How the web works

Table of contents

This section is about what happens when somebody connects to your web site, and what statistics you can and can't calculate. There is a lot of confusion about this. It's not helped by statistics programs which claim to calculate things which cannot really be calculated, only estimated. The simple fact is that certain data which we would like to know and which we expect to know are simply not available. And the estimates used by other programs are not just a bit off, but can be very, very wrong. For example (you'll see why below), *if your home page has 10 graphics on, and an AOL user visits it, most programs will count that as 11 different visitors!*

This section is fairly long, but it's worth reading carefully. If you understand the basics of how the web works, you will understand what your web statistics are really telling you.

The basic model

Let's suppose I visit your web site. I follow a link from somewhere else to your front page, read some pages, and then follow one of your links out of your site.

So, what do you know about it? First, I make one request for your front page. You know the date and time of the request and which page I asked for (of course), and the internet address of my computer (my host). I also usually tell you which page referred me to your site, and the make and model of my browser. I do not tell you my username or my email address.

Next, I look at the page (or rather my browser does) to see if it's got any graphics on it. If so, and if I've got image loading turned on in my browser, I make a separate connection to retrieve each of these graphics. I never log into your site: I just make a sequence of requests, one for each new file I want to download. The referring page for each of these graphics is your front page. Maybe there are 10 graphics on your front page. Then so far I've made 11 requests to your server.

After that, I go and visit some of your other pages, making a new request for each page and graphic that I want. Finally, I follow a link out of your site. You never know about that at all. I just connect to the next site without telling you.

Caches

It's not always quite as simple as that. One major problem is caching. There are two major types of caching. First, my browser automatically caches files when I download them. This means that if I visit them again, the next day say, I don't need to download the whole page again. Depending on the settings on my browser, I might check with you that the page hasn't changed: in that case, you do know about it, and analog will count it as a new request for the page. But I might set my browser not to check with you: then I will read the page again without you ever knowing about it.

The other sort of cache is on a larger scale. Almost all ISP's now have their own cache. This means that if I try to look at one of your pages and *anyone else from the same ISP* has looked at that page recently, the cache will have saved it, and will give it out to me without ever telling you about it. (This applies whatever my browser settings.) So hundreds of people could read your pages, even though you'd only sent it out once.

What you can know

The only things you can know for certain are the number of requests made to your server, when they were made, which files were asked for, and which host asked you for them.

You can also know what people told you their browsers were, and what the referring pages were. You should be aware, though, that many browsers lie deliberately about what sort of browser they are, or even let users configure the browser name. Also, a few browsers send incorrect referrers, telling you the last page that the user was on even if they weren't referred by that page. And some people use "anonymizers" which deliberately send false browsers and referrers.

What you can't know

1. *You can't tell the identity of your readers*. Unless you explicitly require users to provide a password, you don't know who connected or what their email addresses are.
2. *You can't tell how many visitors you've had*. You can guess by looking at the number of distinct hosts that have requested things from you. Indeed this is what many programs mean when they report "visitors". But this is not always a good estimate for three reasons. First, if users get your pages from a local cache server, you will never

know about it. Secondly, sometimes many users appear to connect from the same host: either users from the same company or ISP, or users using the same cache server. Finally, sometimes one user appears to connect from many different hosts. AOL now allocates users a different hostname for *every request*. So *if your home page has 10 graphics on, and an AOL user visits it, most programs will count that as 11 different visitors!*

