Methodologies for Understanding Web Use with Logging in Context

Don Turnbull
University of Texas at Austin
School of Information
1 University Station, D7000
Austin, TX 78712
donturn@ischool.utexas.edu

ABSTRACT

This paper describes possible approaches of data collection and analysis methods that can be used to understand Web use via logging. First, a method devised by Choo, Detlor, & Turnbull (1998, 1999 & 2000) that can be used to offer a comprehensive, empirical foundation for understanding Web logs in context by gaining insight into Web use from three diverse sources: an initial survey questionnaire, usage logs gathered with a custom-developed Web tracking application and follow-up interviews with study participants. Second, a method of validating different types of Web use logs is proposed that involves client browser trace logs, intranet server and firewall or proxy logs. Third and finally, a system is proposed to collected and analyze Web use via proxy logs that classify Web pages by content.

Triangulation: browser history, firewall logs and intranet server logs.

Keywords

Web Use; Survey; Questionnaire; Client Application; WebTracker; Interview; Methodology; Logs; Server Logs; Proxy; Firewall; Analytics; Content Classification; Client Trace; Transaction Log Analysis

1. INTRODUCTION

Much research into understanding Web use is collected in contrived, non-native settings such as a usability lab or via broad surveys. To progress beyond simple, descriptive understanding more contextual, empirical methods must be used. Initial studies of real-world Web use, most notably the Georgia Tech Web surveys begun in 1994 (Pitkow & Recker, 1994) cover a wide range of Web users, mostly students and home users (Kehoe, Pitkow & Rogers, 1998). Other studies of Web users, such as Jansen, et. al (2000) expand on this by examining search behaviors of relatively anonymous Web search engine users by relying on more empirical activity logs. However, collecting log data is not enough. To achieve truly insightful view of Web use, it is proposed that a combination of contextual methods be used in concert with log collection to focus in on specific types of users in specific settings. The methods proposed in this paper aim at smaller, more homogeneous sets of Web users such as knowledge workers in their native workplace or library patrons using public computers to hopefully gain richer, more complex views of overall Web use behavior by relying on a set of complementary instruments including log files.

2. A CONTEXTUAL METHODOLOGY

In the first sample study presented in this paper, several organizations were asked to contribute several Web users to the study. Once permission was obtained to study an organization, an initial briefing for all participants was conducted at each organizational work site. At these briefings, participants were told of the purpose of the study, their personal involvement, the confidential nature of the study; and how the custom-developed WebTracker application would be installed on each participant's machine to monitor their Web usage activity. At each briefing session, participants were given the opportunity to ask any questions they may have. At the conclusion of each session, a **survey questionnaire** was distributed for the participants to complete. During this time, individual appointments were scheduled to install the **WebTracker software** on each participant's machine. During the installation of the software, participants were given a walk-through of how the software works and shown how to view the log files that recorded their personal Web use. Note that each participant was shown how to disable this monitoring, if the participant so chose. After an agreed-upon monitoring period, use logs were collected and analyzed off-site. Later, a **follow-up interview** was conducted with each participant to discuss significant episodes of Web activity identified in individual tracking logs.

2.1 SURVEY QUESTIONNAIRE

All participants completed the survey questionnaire portion of the study. The survey questionnaire instrument was adapted from a questionnaire devised by Auster & Choo (1993) in a study on the environmental scanning behavior of Chief Executive Officers in two Canadian industries. We felt this questionnaire was applicable to measure the behaviors of knowledge workers using the Web in an organizational work environment.

The survey questionnaire instrument was composed of two broad sections. The first dealt with the perception and use of information sources by participants. The idea was to capture and measure participant perception of the World Wide Web compared to other information sources used in typical work activity. In this section, participants were asked to rate their frequency of usage of each of the twelve sources, and to give their perceptions of each source in terms of its quality and accessibility.

2.2 WEBTRACKER

The second instrument used was WebTracker, an application for gathering Web browsing metrics developed for the Faculty of Information Studies at the University of Toronto. WebTracker was designed because of the potential inaccuracy of using Proxy or Firewall servers (Pitkow, 1997) to study fine-grained Web browser activity and the lack of (then) current, publicly-available browser code for the Windows environment to instrument a browser. Previous studies used XMosaic (Catledge & Pitkow, 1995 and Cuhna, Bestavros, & Corvella, 1995) on UNIX systems, but as this sample study focused on corporate users who predominantly work on Microsoft Windows platforms, we required a different tool. Despite the presence of newer, Windows-specific Web browser source code from the Mozilla project we felt that installing a new, instrumented browser would not allow us to observe the actual behavior of users participating in the study. Even requiring participants to use or modify newer browsers such as Firefox or Opera was seen as interrupting their normal Web behavior. With a supplementary tracking and logging application, participants can simply work on the Web as they did before, with their usual system configurations and browser preferences including bookmarks and toolbar choices.

