Lamprey: Tracking Users on the World Wide Web

Ramon M. Felciano & Russ B. Altman
Section on Medical Informatics • Stanford University
Stanford, California

Tracking individual web sessions provides valuable information about user behavior. This information can be used for general purpose evaluation of web-based user interfaces to biomedical information systems. To this end, we have developed Lamprey, a tool for doing quantitative and qualitative analysis of Web-based user interfaces. Lamprey can be used from any conforming browser, and does not require modification of server or client software. By rerouting WWW navigation through a centralized filter, Lamprey collects the sequence and timing of hyperlinks used by individual users to move through the web. Instead of providing marginal statistics, it retains the full information required to recreate a user session. We have built Lamprey as a standard Common Gateway Interface (CGI) that works with all standard WWW browsers and servers. In this paper, we describe Lamprey and provide a short demonstration of this approach for evaluating web usage patterns.

MOTIVATION

The potential value of the World Wide Web (WWW) to health care has been widely recognized. Academic and industrial institutions are making information available on-line and developing WWW front ends to existing software. For example, researchers are designing WWW-based access to oncology resources(1), clinical patient data(2), clinical practice guidelines(3), medical education software(4), and the Medline literature database(5, 6). Despite the popularity of using the WWW as a common interface to healthcare resources and information, there are few mechanisms available for evaluating these web-based interfaces. It is particularly difficult to evaluate the ways in which users reach various pages, the time they spend on individual pages, and the sequence in which they find required information. Clearly, these questions are crucial for medical applications that are targeted to busy healthcare providers. In fact, any institution with an investment in creating web pages would like to know the ways in which these are used, and how they can be improved.

Current WWW servers provide access logs that track how often individual pages are requested by users. These logs help to answer questions that focus on the use of a particular page (e.g., "How many people visited my page today?" or "What is the most often visited page on our site?"). More complex questions about how users navigate to, from, and through a WWW site are difficult, if not impossible, to answer because the standard logging mechanisms track pages, not users (although the IP addresses of clients are available). Since the logging programs are part of the server software, the view obtained from these programs reflects only the input/output experience of the server. Clients, however, are free to navigate in much more complex routes that involve multiple servers. Information about these routes are not easily derived from the server logs.

We have developed Lamprey, a standardized engine for tracking users as they navigate through the World Wide Web. Lamprey's key benefits are that (1) it focuses on the links used to navigate the web rather than individual web pages (thus focusing on the user rather than the site), (2) its tracking behavior is transparent to users, and (3) it tracks usage across the entire Web using any client and any server.

SYSTEM ARCHITECTURE

Lamprey tracks users by rerouting all of their web navigation through a central tracking gateway. The routing is done quickly and transparently to users, with no change in the appearance of the Web pages and little performance degradation (typically under one second delay). The gateway is implemented as Common Gateway Interface (CGI)(7) script written in Perl.

The central mechanism to Lamprey's tracking system is the parsing of HTML pages and embedding of tracking information in every hypertext link in the page. We call this process the "Lamprefication" of Uniform Resource Locators (URLs (7)). A URL represents the address of a WWW page, and typically includes the address of the WWW server as well as the location of the desired page on that server. When a user being tracked by Lamprey requests a page, our system fetches it and changes every URL in that page to reroute it through Lamprey. Thus, instead of linking directly to the original Web site, the link now points to Lamprey. A "Lamprefied" URL includes all the parameters necessary for Lamprey to fetch the original page and return it to the user.

0195—4210/96/$5.00

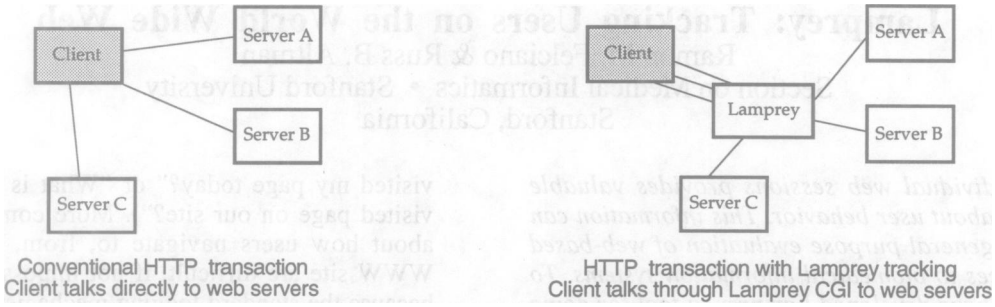

Fig. 1: Lamprey system architecture compared with conventional web interactions.