3. *You can't tell how many visits you've had*. Many programs, under pressure from advertisers' organisations, define a "visit" (or "session") as a sequence of requests from the same host until there is a half-hour gap. This is an unsound method for several reasons. First, it assumes that each host corresponds to a separate person and vice versa. This is simply not true in the real world, as discussed in the last paragraph. Secondly, it assumes that there is never a half-hour gap in a genuine visit. This is also untrue. I quite often follow a link out of a site, then step back in my browser and continue with the first site from where I left off. Should it really matter whether I do this 29 or 31 minutes later? Finally, to make the computation tractable, such programs also need to assume that your logfile is in chronological order: it isn't always, and analog will produce the same results however you jumble the lines up.
4. *Cookies don't solve these problems.* Some sites try to count their visitors by using cookies. This reduces the errors. But it can't solve the problem unless you refuse to let people read your pages who can't or won't take a cookie. And you still have to assume that your visitors will use the same cookie for their next request.
5. *You can't follow a person's path through your site*. Even if you assume that each person corresponds one-to-one to a host, you don't know their path through your site. It's very common for people to go back to pages they've downloaded before. You never know about these subsequent visits to that page, because their browser has cached them. So you can't track their path through your site accurately.
6. *You often can't tell where they entered your site, or where they found out about you from*. If they are using a cache server, they will often be able to retrieve your home page from their cache, but not all of the subsequent pages they want to read. Then the first page you know about them requesting will be one in the middle of their true visit.
7. *You can't tell how they left your site, or where they went next*. They never tell you about their connection to another site, so there's no way for you to know about it.
8. *You can't tell how long people spent reading each page*. Once again, you can't tell which pages they are reading between successive requests for pages. They might be reading some pages they downloaded earlier. They might have followed a link out of your site, and then come back later. They might have interrupted their reading for a quick game of Minesweeper. You just don't know.
9. *You can't tell how long people spent on your site.* Apart from the problems in the previous point, there is one other complete show-stopper. Programs which report the time on the site count the time between the first and the last request. But they don't count the time spent on the final page, and this is often the majority of the whole visit.

Real data

Of course, the important question is how much difference these theoretical difficulties make. In a recent paper (*World Wide Web*, **2**, 29-45 (1999): PDF 228kb), Peter Pirolli and James Pitkow of Xerox Palo Alto Research Center examined this question using a ten day long logfile from the xerox.com web site. One of their most striking conclusions is that different commonly-used methods can give very different results. For example, when trying to measure the median length of a visit, they got results from 137 seconds to 629 seconds, depending exactly what you count as a new visitor or a new visit. As they were looking at a fixed logfile, they didn't consider the effect of server configuration changes such as refusing caching, which would change the results still more.

Conclusion

The bottom line is that HTTP is a stateless protocol. That means that people don't log in and retrieve several documents: they make a separate connection for each file they want. And *a lot of the time they don't even behave as if they were logged into one site*. The world is a lot messier than this naïve view implies. That's why analog reports requests, i.e. what is going on at your server, which you know, rather than guessing what the users are doing.

Defenders of counting visits etc. claim that these are just small approximations. I disagree. For example, almost everyone is now accessing the web through a cache. If the proportion of requests retrieved from the cache is 50% (a not unrealistic figure) then half of the users' requests aren't being seen by the servers.

Other defenders of these methods claim that they're still useful because they measure *something* which you can use to compare sites. But this assumes that the approximations involved are comparable for different sites, and there's no reason to suppose that this is true. Pirolli & Pitkow's results show that the figures you get depend very much on how you count them, as well as on your server configuration. And even once you've agreed on methodology, different users on different sites have different patterns of behaviour, which affect the approximations in different ways: for example, Pirolli & Pitkow found different characteristics of weekday and weekend users at their site.

I've presented a somewhat negative view here, emphasising what you can't find out. Web statistics are still informative: it's just important not to slip from "this page has received 30,000 requests" to "30,000 people have read this page." In some sense these problems are not really new to the web -- they are present just as much in print media too. For example, you only know how many magazines you've sold, not how many people have read them. In print media we have learnt to live with these issues, using the data which are available, and it would be better if we did on the web too, rather than making up spurious numbers.

Acknowledgements and further reading

Many other people have made these points too. While originally writing this section, I benefited from three earlier expositions: *Interpreting WWW Statistics* by Doug Linder; *Getting Real about Usage Statistics* by Tim Stehle; and *Making Sense of Web Usage Statistics* by Dana Noonan. (The last two don't seem to be available on the web any more.)

Another, extremely well-written document on these ideas is *Measuring Web Site Usage: Log File Analysis* by Susan Haigh and Janette Megarity. Being on a Canadian government site, it's available in both English and French. Or for an even more negative point of view, you could read Why Web Usage Statistics are (Worse Than) Meaningless by Jeff Goldberg.

