Exploring User Navigation during Online Health Information Seeking

Laurel Graham, Tony Tse, Alla Keselman
Lister Hill National Center for Biomedical Communications,
National Library of Medicine, NIH, DHHS, Bethesda, MD

ABSTRACT

Understanding online user behavior is essential for designing user-friendly consumer health Web sites. Transaction log analysis (TLA) provides a way to extract aggregate data about online behavior. This paper describes prevalent user navigation trends using TLA methods at ClinicalTrials.gov. Preliminary results suggest that users typically access low-level pages directly from Web-based search engines and consumer health sites/portals. A pilot user study is presented to illustrate a complementary research method that might be integrated with TLA to provide a multidimensional view of online health information-seeking behavior. Implications of the observed navigation behavior on the design of consumer health Web sites from TLA and users studies are discussed.

INTRODUCTION

Nearly two-thirds of all online adults have reported seeking health information [1], but relatively little is known about these consumer behavior patterns. Further, considerable variation has been observed in consumer information needs, ability to access information, and search strategies [2]. Thus, tools for analyzing online behavior trends support (1) improved understanding of typical search and navigation patterns and (2) user-oriented (re-)design to improve search outcomes and user satisfaction.

Transaction log analysis (TLA) is a non-intrusive method for investigating online behavior using data extracted from log files. It has been used to study online search behavior [3], query failures [4], navigation [5,6], and browsing strategies [7]. Navigation research focuses on user actions at a Web site during an information-seeking episode, such as clicking on links, initiating queries, and other task-oriented actions (e.g., logging into a system).

Navigation research provides information about how users traverse Web sites and use site-based tools (e.g., hierarchical trees and search). Chen and Cimino [5] applied pattern discovery to examine physician behavior on a clinical information system for guiding Web site design. Jansen and McNeese [6] evaluated the impact of automated assistance on user search performance by analyzing user-system interaction patterns. Catledge and Pitkow [7] broadly classified user navigation by path length and frequency, noting that users tend to exit a site by backtracking through the entry path via the *Back* button, creating a spoke-and-hub traversal structure, where no more than two layers were traversed before returning to the hub.

TLA, however, relies on retrospective, archival data (i.e., "footprints") that do not address *how* or *why* users visited Web sites (e.g., motivations), the quality of their online experience, and other behaviors (e.g., multitasking, non-browser interactions) [8]. User studies [9,10] provide rich information on user experience and goals. Thus, multi-dimensional studies using both TLA and studies involving users show a more complete picture than TLA alone.

This study applies TLA to investigate online user navigation behavior patterns at the National Library of Medicine (NLM) consumer health Web site, ClinicalTrials.gov. Following a description of the methods, we present navigation trends. The preliminary results offer insights into online user behavior at the site, suggesting ways to improve access. Finally, we describe the results of a pilot user study on seeking clinical trials and discuss ways in which it might enhance TLA by creating a more complete picture of online health information seeking behavior.

METHODOLOGY

We obtained ClinicalTrials.gov log data over a three-month period (July-September 2005). IP addresses were hashed to ensure user privacy.

Parsing: Log data were parsed, converted into data structures, and written out as XML. Data structures:

- Client: represents a unique IP and is associated with one or more sessions
- Session: coherent, sequential actions conducted in response to an information-seeking goal or task
- Request: a user action *and* its corresponding Web server response (e.g., both a user click on a link and returning the requested page)

Preprocessing: A Java program corrected erroneous session data, filtered out Web crawlers, and inserted the remaining log data into a MySQL database. Temporary cookies set in the client Web browser were used to estimate session boundaries. Sessions less than 10 minutes apart were conflated into one session, and a single session with requests 30 minutes or greater apart was split into two sessions. Web

crawlers were identified using header information and a predictive algorithm (i.e., session length and time between requests). Sessions lacking referring page or browser type information were also filtered.

Analysis: We assessed high-level Web usage statistics (i.e., page view and referral frequency) as well as page transitions and navigation path frequency. To facilitate data analysis, pages were aggregated into functional categories (Table 1).