Original URL:

```
<A HREF="http://site.host.edu/
    projects/index.html">Our Project</A>
```

URL After Lamprefication:

```
<A HREF="http://www.lamprey.stanford.edu/
    trackercgi?url=http://site.host.edu/
    projects/index.html&user=smith@stanford.e
    du&dts=09:10:15">Our Project</A>
```

Once the user sends a URL to the Lamprey application, all subsequent activity is logged (until the browser window is closed). In our current use of Lamprey, we track user information by asking them to log in as part of our experiment, although this is not a technical requirement. The anchor text that the user sees is unchanged: for both URL anchors above, the user sees "Our Project" as a hyperlink to click on. Only the underlying destination of the link is changed to send all URL traffic through Lamprey.

This process currently only tracks hyperlink and forms submitted across HTTP. Lamprey's parser recognizes a subset of HTML 3.0(7) and assumes rigid adherence to this specification. URLs that use FTP, Gopher, S-HTTP and other protocols are left intact and not tracked. Lamprey does track forms and search requests, but does not track the data entered into password fields for restricted access pages.

Using this method of altering the HTML source, Lamprey tracks a number of statistics:

- The id of the user being tracked (typically an e-mail address obtained through a log-in page).
- The page they are coming from, including when they arrived at that page.
- The link that was activated (i.e. the page the user is going to), and when that link was activated.
- Additional client information such as the WWW browser they are using, and their IP address.

We have developed three methods of viewing Lamprey log files. All three methods are available through a WWW-based interface and can be viewed from remote locations.

Tabular report (Figure 2)
A formatted tabular listing of log entries is useful for examining the raw data at its finest granularity. A Lamprey tabular report includes a simplified version of the basic time-stamp data in the log file.

Footprints (Figure 3)
The "footprints" report shows a formatted history of the user's actions. The reporting mechanism performs abstractions on the log data to simplify the format. For example, although Lamprey does not directly track the usage of client-side interactions such as clicking on the "Back" button (these actions are not transmitted over the network and so can not be captured by this

Time	From	To
13:48:25	manually entered	http://www-camis.stanford.edu/
13:49:00	http://www-camis.stanford.edu/ (13:48:25)	http://www-camis.stanford.edu/smi/projects/
13:50:07	http://www-camis.stanford.edu/smi/projects/ (13:49:00)	http://www-camis.stanford.edu/smi/projects/history.html
13:50:39	http://www-camis.stanford.edu/smi/projects/history.html (13:50:07)	reload
13:51:03	http://www-camis.stanford.edu/smi/projects/ (13:49:00)	http://camis.stanford.edu/projects/shine.html
13:51:11	http://camis.stanford.edu/projects/shine.html (13:51:03)	http://camis.stanford.edu/projects/shine/shine.html
13:51:28	http://www-camis.stanford.edu/smi/projects/ (13:49:00)	http://camis.stanford.edu/projects/intermed-web/
13:52:09	http://www-camis.stanford.edu/smi/projects/ (13:49:00)	http://camis.stanford.edu/people/ehs/
13:52:44	http://camis.stanford.edu/people/ehs/ (13:52:09)	http://www-ksl.stanford.edu/abstracts_by_author/Shortliffe_E.papers.html
13:53:02	http://www-camis.stanford.edu/smi/projects/ (13:49:00)	http://www-camis.stanford.edu/people/bio/shahar.html