WebTracker runs on 32-bit Windows environments and has standard Windows controls and behaviors including normal application install procedures, standard (non-hidden) systems processes, and can therefore be used, understood, observed and uninstalled easily by participants and has no appearance of surreptitious malware or spyware.

Primarily, WebTracker watches the Web browser and logs menu choices, button bar selections, and keystroke actions. These actions are associated with the open Web page (URL), tagged with a date-time stamp and recorded in a daily log file. This tracking method enables log analysis that can essentially reconstruct move-by-move how participants looked for information on the Web. Currently, other similar tools are available for logging user data in browsers including The Wrapper (Jansen 2005).

During the initial stages of the study, we physically visited the users' individual work environments and installed WebTracker to run at system startup as a minimized application. (By developing WebTracker as a standalone, typical Windows application, participants could “see” it running, and have WebTracker available for suspending or viewing their usage logs.) After verifying that WebTracker was functional, we again explained how WebTracker works by showing the few user functions available. These included the option of turning WebTracker logging off by selecting the "Web Tracker is INACTIVE" radio button.

Next, we showed each participant how to enter a personal identifier string used thereafter in the WebTracker log file. This is the only actual interaction with WebTracker required by the participant. Once configured to load at system startup as minimized, WebTracker runs without any additional intervention for the duration of the study. Once WebTracker had been demonstrated, participants were encouraged to use their Web browsers as they normally would.

2.3 INTERVIEWS

At the conclusion of the study and after the logs were analyzed, one-on-one interviews were conducted with participants. These interviews served two broad purposes. The first was to better understand the context behind individual Web usage activity recorded in the tracking logs. The interview format was based on the principles of the Critical Incident Technique (Flanagan 1954), in which the ‘incident’ to be studied should be recent, sufficiently complete, and its effects or consequences suitably clear. In the interviews, participants described two ‘critical incidents’ of Web information seeking and use. Where appropriate, participants were prompted with the names of Web sites that were indicated in their WebTracker log files. Additional questions were asked about behavioral regularities noted in the log files as well as isolated, unique log entries to gain further understanding of typical and atypical Web use, which often lead to additional discussion.

The second broad purpose of the interviews was to obtain participant perception of the Web in general. To do this, participants were invited to comment more broadly on their use of the Web, including their general Web use strategies and preferences, as well as what they perceived to be both the positive and negative aspects of Web use. These interviews provided insight into the context behind each individual participant's Web use logs within their organizational settings.

3. A LOG VALIDATION METHODOLOGY

In this second illustrative study, a very large corporate organization’s Web use was studied with the goal of understanding individual and aggregate patterns of Web use including overall behavior. The participating organization allowed for tuning their existing technology infrastructure for data collection including **outgoing Web use logs** and access to **internal Web server logs**. The overall questions addressed in the study fit with this method of collecting and understanding Web use from multiple perspectives, both to get a wider view of Web use, but also to validate the possibility of collecting truly accurate data with such methods. Additionally, the use of the firewall, or more appropriately a proxy server with automated content classification of requested Web resources could yield insights into topics of interest and insure that logged content is relevant.

3.1 FIREWALL LOGS

All outgoing broadband Web access in the organization was controlled and permitted through a set of firewall applications. The existing firewalls were customized to log all outgoing access including both requested and resolved pages, blocked requests and protocols and ports beyond HTTP conventions as well as to rely on static IP addressing or other identifiers to provide large scale comprehensive and consistent views of Web use by well over 1000 knowledge workers. For the most part, collecting this scale of general Web use would not likely be applicable with a client-based logging tool that could have difficulties in installation on large numbers of individual systems, advocating its continued use by users or in possible data collection practices.

3.2 INTRANET ACCESS LOGS

It is often thought that in some configurations, client browsing application local caching settings may influence server-based logging accuracy. If it is not efficient to modify each study participant's browser settings (or that temporarily modifying participants browser settings for the study period affects true Web use) a method of factoring in what may be lost due to local cache may be applied. In this study, local browser cache settings were not modified but logs from the organization's own intranet Web servers were collected. By tuning intranet server logging settings and collecting and analyzing these logs, some initial measurement of the differences that client browser caching makes in accurate firewall logs can be made. Comparisons to access on the organizations intranet Web server logs such as total page requests per page, time to load, use of REST or AJAX interaction and consistent user identification can be made to the more raw logging from the firewall logs collected. Of course, this method does not guarantee universal discrepancies but can provide insight into a study's own user base of particular browser applications and their settings on the overall comprehensiveness of log data collected via the firewall logs.