Function	ClinicalTrials.gov Page(s)
Search	Basic and Focused search
Browse	Browse hierarchy
View Results	Results (search or browse)
View Study	Specific clinical trial(s)
Static Pages	Web site help pages
External Link	Redirect to an external site
Query Details	Query assistance
Map Results	Narrow search by location

Table 1. Categories for page-based analyses

A single-transitions table was built by computing the frequency of users moving from one page to another within sessions. Artificial "start" and "end" states were added to ClinicalTrials.gov sessions for clarity. A graph of the network was used to depict prevalent user-site interaction patterns.

Navigation paths, or contiguous moves over a session, were determined algorithmically. Paths were "condensed" to fold multiple consecutive page occurrences into a page-plus-fold marker (e.g., *View Results* → *View Study* → *View Study* folded into *View Results* → *View Study(N)*). Folded paths were clustered and analyzed manually. (Multiple, consecutive *View Study* pages reflect use of the *Back* button, which is not recorded in Web server logs.)

Pilot User Study: We conducted a structured assessment to explore how consumers search for clinical trials using hypothetical scenarios on sleep apnea and Parkinson's disease. Participants, assigned one of the two scenarios (alternated,) were asked to find information from their choice of online resources using Internet Explorer and thinking aloud. TechSmith's Morae™ captured user-computer interactions and verbal responses. The sessions ended when participants felt that relevant information had been found or believed it did not exist online. We analyzed click stream data (e.g., Web page changes and keystrokes), participant comments, usage statistics, and navigation data.

RESULTS

The unprocessed ClinicalTrials.gov log files consisted of 4.9 million sessions. After parsing and preprocessing the logs, the remaining 1.4 million sessions—5.5 million requests for 675,000 unique clients—were analyzed. *View Study* pages containing trial summaries were the most frequently requested and the most common *entry* and *exit* points to ClinicalTrials.gov (Table 2).

Usage Type	Category	Freq.
Page View	View Study	40%
	View Results	25%
	Opening Screen	10%
	Browse	6%
	Search	6%
	Static Page	5%
	Query Details	5%
	External Link	3%
	Map Results	1%
Session Initial Page	View Study	39%
	Opening Screen	28%
	View Results	24%
Session Ending Page	View Study	57%
	View Results	17%

Table 2. ClinicalTrials.gov usage statistics

External Web sites such as search engines and government sites initiated 69% of all sessions; Google and other search engines were half of these. Table 3 lists the top five referring Web sites, accounting for 66% of all referrals.

Referrer Site	Sessions	Freq.
Google*	541,345	41%
NIH.gov domain	219,170	17%
Yahoo*	63,459	5%
MSN*	30,305	2%
AIDSinfo	16,967	1%
Total	871,246	66%

Table 3. Top ClinicalTrials.gov referrers
*** Includes national variants (e.g., www.google.com.uk)**

Figure 1 (next page) is a transition network illustrating the most frequent page moves by users, where the 94% of all moves are represented. The diagram shows 3 pages serving as Web site entry and exit points: *View Study*, *View Results* (list of trial records resulting from a search), and *Opening Screen* (homepage). The most frequent moves between pages are: users viewing two studies in a row, viewing a study and exiting, and clicking on a specific study in the results list. In general, this diagram shows users entering the site and moving between a limited set of pages rather than traversing a structured hierarchy or starting at the homepage.

Analysis of the 100 most frequent user paths yielded 8 common patterns representing 75% of all sessions (Table 4, next page). Each pattern represents the main task but may include clicks on help pages or external links that do not detract from the task. These patterns are consistent with the transition network and descriptive statistics: users generally accessed ClinicalTrials.gov directly at the study level.