Information available at www.analog.cx

Why web usage statistics are (worse than) meaningless

Note: This document was originally written in 1995 to explain the stats situation at Cranfield University where I worked at the Cranfield Computer Centre. It was initially intended for local users there, but quickly gained popularity (notoriaty?) elsewhere. The content has had little changes and updates since then.

1. There is no discussion of cookie tracking (yet) in this document.
2. I had over-estimated the extent to which caching (and hierarchical caching) would be used.
3. Cranfield University has a proud history of leadership in the web. It was one of the very first UK sites to have a webserver at all (in 1993), was at the forefront of the UK caching effort, and in enabling individual users to publish on the web (early 1994). I am grateful that they permitted what way essentially my personal rant to be hosted so prominantly for so long, and have provided a long term redirect to this page.
4. On re-reading my original, I see that this document is a bit hyperbolic. So be it. It is after all an acknowledged rant.

Web usage statistics, such as those produced by programs such as analog cannot be used to make strong inferences about the number of people who have read a website or webpage. Although those who compile these statistics usually try to make this clear, people still insist on misusing them to make overly strong inferences. Attaching meaning to meaningless numbers is worse than not having the numbers at all. When you lack information, it is best to know that you lack the information. Web statistics may give the user a false sense of knowledge which can be worse than being knowingly ignorant.

A useful analogy is with putting up advertising posters. You will never really know how many people have noticed them or read them.

It is not enough to say that the statistics should be taken with a grain of salt; they should be taken with a salt lick. If you want to understand why *no inference about the number of people reading this pages can be made from web statistics* read on. Otherwise, you may wish to just trust that statement or may wish to skip to the section on Quick Questions and Answers.

What web stats are really good for?

Web stats are useful for web administrators to get a sense of the actual load on the server. This is useful for diagnostics and planning, and for detecting unusual behaviour that may require planning action. The goal of the administrator is to keep the server running smoothly under expected loads, while improving the speed and reliability of obtaining documents from the site. The best way to achieve this is to have browsers retrieve documents from places closer to where they will be used (and even from memory) than to get them from the disk on the server. It is only when the file is retrieved from the server that the server has the ability to keep track of the access.

Caching: Essential for the web and disastrous for statistics

Let's take a fictitious example of what might happen when someone in Nome, Alaska, say at Nome Community College (this would be a polytechnic in the UK), wants to read Cranfield's Prospectus. The user would somehow select the URL with his/her browser, which would then try the following.

Browser Cache

The particular instance of the browser will look in its own memory (or what it may have saved on the its local disk).

If it finds the page corresponding to the sought for URL there it will not go any further, and our site will never know that the request was made.

Local site cache

If the page was not in the browser cache, the browser may look to its site cache. That is, if someone at the user's same site recently retrieved the page, it may be available to the user there.

If it finds the page corresponding to the sought for URL there it will not go any further, and our site will never know that the request was made.

Local regional cache

The site cache may be configured to look in a local regional cache, say at the University of Alaska, Nome campus which might provide a caching service for smaller sites around Nome.

If it finds the page corresponding to the sought for URL there it will not go any further, and our site will never know that the request was made.

Large regional cache

The local regional cache may be configured to look in a large regional cache, say in Fairbanks Alaska, which might provide caching for sites in Alaska that use it.

If it finds the page corresponding to the sought for URL there it will not go any further, and our site will never know that the request was made.

The Cranfield accelerator

An accelerator is an out-going cache for a site. When a document is requested from the site, the accelerator sees whether it has it stored (it stores them in ways much faster to find and retrieve then the server does with files in the directory structure) and serves that up.

While it would be possible to have the accelerator keep a record of which files it served up and to whom, this would defeat the purpose, because it would require a disk operation to make that record.

> In addition to over-estimating the degree of caching that would be in place, this last step about accelerators is also no longer relevant. The accelerator was needed when Cranfield was running the original CERN server over an AFS filesystem. Given the nature of modern web server set-ups, accelerators are no longer needed.

Now that you have an idea of what caching is, you are in a better position to understand why it is impossible to make any inference about numbers of people reading your pages from web statistics. But there is more to come described in the section on multiple hits per users. What is necessary to understand about caching is that some users may go through a long and efficient cache chain (as described in the example) and other users may not. Much of this depends on how their site is set up or how they set things up themselves.