3.3 PROXY LOGS FOR TOPIC IDENTIFICATION

This second study did not immediately or directly attempt to classify study participant Web content requests, but it is suggested that logging and classifying Web content as it is being requested is essential for a large-scale accurate view of Web use. In this study, the firewall logs were used to request and collect Web page content and other requested resources after the study period was complete. This made full collection of logged requested content difficult as the dynamic nature of page generation, changes in pages over time and of course the continued availability of Web content is problematic. Therefore, it is suggested that a proxy server be used to log and save all requested content at request time and be used as a data set for classifying the content in a taxonomy of topics or keyword identifiers as appropriate. A study that can separate non-user requested content such as advertising content, redirected Web pages, pop-ups and even junk email can yield a much more detailed view of what Web users activity involves. The use of the proxy server, in this proposed case, a modification of the SQUID server is also useful (Rousskov 1999) as it allows for removing the load from a firewall server, which could also be used for this purpose. The collected proxy logs and content data can also be archived and processed offline and for future comparison to additional studies. Moreover, in some cases the proxy itself could provide the main data logs for the study depending on the existing study site's technology environment and specific study goals such as focusing on studying Web use in a laboratory or library setting that does not seek to use a firewall to restrict any Web accesses.

4. SUMMARY

This paper outlines two methods of empirical investigation for studying Web use in organizational contexts. The combination of survey questionnaire, client Web logging software and personal interviews provided complementary methods of collecting contextual qualitative and quantitative data. The second, validated logging method including firewall, intranet server and the potential for requested content identification via proxy logs yields a more scalable, multi-faceted and more verifiable context than single source log collection perspectives may provide. Of course, combinations of these two methods may be used depending on both technical and organizational constraints of the study site. Finally, the depth of context required for a study's hypotheses such as usability testing for Web pages and/or Web browser applications; evaluation of how Web-based networks such as corporate intranets are utilized; or user profiling studies by examining the behavior of Web users during both browsing and/or searching activity may also guide researchers into particular blends of logging systems and contextual data collection as reviewed.

5. REFERENCES

Auster, E., & Choo, C. W. (1993). Environmental scanning by CEOs in two Canadian industries. *Journal of the American Society for Information Science, 44*(4), 194-203.

Catledge, L. D., & Pitkow, J. E. (1995). Characterizing Browsing Strategies in the World-Wide Web. *Computer Networks and ISDN Systems, 27*, 1065-1073.

Choo, C.W., Detlor, B. & Turnbull, D. (1998). A Behavioral Model of Information Seeking on the Web — Preliminary Results of a Study of How Managers and IT Specialists Use the Web. *Proceedings of the 61st Annual Meeting of the American Society of Information Science,* 290-302.

Choo, C.W., Detlor, B. & Turnbull, D. (1999). Information Seeking on the Web - An Integrated Model of Browsing and Searching. *Proceedings of the 62nd Annual Meeting of the American Society of Information Science*, Washington, D.C.

Choo, C.W., Detlor, B. & Turnbull, D. (2000). *Web Work: Information Seeking and Knowledge Work on the World Wide Web*. Dordrecht, The Netherlands, Kluwer Academic Publishers.

Cuhna, C.R., Bestavros, A. & Crovella, M.E. (1995). Characteristics of WWW Client-Based Traces. Technical Report #1995-010. Boston University, Boston MA.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin 51(4)*, 327-358.

Jansen, B. J., Spink, A. & Saracevic, T. (2000) Real life, real users, and real needs: a study and analysis of user queries on the Web. *Information Processing & Management*, Volume 36, Issue 2, pp 207-227.

Jansen, B. J. (2005) Evaluating Success in Search Systems. *Proceedings of the 66th Annual Meeting of the American*

Society for Information Science & Technology. Charlotte, North Carolina. 28 October – 2 November.

Kehoe, C., Pitkow, J. & Rogers, J. (1998). GVU's Ninth WWW User Survey Report. http://www.gvu.gatech.edu/user_surveys/survey-1998-04.

Pitkow, J. and Recker, M. (1994). Results from the first World-Wide Web survey. *Special issue of Journal of Computer Networks and ISDN systems*, *27*, 2.

Pitkow, J. (1997, April 7-11). *In Search of Reliable Usage Data on the WWW.* Sixth International World Wide Web Conference Proceedings, Santa Clara, CA.

Rousskov, A. & Soloviev, V. (1999) A performance study of the Squid proxy on HTTP/1.0. *World Wide Web*., 2, 1-2, pp 47 – 67.