AMCIS 1997 Proceedings

Americas Conference on Information Systems
(AMCIS)

8-15-1997

Using Log Files to Assess Web-Enabled Information System Usage

Kristin Eschenfelder
Syracuse University

Steven K. Wyman
Syracuse University

John Carlo Bertot
University of Maryland Baltimore County

William E. Moen
University of North Texas

Follow this and additional works at: http://aisel.aisnet.org/amcis1997

Using Log Files to Assess Web-Enabled Information System Usage

Kristin Eschenfelder and **Steven K. Wyman**,
School of Information Studies, Syracuse University
John Carlo Bertot,
Department of Information Systems, University of Maryland Baltimore County
William E. Moen,
School of Library and Information Science, University of North Texas
Charles R. McClure,
School of Information Studies, Syracuse University

Abstract

This article presents a beginning point for understanding log file assessment techniques. This paper identifies and discusses measurement techniques and issues for the analysis of webserver and CGI interface logfiles, describes the "lessons learned" from two projects which assessed use of a web-enabled tool at a large federal agencies, and postulates about other types of research which might benefit from the employment of log file analysis techniques.

Introduction

Public use of the World Wide Web continues to increase at an explosive rate (Network Wizards, 1997). Multimedia and hypertext formats, along with point and click navigability, make the Web an appealing information resource. Additionally , the low cost of publishing on the Web make the Internet and "intranets" attractive alternatives to traditional paper based information dissemination methods. Furthermore, many software producers are creating web accessible software packages which allow users to employ the package through their web browser, regardless of their desktop platform.

As more organizations come to rely on web-enabled tools, including web pages and web-enabled software applications, several important questions emerge. Who are the users of the web-enabled tools? To what extent do such tools meet the needs of its users? What parts of the tools are most used? How do the users use the tools? With answers to these questions, managers could better evaluate the effectiveness and efficiency of the given tools, and designers could better design new tools to satisfy user needs. The authors see log file analysis as one component of a user-oriented analysis program designed to answer these questions.

This paper identifies and discusses measurement techniques and issues for the analysis of webserver and

CGI interface logfiles, describes the "lessons learned" from two projects which assessed use of a web-enabled tools at large federal agencies, and postulates about other types of research which might benefit from the employment of log file analysis techniques.

Webserver and CGI Logfiles

Webserver and CGI script log file analysis offer a crucial window into work patterns and processes. Typical

webserver software collects data into four different types of logfiles:

Access Log:

This file collects data regarding the following variables: the Internet Protocol (IP) number of the assessors, the date of the access, and user actions taken during the access period (e.g. how often a CGI script is invoked to access an application). An example of an access log line appears below:

smx-ca8-50.ix.netcom.com - - [30/Sep/1996:02:57:07 -0400 "GET/Proj/main.html

From this data, researchers can generate information about the percentage of users coming from a specific domain type (.edu, .com, etc.), a specific domain (company.com), or from a specific IP address. Additionally researchers can ascertain peak usage times, and most and least used files within the tool.

Agent Log:

This file provides data regarding the user's chosen web browser and computer operating system. An example of an agent log appears below:

Mozilla/3.0 (Win95;I)

From this data, researchers can generate information about how users access the tool, what Internet browser they use and what desktop operating system they use.

Error Log:

This file provides data regarding errors a user received as a result of using the tool (e.g. broken link, html page unavailable). An example of an error log appears below:

[Sun Nov 3 23:57:00 1996] httpd: send aborted for pm02-23.ct.net, URL:/OWOW/images/new/owpool.gif

From this data, researchers can generate information about what parts of the tool commonly generate error messages because of misfunctions or dead links. Researchers can also define where in the tool users commonly "give up" and log out because of a problem such as an overly lengthy download time.

Referer Log:

This file provides data regarding what other websites on the Internet link to the given website. An example of a line from a referer log appears below:

http://www.altavist.digital.com/cgibin/query?pg=q&what=web&rmt=.&q=SIC+CODE /xxx/html/rcris/rcr_sic_code.html

With this data, researchers can generate information about what outside sites have created links to the tool. The above example indicates site entry from an AltaVista search output.

Path Analysis:

In addition to the above listed data logs, one of the project teams developed a method to analyze a user's path through, and actions taken while connected to, a website. An shortened example of path analysis output appears below:

/ 07:44:17

/epahome/images/2title1n.gif 07:44:21

/epahome/images/newmenu.map? 436,182 07:44:56

/earth100/ 07:45:01

/cgi-bin/waisgateII?WAISdocID =6998+30+0+0 &WAIS action = retrieve 07:46:45

In the given example, the path analysis shows the user entering the Environmental Protection Agency homepage, using the image map to move to a different file called Earth 100, and then using a CGI script function to perform a document search. Researchers interested in employing path analysis techniques should see <headcase.syr.edu/logs>. The Access Log, Agent Log, Error Log, Refer Log, and Path Analysis all provide useful data which can help to answer the questions set out in the introduction.