One user many hits

Imagine (in the extreme case) a user who is doing no caching whatsoever. Now if that user comes across the Cranfield Home Page 20 times while browsing around the Cranfield pages that will count as 20 hits. Remember the statistics are about accesses, not about people.

Big pages are little pages

When comparing hits for different different directories, it is important to note how documents are structured. If you have a directory with a single document on one hand, and on the other you have another directory with the same amount of real content broken in to twenty smaller documents, you will find far more hits into that second section.

Quick Questions and Answers about web statistics

Most of everything listed here is either mentioned above or can be inferred from the explanations above. If there is a question that you would like to see added to this list, or if you have other comments on this document, please use the form at the end to submit queries. [Sorry, that form is now defunct.]

A quick list of the questions is provided here.

?? Can stats be used to assess changes over time?

?? Can stats be used to assess relative popularity in different Internet domains such as .ac.uk, or .jp?

?? Can stats be used to assess relative popularity of different pages?

?? Is there some multiplier which can applied to the stats to get more meaningful results?

?? Can I ensure that my document is never cached?

?? Can I put counters in my page?

?? Can we get stats from the sites that do caching?

?? Can I infer from stats a minimum number of readers?

?? How can I gauge interest in pages?

?? If web stats are so bad why are they kept at all?

?? Then why make the stats public?

?? Is this all an just excuse to avoid the work of maintaining stats?

Can stats be used to assess changes over time?

Not really. The number of individuals and sites using caches is rising all the time, as is the amount of disk space and memory used for caching. When the Cranfield Accelerator goes live (early November, 1995), there should be an actual drop our server stats, while an increase in accesses, due to increased speed and reliability of the server. Caching has been on the rise for more than a year now. Even so, loads on systems (including ours) have gone up dramatically.

Can stats be used to assess relative popularity in different Internet domains such as .ac.uk, or .jp?

Unfortunately not even this is possible. Suppose for example that Japan has a very high level of regional and national caching while Singapore does not (the example is fictitious). Under these circumstances, web statistics might show more accesses from Singapore than from Japan even if more people in Japan read our pages.

A clear example of this is the number of accesses from "numerical domains" that have recently started to top various lists. These are accesses from sites that don't have proper reverse DNS listings. Such sites are probably misconfigured single user machines, where either the particular machine that is used in misconfigured or the organisation they belong to has not straightened out its machine names properly. It is reasonable to assume that those running such misconfigured systems are far more likely to not have configured their proxies correctly, so far less caching will be seen from those sites.

Can stats be used to assess relative popularity of different pages?

Not really. The more popular pages will cache more, meaning that real differences between page hits will be dramatically distorted. It is probably safe to say this if one page shows more hits then another that there really were more accesses to that page, but there are circumstances under which even that weak inference won't be true.

Is there some multiplier which can applied to the stats to get more meaningful results?

Not really. This is because any such multiplier would have to differ from page to page and differ from access region to access region.

Can I ensure that my document is never cached?

Yes you can. There are several ways to do so, and there are some circumstances for which it is even legitimate, but to do so merely to get better stats is seriously misguided. This is for two reasons:

1. You will make your page (much) harder for people to get to and add to network traffic unnecessarily.
2. If someone fails to reach your page at our site, they may give up on the site all together. Thus hard to get at pages (unless there is a clear reason for them being such) will be unfair to other providers at the site.

Quiet embarrassingly, many of the pages on this site don't normally cache properly. This is because I had some technical difficulties with my configuration of server side includes and the so-called "XBitHack". I've fixed that now, but now have to fix dozens of documents to use things properly.

Can I put counters in my page?

You may have noticed some pages with web counters. There are basically two ways to put them in your page: the wrong way and the very wrong way. The wrong way merely doesn't work and will not be more useful than normal statistics. The very wrong way is counter productive because it subverts the caching mechanism which is not a good idea just to get statistics.
Please note that even if you think that statistics can be made useful, counters on individual pages are displayed to the reader, who isn't in the position to make the various adjustments needed to get some sense of true readership.