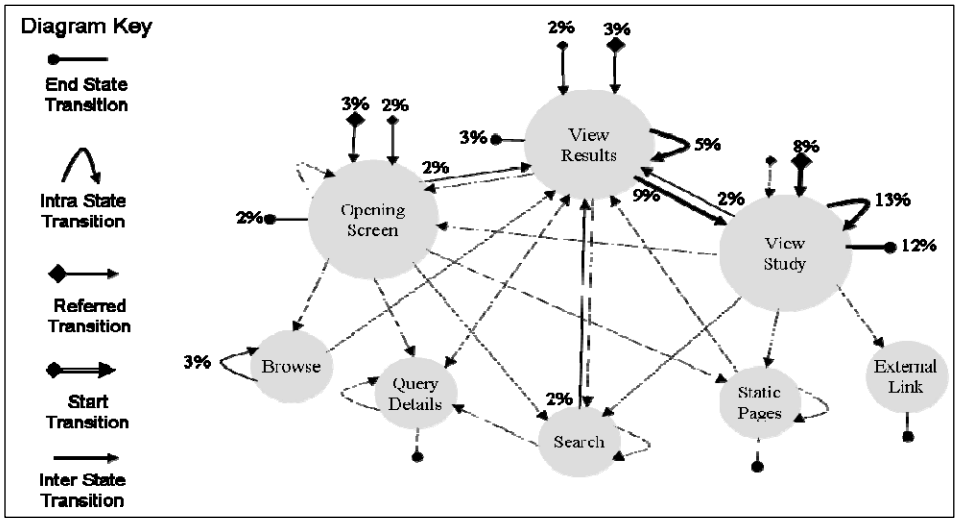

Figure 1. Frequent transitions model from TLA data at ClinicalTrials.gov

Path Pattern* (Main Task)	Freq.
View Study(N)	40%
View Results(N) → View Study(N)	9%
View Results(N)	9%
Opening Screen(N)	5%
Search or Browse → View Study(N)	4%
Search(N) or Browse(N)	4%
Static Page(N) or External Link(N) (only)	3%
View Study(N) → Search(N)	1%
Total	75%

Table 4. User path pattern frequencies
***Includes ancillary tasks such as help page clicks**

Pilot User Study

Seven lay consumers participated in the study during February 2006. While all participants began information seeking using search engines, five viewed at least one page at ClinicalTrials.gov during their session (Table 5).

ID	Referring Page	Entry Page	Exit Page
1	Google	View Results	View Results
2	Yahoo	Opening Screen	View Study
3	MedlinePlus	View Results	View Results
4	MedlinePlus	View Results	View Results
5	Non Profit	View Study	View Study

Table 5. Overview of navigation actions at ClinicalTrials.gov by participant (n=5)

Of the participants who used ClinicalTrials.gov, two were referred directly from search engines, two from MedlinePlus, and one from a non-profit health Web site. Participants most frequently entered and exited at the *View Results* (3 of 5 in both cases) in the pilot study, which differed from the TLA data where *View Study* page was the main entry/exit point.

Overall, participants made rapid relevance judgments regarding search results, quickly scanning hit lists for site information and keywords. Web site relevance was judged primarily by name recognition (e.g., *Mayo Clinic*), "dot-gov" domains, or keywords (e.g., *National* as indicator of an authoritative site) and no participant checked the *About Us* page for any site, consistent with Eysenbach and Köhler [9]. Several participants were confused by the vocabulary at ClinicalTrials.gov, causing one to note:

> I hate to tell you but I'm an LPN [Licensed Practical Nurse] and really, some of this stuff, I have no idea what it is.

While several participants did not complete the scenarios successfully, all felt they found relevant and useful information. Participant satisfaction with sub-optimal online health information seeking outcomes has been observed previously [10].

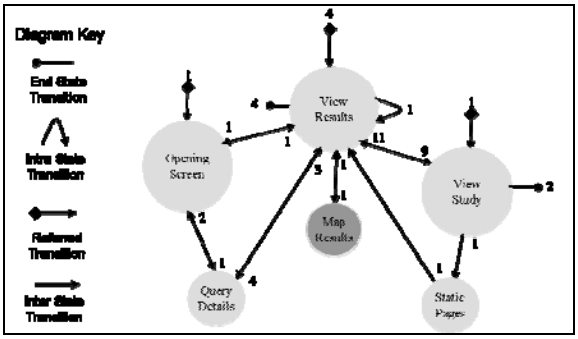

Figure 2. Participant moves on ClinicalTrials.gov

When a single-transition diagram from the user study (Figure 2) was compared to the one representing the three-month log data (Figure 1), both showed a

similar structure of entry and exit points, with most moves occurring within these pages. Further, external Web sites (e.g., search engines) direct users to results pages as well as individual studies. However, details differed, such as pages accessed, as may be expected given the large discrepancy of sample sizes.