Fig 2. Lamprey tabular log

1. 14:13:08: --Started from http://www-camis.stanford.edu/
2. 14:13:18: ----jumped to http://camis.stanford.edu/people/
3. 14:13:32: ------jumped to http://camis.Stanford.EDU/people/kxl/KXL.html
4. ----------: -------- back to http://camis.stanford.edu/people/
5. 14:13:47: ----jumped to http://www-camis.stanford.edu/people/lrw/Index.html
6. 14:14:04: ------jumped to http://www-camis.stanford.edu/people/lrw/Hobbies.html
7. ----------: -------- back to http://camis.stanford.edu/people/
8. 14:14:57: jumped to http://www-camis.stanford.edu/people/bio/vian.html
9. ----------: -- back to http://camis.stanford.edu/people/
10. 14:15:15: jumped to http://cmgm.stanford.edu/~brutlag/
11. 14:15:58: jumped to http://cmgm.stanford.edu/~brutlag/Experience.html
12. ----------: -- back to http://cmgm.stanford.edu/~brutlag/
13. 14:16:10: jumped to http://cmgm.stanford.edu/~brutlag/Research.html
14. ----------: -- back to http://camis.stanford.edu/people/
15. 14:18:29: jumped to http://www-camis.stanford.edu/people/bio/rcarlson.html
16. ----------: -- back to http://camis.stanford.edu/people/
17. 14:18:40: jumped to ftp://db.stanford.edu/www/people/gio.html
18. ----------: -- back to http://camis.stanford.edu/people/

Fig 3. Lamprey "footprints" log

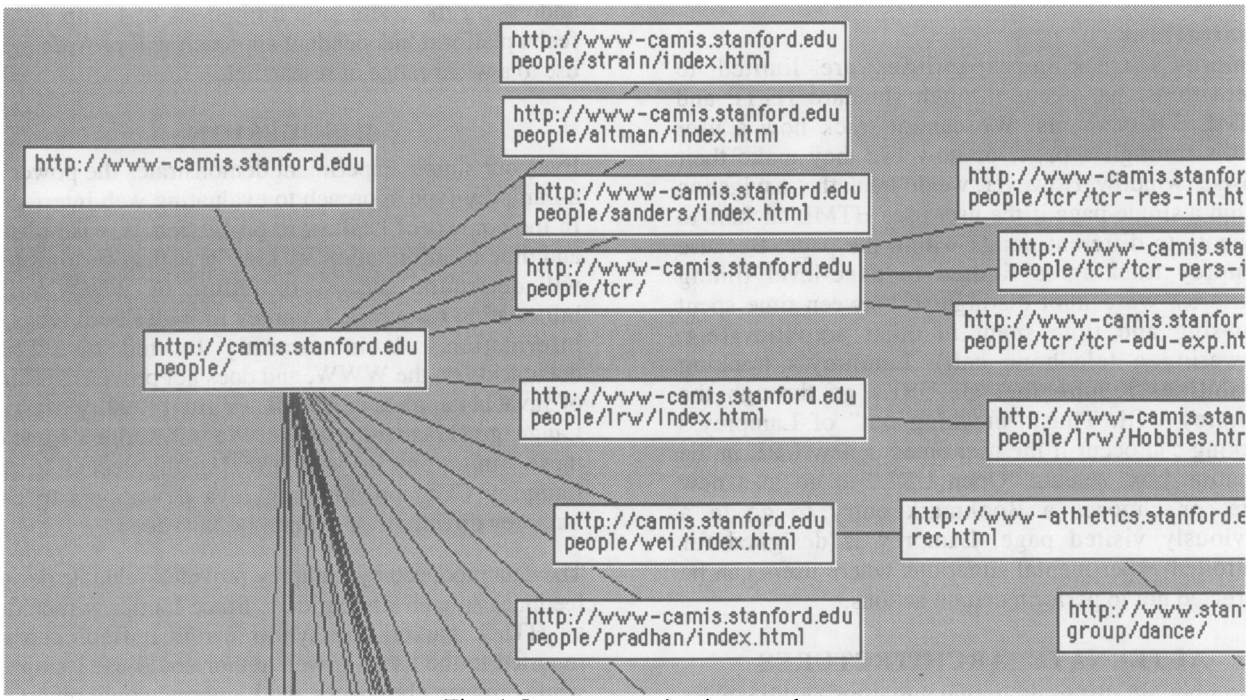

Fig. 4. Lamprey navigation graph

mechanism), we can often infer these actions. Using time stamps, we can distinguish between a user returning to a recent page using the "Back" button, and clicking on a new link that returns to the page.

Graphing user's view of the web (Figure 4)
The third way of viewing the log is to display the user's actions as a graph. Each node on the graph represents a web page, and the vertices represent links between the pages. The vertices are weighted proportionately to how often a link was used, so heavily traveled paths appear as thicker lines. The graph is drawn dynamically using a Java applet. Although this method loses information about the sequence of pages, it summarizes effectively the "web space" that has been traversed by the user.