Can we get stats from the sites that do caching?

Yes and no, but mostly no. There are two reasons for "mostly no". One is simply that there are too many small caches out there which may have cached our stuff (including the browser software internal cache). Clearly not all of these are going to send us records on a regular basis which we would then have to incorporate into all of the other records to process statistics.
The other reason for "mostly no" is that even the large caches are willing to only send a byte count. That is, one major UK cache is considering sending out on a monthly basis how many bytes of data they served up in our name.
We must remember that the caches are doing us a favour by making our pages much easier to reach. We cannot ask them to take on a task that would degrade the service or place an additional administrative, disk, memory and CPU load on them. Without caching, the web would have collapsed long ago.

Can I infer from stats a minimum number of readers?

Yes and no. If by minimum you mean "at least one" then yes. If you have 400 hits from Japan then you can conclude that during that period you had at least one reader from Japan. You *cannot* infer that there were at least 400 readers, because the same reader may hit a page many times in a short period of time.
So, the only certain inference that can be made is that there was at least one from a particular domain, or for a particular page.

How can I gauge interest in pages?

One way is to set up Mail Reply Forms in your pages like the one at the end of this document. Of course many more people will read your pages then will complete the form, but the form can be used to judge serious interest. Most people will, however, not fill out a form unless the

think they will get some sort of useful response, even if they read the document seriously. (Did you fill out the form for this document?).
Setting up these forms is not as difficult to do as it first appears, and courses are offered on it by the computing centre staff.

If web stats are so bad why are they kept at all?

They are useful for system administrators to judge the actual load on the server. The section on what stats are good for contains more information.

Then why make the stats public?

Popular demand. It is not the computer centre's job to deny users some service just because we know the request to be misguided. Attempts to eliminate this statistics from the system have meet with complaints. However, no great effort will be put into maintaining statistics or access to them either. It is hoped that this document will make it easier for the computer centre to withdraw statistics altogether, except for what is required for system maintenance.

Is this all an just excuse to avoid the work of maintaining stats?

No. But you may have noticed that many of the individual problems and difficulties could be *partially* mitigated by collecting and using more information (from some caches for example or times of requests) and using that to make very rough estimates of various correction factors. It would take serious statistic analysis of the sort that professional market research firms may be able to undertake and still the estimates (and relative hits on pages or from regions) would remain iffy. Performing complicated analyses on dubious data only compounds the problem, and the marginal utility would be negative (ie, the large amount of extra effort would not be justified by the tiny gain in meaningfulness of the statistics).

Time to ask your questions

> When this page was hosted by Cranfield there was a form for mailing comments. I have disable that since moving this document to its current location, because (a) I don't have as good a mailform system as was available at Cranfield, (b) there are spam/privacy concerns about collecting unconfirmed email addresses, which I hadn't considered in 1995 for what was initially intended as an internal document, (c) this was partially an attempt to promote the use of Mailforms at Cranfield, and (d) history has shown that I am often not very good at responding to the queries that I get.

Information available at www.goldmark.org

Measuring Web Site Usage: Log File Analysis

by Susan Haigh and Janette Megarity

Network Notes #57
ISSN 1201-4338
Information Technology Services
National Library of Canada
August 4, 1998

1.0 Introduction

As more organizations view the Web as an integral part of their operations and external communications, interest in the measurement and evaluation of Web site usage is increasing.

Server logs can be used to glean a certain amount of quantitative usage information. Compiled and interpreted properly, log information provides a baseline of statistics that indicate use levels and support use and/or growth comparisons among parts of a site or over time. Such analysis also provides some technical information regarding server load, unusual activity, or unsuccessful requests, and can assist in marketing and site development and management activities.

2.0 Web Site Usage: The Broader Picture

Usage analysis could involve detailed study of a sweeping range of questions: not just what, when, and by whom, but also how and why the information was sought and used (or not). Assessing Web site use in a meaningful manner is not a trivial undertaking. It is essential to begin by determining the questions on usage that must be answered, then to choose one or more appropriate evaluation mechanisms to provide meaningful answers.

