WWW Server Transaction Logs

David S. Carter
University Library Associate Project
August 18, 1995

Statement of the Problem

Many libraries are creating World-Wide Web documents as information services for their patrons. In order to effectively gauge the effectiveness of these documents, librarians should perform research on their usage, as has been done in the past with other types of information products (i.e. OPACs, CD-ROM databases, etc.). However, the tools available for analyzing data from HTTP (Web) servers (such as *wwwstat*, Fielding, 1994) are at best crude and at worst useless. In the past, librarians have found it helpful to conduct transaction log analyses to examine user behavior with information systems. For my University Library Associate Project, I have written a program, *Clark*, which creates transaction logs from HTTP server access logs, and have written a sample program, *statanal*, which demonstrates some simple analysis which can be done using the transaction logs.

Literature Review

Transaction Logs

Transaction logs have been used by librarians for over a quarter of a century to unobtrusively monitor user behavior with information systems (Peters, Kaske & Kurth, 1993). A transaction log is the output product of transaction monitoring. The transaction monitoring of an information system is defined as "the automatic logging of the type, content, or time of transactions made by a person from a terminal with that system" (Peters, Kurth, Flaherty, Sandore, & Kaske, 1993, p. 38).

To date, most transaction log analysis (TLA) has been done with OPACs and CD-ROM databases (Peters, 1993). However, Peters (1993) points out that "aggregate usage patterns of new types of IR systems, such as Gopher, are useful and enlightening" (p. 46). Peters goes on to say that "studies of reference service providers' use of IR systems still need to be undertaken" (p. 57) and that TLA studies need to move into examining the use of IR systems over the Internet.

Why use TLA? Kaske (1993) writes that "The central goal of [TLA] is to acquire new knowledge, which will help the library managers, systems designers/developers, and researchers better understand how online information systems are used by library patrons and staff" (p. 79). Crawford (1987) identifies the two main purposes of TLA to be performing statistical analysis of system performance and use, and undertaking analysis of searching behavior and problems. Sandore (1993) identifies many ways in which the results of TLA can be applied to improve information systems. These include anticipating the evolution of system use and demands, determining user preference for experimental changes, monitoring the use of help systems, determining instructional needs, and monitoring user searching patterns. Wallace (1993) demonstrated how TLA can identify bibliographic instruction needs and point out weaknesses in information system design. Young (1992) illustrated the use of TLA as a collection management tool.

The unobtrusive nature of TLAs, while in many respects a strength, can also be a weakness. Kurth (1993) states: "Transaction log data effectively describe what searches patrons enter and when they enter them, but they don't reflect, except through inference, who enters the searches, why they enter them, and how satisfied they are with their results" (p. 98). Kurth further goes on to explain that errors in TLA can arise through limitations of the online system, the inability to isolate and characterize individual users, and decisions and biases of the researcher analyzing the logs. To account for some of the shortcomings in TLA, Cochrane & Markey (1983) suggest combining TLA with another type of analysis (either questionnaire or protocol) to provide a more complete picture which can draw on the strengths of both types of studies.

The World-Wide Web

The World-Wide Web (WWW, or simply the Web) is a network hypertext protocol which employs hypertext markup language (HTML) to link documents to each other (Nickerson, 1992). It was developed back in 1989 by Tim Berners-Lee, a research at the Swiss research facility CERN, but the use of the WWW took off in 1993 with the introduction by the National Center for Supercomputing Applications' (NCSA) *Mosaic* program, a graphical WWW client (or browser) which was available for UNIX, Macintosh, and Windows systems and, most importantly, was free (Andreessen & Bina, 1994).

Polly & Cisler (1994) point out two weakness of the use of the Web as an information system: slowness and “chaotic disorganization [sic]” (p. 34). While the issue of speed will have to be taken up by computer scientists and engineers, the disorganization of the Web is a prime target for librarians to tackle. Powell (1994) was one of many to identify the uses to which libraries could put the WWW in the creation of library information systems. To date, hundreds of libraries around the world have created and mounted various documents on the Web, from simple informational ‘handouts’ to Internet resource subject guides to a library (the Internet Public Library) which exists solely on the Web (Goldberg, 1995). It is hoped that applying TLA to WWW systems will go a step or two towards evaluating and improving library WWW information systems.

