

Library and Archives
CanadaBibliothèque et Archives
Canada

Canada

Français	Contact Us	Help	Search	Canada Site
Home	About Us	What's New	What's On	Publications

Publications

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 /5/3/a3-160-e.html HTTP/1.0" 200 2308
```

```
"http://www.amicus.collectionscanada.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

"www.collectionscanada.ca/women/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>