Log analysis is only one of several such mechanisms. Qualitative methods of data collection, such as user surveys, focus groups, and other feedback mechanisms, can gather user opinions on site content, navigation, or look-and-feel, as well as assess user satisfaction and the reasons that users visited the site or navigated as they did. A site's usability--which will affect both rate and manner of use--can be evaluated through various methods to reveal whether the site is accessible, easy to navigate and appealing to users. Benchmarks against which to compare or evaluate use figures makes them more meaningful. How does my site's growth compare with overall Web growth rates? What volatility of use levels is normal, or how much can be attributed to our promotional efforts? Can we find sites that are comparable to ours, and which use their server logs with similar parameters and care?

This *Network Notes* focuses on log file analysis as one quantitative research method for usage analysis, providing an overview of what can and can not be mined from the data, and the software tools that are currently available to support log analysis.

3.0 What's in a Log File

Every communication between a client browser and a Web server results in an entry in the server's log recording the transaction. A busy Web site, such as that of the National Library of Canada, generates hundreds or thousands of log entries per hour and compiles them in a log file. The data captured in a log file vary according to the type of server used and the log file format(s) it supports. Most widely employed are the common log file format and the combined or extended log file format. In general, a log file entry contains:

the address of the computer requesting the file

the date and time of the request

the URL for the file requested

the protocol used for the request

the size of the file requested

the referring URL

the browser and operating system used by the requesting computer.

Two log file entries are shown below. The first is a request for a copyright message made from a bibliographic record displayed from National Library's catalogue, resAnet. The second requests an image embedded on a page in the National Library's "Celebrating Women" digital product. Both requests were logged at four seconds after midnight on July 24, 1998.

```
192.117.240.3 - - [24/Jul/1998:00:00:04 -0400]
"GET /10/3/a3-160-e.html HTTP/1.0" 200 2308 "http://www.amicus.nlc-
bnc.ca/wbin/resanet/itemdisp/l=0/d=1/r=1/e=0/h=10/i=11683503"
"Mozilla/2.0 (compatible; MSIE 3.01; Windows 95)"
208.145.1.13 - - [24/Jul/1998:00:00:04 -0400]
"GET /icons/etb3.gif HTTP/1.0" 200 443
"http://www.nlc-bnc.ca/2/12/h12-221-e.html" "Mozilla/4.0 (compatible; MSIE 4.01; Windows 95)"
```

4.0 What Can You Learn From a Log File?

Data available from a log file can be compiled and combined in various ways, providing statistics or listings such as:

number of requests made ("hits")

total files and kilobytes successfully served

number of requests by type of file, such as HTML page views

distinct IP addresses served and the number of requests each made

number of requests by domain suffix (derived from IP addresses)

number of requests for specific files or directories

number of requests by HTTP status codes (successful, failed, redirected, informational)

totals and averages by specific time periods (hours, days, weeks, months, years)

URLs from which user came to the site (referring pages)

browsers and versions making the requests.

5.0 What Can't You Learn from a Log File?

The shortcomings of log files as usage indicators fall into three main categories: certain types of usage data are not logged; the data that are logged may be incomplete; and it is tempting to draw unsound inferences from some of the data.

5.1. Data not captured in the logs

?? Individuals' identities: Except for transactions that have required authorization (passwords), no data recorded in server logs reveal an individual user's name or any other individual identifier, an e-mail address, for example.

?? Number of users: A "user", as reflected in a log, is an IP address--a computer. This does not necessarily correspond in a one-to-one ratio with an individual person. An IP address can represent:

? a spider or other agent--not a person at all but an automated browser;
? a cache, a proxy server such as a firewall, or an Internet Service Provider--all of which may represent the use of multiple individuals;
? an individual PC user executing commands on his browser.

?? Qualitative data: Log file data shed no light on the reasons requests were made, user motivations for visiting a site, reactions to site content, actual use made of files served, and all other qualitative aspects of use.

?? Files not viewed: Log files have no record of files in which no activity occurred. Thus, a log analysis report "Least used pages" will not reflect unused pages.

?? Where the user went next: This transaction would be recorded only in the log of the subsequent site visited.