The Internet community is also starting to recognize the need for Web server TLA (though they may not know the term). Cutler & Hall (1995) point out that business with Web sites “want to know the answers to some relatively simple questions:

- How many people are coming to our site?
- Who are they?
- Where are they coming from?
- How long do they stay on our site?
- How do they move through the site?
- Where are they going when they leave?
- How does our information appear to them?” (p. 22)

All questions which, to one extent or another, can be answered using Web server TLA. Although some commercial products that meet these needs from a business aspect are being developed (e.g. *Intersé Market Focus*; *Analysis Software*, 1995) there is a need for a public domain, customizable, multi-platform, research oriented program, such as *Clark*, the program I have developed for this project.

Archimedes

Before the arrival of the WWW on the big scene, the Engineering Library at the University of Michigan developed a *HyperCard* stack called *Archimedes* which went into public release in March 1991 (Ottaviani, 1995). *Archimedes* used *HyperCard*, a Macintosh hypertext authoring program, to present users with information about the library and the collection. Importantly, the versions of *Archimedes* which resided on dedicated terminals on the library floor kept track of its use by patrons and generated a transaction log and summary statistics. In 1994 the staff of the Engineering Library developed a set of WWW documents which, among other functions, duplicated the content of *Archimedes* in the new Web environment; in 1995 the public *Archimedes* stations were replaced by a dedicated Netscape station on the library floor which defaulted to the library's new WWW home page and information system. It was the desire to be able to gather the same transaction information and statistics as was possible with *Archimedes* in the new WWW system which was one of the main impetuses for this project.

Goals & Methodology

This project had three goals:

1. Devise a program to extract transaction logs from NCSA HTTP server access logs, containing the following information:

- accessing host (and any available user information)
- start and end time of transaction
- test for image loading
- number of requests and length of transaction
- document path and results of requests

The program, *Clark*, allows the optional selection or exclusion of certain hosts from the transaction log (e.g. to exclude staff workstations, etc.), and uses host information combined with a time gap of inactivity (user-definable or default setting) to determine that start and end points of a transaction.

All programming was done in Perl, a widely used, relatively easy, cross-platform, free language (Potter, 1995). The program was designed to use the access logs generated by NCSA's HTTP server, one of the most popular and widely used Web servers; this access log format has been adopted as a common log format for many other HTTP servers. *Clark* has been tested on access logs from the Internet Public Library and the University of Michigan

Engineering Library's WWW sites. The program works well, though is slow and still has a few minor things that need to be ironed out. See the online documentation for *Clark* (http://www.sils.umich.edu/~superman/AP/Clark.html) for more information.

2. Devise one or more programs which use the transaction log to analyze the information contained therein. Ideas include:

- exploratory statistics (raw numbers, averages, distributions)
- entry and exit point analysis
- compare public 'on-floor' usage with general usage

A program called *statanal* performs exploratory statistical analysis for the number of events in each transaction and transaction length, then outputs ordered data sets for future use (in a more robust statistical analysis software package). *statanal* was tested on transaction logs generated by Clark based on data from the Internet Public Library and the University of Michigan Engineering Library. Other analysis programs are forthcoming.

3. Develop documentation so that others can use the programs, write analysis programs of their own, and modify the transaction log generating program as their own research needs require.

The documentation includes this report, manual pages for *Clark* and *statanal*, and comments within the code, all of which can be found at the *Clark* distribution site on the World-Wide Web (http://www.sils.umich.edu/~superman/AP/). The *clark.pl* and *statanal.pl* Perl scripts can also be downloaded from this site.

Limitations

Besides the limitations inherent in transaction logs, the transaction logs I develop will be hindered by two additional limitations inherent in WWW systems:

1. Caching: Most browsers cache text and images, so when a user returns to a previously viewed page, the server is often not accessed; thus, the

server access log contains no record of the return to the cached document and this action cannot be registered in the transaction log. This means that the transaction logs will, by necessity, be incomplete.