SAMPLE USE

We performed a pilot demonstration of Lamprey by asking seven users to browse the web site for the Center of Advanced Medical Informatics at Stanford (CAMIS) to find and vote on their favorite home page, among the 44 affiliated faculty, staff and students. They were asked to start at the CAMIS WWW site (http://www-camis.stanford.edu/). The data from the seven sessions were displayed using the methods described above. Each individual trace was quite different (and, in fact, can be simulated faithfully using

the information collected by Lamprey). We found that the average time from start of the Lamprey session to completion was 15.4 minutes (range: 6.2-24.5). The average time taken to reach the home page that the user eventually selected as the best was 6.7 minutes (range: 2.8 - 17.2). The average number of home pages visited by the users before finishing the session was 18 (range: 4-29). We also observed that the users averaged 1.9 (range:1.34-3.38) URL traversals per non-URL traversal (using the backward, forward or go facility in browser). In addition to these quantitative statistics, we were able to make qualitative observations about the nature of these sessions. Some users were noted to do a breadth-first examination of pages, while others examined home pages in a depth-first manner. The ability to gather these data easily suggested to us that Lamprey would be suitable for comparative evaluation of different web site organizational schemes.

Limitations

Lamprey's tracking capabilities are limited to interactions that occur through standard HTTP and HTML 3.0 elements. We cannot track how a user scrolls through a page, or how big they make their browser window. However, we do track that navigation within a single page if the user uses HTML anchors to navigate to different points within the page. Because Lamprey tracking is limited to page-level timing measures, we cannot distinguish between time spent browsing and time spent in other activities (e.g. answering a telephone call). Lamprey's tracking capabilities require that all URLs go through the Lamprey CGI. Thus, "breaking out" of Lamprey's tracking can occur if the user enters a new URL in the Location box, selects "Open URL" to go to a new page, or chooses a Bookmark entry to go to a previously visited page. Lamprey is designed for controlled experimental situations where users can be instructed not to perform certain actions.

ALTERNATE ARCHITECTURES

A number of different implementation architectures could support a Lamprey-like system.

Server extension

One approach is to install Lamprey at the level of the HTTP server, and track all user navigation within the web site served. Netscape Communications' servers are based on extensible architecture that allow developers to add "wrapper" functions around the server. The SiteTrack system from Group Cortex, Inc.(8) implements similar functionality to allow tracking of pages on a particular site. Lamprey differs in that it tracks all web traffic across all sites.

Proxy server

A proxy server "provides access to the Web for people on closed subnets who can only access the Internet through a firewall machine"(9). Because firewalls block access to sites outside of the firewall, proxy servers function as routers to fetch requested web documents and send them back to users. Usage of a proxy server could be explicitly turned on by user (typically as a preference to their client software). Lamprefication of documents could occur transparently at this phase. Once turned on, the Proxy architecture has the advantage that users cannot accidentally "break out" of the tracking session (see "Limitations").

Client-side scripting

A third approach to providing Lamprey-like tracking is through client-side scripting architectures such as Netscape Communications' JavaScript(10). JavaScript allows users to embed scripts in their web pages that are executed once the page is downloaded by a web client. While this approach may hold some promise for addressing the client-side limitations of Lamprey, we feel a platform independent approach will provide more use to a wider range of researchers.

DISCUSSION

Even our simple experiment demonstrates the power of a Lamprey-type approach to evaluating web interfaces. In the context of healthcare professionals, ease-of-use and time-to-information are clearly critical features that will determine success or failure of WWW-based information resources. Lamprey provides both types of information. It is currently limited to HTML documents on the WWW, and does not provide tracking of Java(11) applets or VRML (Virtual Reality Markup Language)(12) manipulations. We will address some of these limitations by standardizing the access to the Lamprey CGI and allowing Java developers to call Lamprey directly from within the Java code.

The data produced by Lamprey provides valuable design feedback to web site builders. Since Lamprey includes an HTML parser, any syntax errors it finds can be reported to the user—a web author could use Lamprey while browsing her pages and Lamprey would report on any bad HTML it found. During our evaluation, several pages on the CAMIS WWW site were found to have malformed URL expressions. Further, because Lamprey tracks the actual browsing steps a user takes, including page re-visits, web site builders can see which pages users return to regularly as they navigate a site, and which ones are rarely visited. The order and frequency in which pages are visited provides valuable qualitative information to site designers about the efficiency of the site organization.