5.2 Data that are logged but inherently incomplete

Number of requests (and all other statistics based on that figure): Server logs provide an incomplete picture of use because of caching. A downloaded page is automatically cached on the client for a period (determined by the amount of memory allocated to this function). Thus, a frequently requested document may be drawn directly from the cache, and the server has no record of its having been viewed. The server records instances only when the cached document is compared with the server version for currency; if, or how often, this occurs depends on browser settings. The clearest example of what is counted is page "re-views" within a browser session: those using Back and Forward buttons or Go features are not counted at the server, while those using the Reload button are counted.

Throughout the Internet, large-scale memory banks, or caches, are increasingly used to reduce response time. This means a file may be cached at various other points in the network en route between the server and the browser, such as a site cache, local regional cache, a service provider's cache, or even a national cache. If the browser finds the file at any intermediary cache, the server has no record of when the file was viewed.

These factors reduce the quantity of use recorded by the server to an unknown extent. Log file totals are, therefore, no more than indicators of the amount of use captured in the logs.

5.3 Unsound inferences from data that is logged

Log files can not support the following inferences, although they are tempting, widespread and, to a greater or lesser degree, encouraged by most of the log analyzer software:

?? That hits are equal to use: "Hits" are all exchanges between the client and the server. In order to present an HTML page to a user, the server serves the HTML file, plus all image files embedded on that page (unless the user has turned images on their browser off). This makes "hits" a highly inflated figure.

?? That "user sessions" can be isolated and counted: "User sessions" are calculated by some log analyzer products by tracking requests received from an IP address until a period of inactivity (say 30 minutes) indicates to the software that the "session" has ended. As this calculation is based on two unsound assumptions--that a host corresponds to an individual, and that the individual would not normally pause (whether to go to another site or another task) within a site visit--user sessions are, at best, gross estimates.

?? That average page views per session, average length of session, average length of a page view, top entry and exit pages, single use pages, and top paths through a site can be calculated: These statistics are derived from the artificial construct of a "user session". Also, because more frequently requested files may be obtained from a cache, the first file logged as requested might, in fact, be in the middle of a user's actual site visit.

?? That all use is the same: An assumption inherent in totaling log file entries into a single usage figure is that all use is the same. Requests made by spiders (automated browsers) are included in the server logs, although these do not constitute a form of use comparable to that of Web browsers (i.e. computers with people behind them). Some log analyzer products can provide reports isolating users that the software recognizes as spiders. Nevertheless, spider use tends to be included in overall use indicator totals automatically.

?? That users' geographic distribution and type of organization can be accurately extrapolated: Log files do not provide a sound basis upon which to categorize the type of user or to track geographical distribution. As noted above, an IP address is a unique number attached to a machine rather than an identifier of people. Secondly, Web log analysis packages tend to base their geographic statistics on where an IP address was registered. But a user's PC may be located in a different geographic location from where its IP address was registered. This is typically the case with Internet Service Providers. For example, individuals from across North America accessing a site through America Online are captured in the log file as being from the state of Virginia. The structure of the Domain Name System causes problems in designating the geographic location and for the organization-type of user because, in effect, the system mixes the two. Geographically, domain suffixes such as .com, .org, and .net could refer to commercial enterprises, organizations and networks from any country. Other suffixes, such as .edu and .gov, when used as top-level domain suffixes, refer specifically to U.S. domains (namely higher education and federal government domains respectively). In terms of Canadian statistics, a major shortcoming for both geographic and organization-type categories is that Canadian companies may use either the geographical .ca suffix, or the .com suffix, but they may not use both. Resolution of this problem would require an amendment to the domain name structure and universal adoption of the revised schema–an unlikely scenario. Therefore, log analyzer reports presenting geographic distributions and organization type breakdowns as separate tables are very misleading.

?? Finally, in most log files, a significant percentage of hits may be unresolved in terms of reverse DNS look-ups (converting IP numbers to domain names, thereby providing the necessary suffix to interpret). These remain numerical addresses of largely unknown origin, although a high level of use from an unresolved IP number may indicate that it is a spider.