2. Links to other servers: It is in the nature of WWW documents that links are often made to other documents which do not reside on the same server. When a user follows such a link, no indication is made to the server, and thus the user's behavior outside of the particular document space cannot be monitored or studied.

In a personal conversation, Kerr (1995) suggested the following solutions to these two problems:

1. Caching: Put an expired tag for a very short period of time in the document header. Many browsers will recognize this tag and thus go back to the server every time the document is accessed, rather than get it from the cache.

2. Links to other servers: Instead of linking to an external server directly, run the request through a CGI script, which will then send the user off to the proper place. Thus the requests to the CGI script will be noted in the log.

It remains to be seen how detrimental of an impact either or both of these factors will have on Web server transaction log analysis. Hopefully this project will help to illuminate the possible effects of these limitations.

Future Work

Clark is in a good, usable state right now, but can certainly use more work. Here's what I'd like to accomplish in the near future:

- Increase speed
- Add in pattern matching capability for domain exclusion and inclusion
- Develop an ‘idiot-proof’ interface, so that Clark won’t crash or do odd things when used improperly.
- Implement an HTML forms interface so that it can be used from directly within a site without hopping over to UNIX command lines.

It is hoped that this project will help instill in others the desire to do quantitative and qualitative research on the use and impact of WWW based information products, via TLA or other types of analysis.

Cited References

Analysis Software. (1995). http://www.interse.com/marketfocus/

Andreessen, M. & Bina, E. (1994). NCSA Mosaic: a global hypermedia system. *Internet Research*, 4(1), 7–17.

Cochrane, P. A. & Markey, K. (1983). Catalog use studies—since the introduction of online interactive catalogs: impact on design for subject access. *Library & Information Science Research*, 5(4), 337–363.

Crawford, W. (1987). *Patron access: issues for online catalogs*. Boston: G. K. Hall.

Cutler, M. & Hall, D. (1995). Sizing 'em up. *Internet World*, 6(8), 22–24.

Fielding, R. (1994). *wwwstat -- distribution information*. http://www.ics.uci.edu/WebSoft/wwwstat/

Goldberg, B. (1995). Virtual patrons flock into the Internet Public Library. *American Libraries*, 26, 387–388.

Kaske, N. K. (1993). Research methodologies and transaction log analysis: issues, questions, and a proposed model. *Library Hi Tech*, 11(2), 79–85.

Kerr, E. (1995). Personal Communication, June 21, 1995.

Kurth, M. (1993). The Limits and limitations of transaction log analysis. *Library Hi Tech*, 11(2), 98–104.

Nickerson, G. (1992). World Wide Web: hypertext from CERN. *Computers in Libraries*, 12(11), 75–77.

Ottaviani, J. S. (1995). Archimedes: analysis of a HyperCard reference tool. *College & Research Libraries*, 56 (2), 171–182.

Peters, T. A. (1993). The History and development of transaction log analysis. *Library Hi Tech*, 11(2), 41–66.

Peters, T. A., Kurth, M., Flaherty, P., Sandore, B., & Kaske, N. A. (1993). An Introduction to the special section on transaction log analysis. *Library Hi Tech*, 11(2), 38–40.

Peters, T. A., Kaske, N. K., & Kurth, M. (1993). Transaction log analysis. *Library Hi Tech Bibliography*, 8, 151–183.

Polly, J. A. & Cisler, S. (1994). What's wrong with Mosaic? *Library Journal*, 119(7), 32–34.

Potter, S. (1995). *comp.lang.perl.* FAQ*. http://www.cis.ohio-state.edu/hypertext/faq/usenet/perl-faq/top.html

Powell, J. (1994). Adventures with the World Wide Web: creating a hypertext library information system. *Database*, 17(2), 59–66.

Sandore, B. (1993). Applying the results of transaction log analysis. *Library Hi Tech*, 11 (2), 87–97.

Wallace, P. M. (1993). How do patrons search the online catalog when no one's looking? Transaction log analysis and implications for bibliographic instruction and system design. *RQ*, 33 (2), 239–252.

Young, I. R. (1992). The Use of a general periodicals bibliographic database transaction log as a serials collection management tool. *Serials Review*, 18(4), 49–60.