CGI scripts are often used as part of the web-enabling process. They generally perform two roles. First they may act as intermediaries between the webserver and applications allowing web access to the application. A CGI script residing on the webserver takes query requests typed into a webpage and passes them on to the application. At the other end, the CGI script takes the output data from the application and transposes them back to the Webserver. Secondly, CGI programmers can write scripts that reside on the applications server to provide reports on applications use. For instance, programmers can write a CGI script to keep track of what user logged onto a database, what query term the user employed, and what documents the user downloaded from the database.

CGI scripts are extremely flexible, and do not have "standard" reports like webserver log files (Dwight and Erwin, 1996).

Issues Confronting Researchers Using Log File Analysis

Researchers using log file analysis should consider several key issues during the project planning period.

- *Understanding what, exactly, the data reflect*. For Webserver logfile analysis, the distinction between ìhitsî (downloads on an html page) and accesses (a downloaded html page) is critical. Software that counts only ìhitsî will not reflect the true nature of the siteís use.

- *Selecting and/or developing appropriate analysis software*. Researchers need to plan for the analysis of Webserver and CGI log files. The types of information desired should drive the selection and/or development process of log analysis software.

- *Obtaining the cooperation of server administrators and IS managers*. It is important to gain the cooperation of those individuals and/or entities that have direct control over the log files. The lack of such cooperation will have a negative impact on the ability to attain usage data.

- *Preserving the privacy and confidentiality of server users*. In some cases, it is possible to trace directly back to a user, depending on the method of access a user has to the webserver. Researchers need to develop policies as to how such data, if at all, will be used. This issue is particularly troublesome for public sector organizational research, as such capabilities may violate privacy laws.

- *Educating IS managers and decision makers as to the benefits of log file analysis*. Log file analysis is only just beginning to gain popularity. IS managers, server administrators and decision makers need to understand the types of data that log files can generate, the application of that data in an information systems assessment strategy, and the incorporation of such data into management activities.

- *Managing the log analysis process*. Gaining access to and analyzing webserver and CGI script log files requires planning, coordination, and accountability. To engage in log file analysis activities, there needs to be a delegation of responsibility for making the files available with size indictors (on-site or remotely), performing the analysis (on-site or remotely), interpreting the analyzed data, and reporting the findings.

These issues serve as a beginning point for understanding log analysis. Other issues exist, and still more will develop as intranet usage increases and log analysis techniques become more sophisticated.

Overview of Federal Agency Tool Assessment Studies

Two research teams, one based out of Syracuse University, the other based out of University of North Texas, employed basic log file analysis techniques in projects assessing federal agency web-enabled tools (Moen and McClure)(Wyman et al.). Both projects used log file analysis as part of a larger user-oriented data collection methodology including on-line data collection forms, situated user assessments, interviews and focus groups. The goal of both projects was to assess to what extent federal web-enabled tools met the needs of their users.

The projects demonstrated the need to begin web-enabled tool evaluation projects with the log file analysis in order to identify high use or problematic parts of the site or tool. Researchers can then develop more effective within-site webpage sampling techniques. Without sampling techniques, researchers are faced with the choice of evaluating the entire site or a selection of questionably representative pages. Using log file analysis, researchers can develop code categories of high or low use, and troublesome pages, and then take random selections from within these defined categories.

The projects also demonstrated the need to plan for different data formats. Researchers should not assume that the site under study uses standard file name formats or even that the site collects the data in the standard four website log files.

Suggestions for Further Research

It is difficult for the researchers to make suggestions about specific research projects which could make use of log file analysis. The purpose of this paper is to explain log file analysis and the types of questions which log file analysis can answer. It is up to the individual researcher to decide if log file analysis can inform their research question. The authors do believe however, that log file analysis would be particularly helpful in the areas of Bibliometrics research which could use the referer log in an electronic form of citation analysis. Additionally, the researchers believe that log file analysis would be useful in the study of user behavior. By employing path analysis, researchers can study and classify types of user movements through a website. Finally, the researchers think that log file analysis can be used to study the adoption rate of web-enabled technologies. By comparing the access log's list of user IP addresses and user actions taken on the site, and assuming that IP addresses are assigned geographically, researchers can ascertain what areas or departments in an organization make the most use of a web-enabled tool.

As public use of the World Wide Web continues to increase and more organizations come to rely on web-enabled tools, researchers need to address several important questions. Who are the users of the web-enabled tool? To what extent do such tools meet the needs of its users? What parts of the tools are most used? How do the users use the tools? Information gained from research employing log file analysis could result in more effective, efficient and more user-oriented web enabled tools. The authors see log file analysis as one component of a user-oriented analysis program designed to achieve these goals.

References

Dwight, Jeffrey and Erwin, Michael *Using CGI* (Indianapolis: Que) 1996.

Moen, William E. and McClure Charles R. *An Evaluation of the Federal Government's Implementation of the Government Information Locator Service (GILS)* (Denton, TX: University of North Texas, School of Library and Information Sciences) Final Report Due Summer 1997.

Network Wizards. *Internet Domain Survey*, January 1997 <http://www.nw/com>.

Wyman, S., C.R. McClure and J.C. Beachboard User and *System-Based Quality Criteria for Evaluating Information Resources and Services Available From Federal Websites: Final Report* (Dublin OH: OCLC Online Computer Library Center, Office of Research) 1997.