6.0 Other Considerations in Using Log File Data

?? Inclusions and exclusions in reports: Most packages allow specific types of files (e.g. image files), directories, IP addresses (e.g. internal users) or any other data string, to be filtered out of the total log. Conversely, multiple server logs or parts of logs can sometimes be combined into a single report. Such exclusions or inclusions must be executed properly and then made clear to those interpreting the statistical reports or comparing one site's use with that of another site.

?? Site mirrors: If a site is mirrored, log files from all sites should be compiled to measure use of the same content at various sites.

?? Size of the site: Page views are log entries for HTML pages only; other file types (such as images, PDF files, text files, and executables) are excluded. But to be used for meaningful comparison among sites or products (i.e. as an overall indicator of "rate of use"), such a figure should be considered in relation to the number of possible HTML pages, i.e. the size of the site.

?? Structure of the site: Intimate knowledge of the structure of a Web site is crucial to produce accurate log analysis reports. In order to analyze only specific directories and/or files of a site, a log file must be accurately parsed or "filtered". In a complex server environment, or for a large and busy site, it is all too easy to produce plausible but inaccurate figures by making errors in the data compiled.

?? Web traffic volatility: Short-term Web traffic is extremely volatile, so that one week's figures may be double, or half, the previous week's (Nielsen). Such fluctuations mean that trends in site traffic emerge only with long-term data analysis.

7.0 Log Analysis Software

Many log analysis packages containing a variety of features are on the market. Some vendors include log analysis as part of an overall Web management software suite that also performs link analysis and performance.

Log analysis tools typically provide the following features:

User-friendly interface

Variety of output formats (HTML, Word, Excel, text, e-mail)

Robust reporting capabilities

Support for a variety of log file formats

Many filtering options

Real-time analysis

Zipped log file processing

Built-in summary database

Remote access to the software

Proxy analysis reporting

Automatic report scheduling

Reverse DNS lookups

A list of software reviews of Web log analysis tools is provided at the end of this paper.

8.0 Conclusion

Currently, log file analysis is perhaps best viewed as an art disguised as a science. The limitations of log file data, Web log analysis software, and the inherent nature of the Web mean that log file statistics should be scrutinized closely and interpreted extremely cautiously. In the future, as the use of caches and agent software within the network increases, the accuracy of log files as use indicators will diminish further. On the other hand, increasing use of cookies and/or new communications protocols and servers may shed more light on users and usage.

For now, it is essential to remember that the true extent of use, and the true number of individual users of the site, remain unknown. However, properly compiled and knowledgeably interpreted, Web server log files can still provide some meaningful statistical indicators of Web site usage.

Selected Readings

Goldberg, Jeff. Why web usage statistics are (worse than) meaningless.

http://www.cranfield.ac.uk/docs/stats/

Linder, Doug. Interpreting WWW Statistics.

gopher.nara.gov:70/0h/what/stats/webanal.html

Neilsen, Jakob. Tracking the Growth of a Site.

http://www.useit.com/alertbox/980222.html

Stehle, Tim. Getting Real About Usage Statistics.

http://www.wprc.com/wpl/stats.html

Turner, Stephen. Readme for analog 3.0: How the web works.

http://www.statslab.cam.ac.uk/~sret1/analog/docs/webworks.html

Web Log Analysis Software Reviews

Randell, Neil. (1998, March 10). The Results Are In. PC Magazine [online].

http://www.zdnet.com/pcmag/features/webanalysis2/index.html.

Randell, Neil. (1998, March 10). Web Site Analysis Tools: The Under-$100 Crowd. PC Magazine [online].

http://www.zdnet.com/pcmag/features/webanalysis2/sb5.html

Randell, Neil. (1997, October 7). Who Goes There? Seven Inexpensive Web Analysis Tools Can Help You Determine Who's Visiting Your Site. PC Magazine [online].

http://www.zdnet.com/products/content/pcmg/1617/prmg0029.html

Taschek, James. (1997, April). Analyzing Your Website. ZD Internet Magazine [online].

http://www5.zdnet.com/products/content/zdim/0204/zdim0012.html

Zieger, Anne. (1997, October 13). Tracking Tools: Your Next Stop. Internet Week [online].

http://techweb.cmp.com/internetwk/trends/1013a.htm