We are using Lamprey to quantitatively evaluate WWW-based interfaces. We provide users with alternate interfaces using standard Web clients, and measure how long it takes a user to accomplish certain tasks. We can compare performance in terms of actual time spent and

in terms of the number of links traversed. Qualitative impressions from users can also be collected.

The data produced by Lamprey can be used to perform keyword indexing on the pages that a user visits in order to generate a simple model of their interests. For example, the top ten keywords in the pages a user visits could be fed into Web search engines to do daily checks for similar pages of interest. Keywords could be identified using any number of existing information retrieval techniques (13, 14).

PRIVACY AND SECURITY

A tracking technology such as Lamprey gives rise to a number of sensitive privacy and security issues. For example, since all forms submissions go through Lamprey, there exists the possibility of intercepting and recording information included in these forms. While secure protocols such as S-HTTP(10) and SSL(15) can be used to keep this from happening, most web-based forms are still submitted across standard HTTP and are subject to this tracking. It is also possible to hide the fact that Lamprey is active through extensions to commonly used browsers.

The potential for misuse of such tracking technology needs to be addressed explicitly by the Internet community. In our implementation, we announce Lamprey's tracking activity on every page the user sees. A banner inserted across the top of the page reminds the user that activity is being tracked and gives the user an option to stop the tracking mechanism at any time.

Acknowledgments

The authors thank Tom Rindfleisch for guidance and support. Torsten Heycke and Christopher Lane of the CAMIS SSRG provided valuable assistance in implementing Lamprey. This work was conducted with the support of the National Library of Medicine under grant LM-07033. Russ B. Altman is a Culpeper medical scholar, and is supported by LM-05652. Computing facilities were provided by the CAMIS Resource, LM-05305. Please address correspondance to Ramon M. Felciano, Section on Medical Informatics, MSOB 215, Stanford, CA 94305. The authors can be reached via electronic mail at {felciano,altman}@smi.stanford.edu.

References

1. Buhle E Jr., Goldwein JW, Benjamin I. OncoLink: a multimedia oncology information resource on the Internet. Proceedings the Annual Symposium on Computer Applications in Medical Care 1994:103-7.
2. Cimino JJ, Socratous SA, Clayton PD. Internet as clinical information system: application development using the World Wide Web [see comments]. Journal of the American Medical Informatics Association 1995;2(5):273-84.
3. Liem EB, Obeid JS, Shareck EP, Sato L, Greenes RA. Representation of clinical practice guidelines through an interactive World-Wide-Web interface. Proceedings the Annual Symposium on Computer Applications in Medical Care 1995:223-7.
4. Kruper JA, Lavenant MG, Maskay MH, Jones TM. Building Internet accessible medical education software using the World Wide Web. Proceedings the Annual Symposium on Computer Applications in Medical Care 1994:32-6.
5. U.S. National Library of Medicine. Internet Grateful Med®. http://www.nlm.nih.gov/publications/factsheets/internet_grateful_med.html 1995.
6. Detmer WM, Shortliffe EH. A model of clinical query management that supports integration of biomedical information over the World Wide Web. Proceedings the Annual Symposium on Computer Applications in Medical Care 1995:898-902.
7. World Wide Web Consortium. Web Specifications and Development Areas. http://www.w3.org/pub/WWW/#Specifications 1995.
8. Group Cortex. Site Track System. http://www.cortex.net 1995.
9. Luotonen A, Altis K. World Wide Web Proxies. http://www.w3.org/hypertext/WWW/Proxies/ 1994.
10. Netscape Communications Inc. The Internet Application Framework: A White Paper. http://home.netscape.com/comprod/server_central/tech_docs/oif.html 1996.
11. Sun Microsystems Inc. The Java Language Environment: A White Paper. http://java.sun.com/whitePaper/java-whitepaper-1.html 1995.
12. Enterprise Integration Technologies. VRML 1.0 Draft Specification. http://www.eit.com/vrml/ 1995.
13. Salton G. Automatic Text Processing. Reading, Massachusetts: Addison-Wesley Publishing Company, 1988.
14. Fowler J, Maram S, Kouramajian V, Devadhar V. Automated MeSH indexing of the World-Wide Web. Proceedings the Annual Symposium on Computer Applications in Medical Care 1995:893-7.
15. Enterprise Integration Technologies. Secure HTTP. http://www.eit.com:80/creations/s-http/ 1994.