. Sturm. Enhancing the web's infrastructure: From caching to replication. IEEE Internet Computing. March 1997. P. 18-27.

[4] J. Beran. Statistical methods for data with long-range dependance. Statistical Science. Vol. 7. No. 4. p404-427.

[5] J. Boucher. Voice teletraffic systems engineering. Chapter 2. ISBN 0-89006-335-4.

[6] 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.

[7] B.M. Duska, D. Marwood and M.J. Feeley. The measured access characteristics of WWW proxy caches. 1997. URL: http://www.cs.ubc.ca /spider/marwood/Projects/SPA/

[8] James Gwertzman and Margo I. Seltzer: People, Places, and Things: The Next Generation Web. COMPCON 1996: 65-70

[9] I. Marshall and C. Roadknight. Characteristic times for long range dependency in WWW page requests – in preparation

[10] C. Roadknight and I. Marshall. Variations in cache behavior. Proceedings of WWW7 Comp. Nets and ISDN systems, 30 (1998) pp733-735.

[11] G. K Zipf. Human Behavior and the Principle of Least Effort. Addison-Wesley, Cambridge, MA, 1949.