DISCUSSION

TLA provides broad, high-level views of user behavior in naturalistic settings over time. To our knowledge, TLA has not been widely applied at a user session level for exploring consumer health information seeking. In summary, the findings reported here suggest the following profile of user online navigation activities at ClinicalTrials.gov:

- Individual trial records (*View Study*; 40%) are the most frequently viewed page, followed by the query results (*View Results*; 25%).
- Users enter the site at a trial record (39%) more often than the homepage (24%).
- External Web sites refer a majority of users (69%); where Google (41%), other NIH-sponsored sites (18%), and other search engines (7%) are the top referring sites.
- The most common user session is viewing one or more trial record (40%), followed by viewing a results list and clicking on a study or simply viewing a results list (9% each).

Contrary to the initial expectations and intensions of the ClinicalTrials.gov Web designers [11] and Web-design guidelines (e.g., [12]), many users do not (1) enter from the homepage where a detailed menu of options is provided; (2) directly use the search and browse features provided by the site; and (3) spend time exploring the site and refining their search queries for their information needs. However, an original design goal of the ClinicalTrials.gov site, to require a minimal number of clicks from the home page to reach a trial record [11], was supported.

One critical factor may be the increasing use of search engines, particularly Google, regardless of the availability of high-quality domain-specific resources [10]. As a result, the most "relevant" sites/pages indexed by the search engine not only get the most exposure, but the most direct visits. For ClinicalTrials.gov, these are generally individual trial records (i.e., *View Study*) and pre-specified URL-encoded search queries (i.e., *View Results*).

Our preliminary results show that many users only view these pages and then return to the referring site using the Web browser's *Back* button to potentially click on another study link, indicative of the hub and spoke behavior pattern [7]. While such users obtain study information, most do not take advantage of the other site features, such as *Search within Results* (narrowing a query) or *Resources*, (background information on clinical trials). It is likely that users do not realize that additional, highly relevant information might be accessible at the site.

New design strategies are required to aid users in finding features that may help satisfy information needs. Descriptive text and/or site instructions on a consumer health site homepage are not seen by users entering a site at low-level pages. Thus, providing links to background information and other search features on low-level pages (directly) might prompt users to continue searching within the site. Similar to recommendations to add site indexes throughout a document space [7], this approach supports the hub-and-spoke behavior pattern and orients users, increasing the chance for exploring related information. In addition, there is evidence that users of Web sites containing visible local maps are (1) less likely to abandon information-seeking episodes, (2) delve deeper into the site hierarchy, and (3) use the *Back* button less [13]. Different research approaches and tools are needed to confirm these hypotheses.

Pilot User Study: We conducted a pilot user study to provide initial TLA validation by demonstrating convergence of the data from the two methods. An additional goal was to begin evaluating methods to compensate for TLA limitations and illuminate consumers' information seeking goals on Web sites like ClinicalTrials.gov. As with the TLA referral data, participants used search engines to find clinical trials information. Participants also exhibited a hub-and-spoke navigation behavior pattern where search engine results pages and other information pages, such as MedlinePlus, served as hubs.

While four of the five users who arrived at ClinicalTrials.gov entered at low-level pages, three entered at *View Results*. TLA data indicated most users entered at the individual trial record level (*View Study*), so the two methods provided differing data. To investigate this trend, we reviewed queries for users referred to *View Study* pages associated with the scenario disorders to look for similarities. We found most queries used specific terms, (e.g. "provigil,") or were complex queries, (e.g. "sleep apnea mirtazapine,") linking to specific trials. Additionally, queries for users referred to *View Results* pages contained general terms "apnea" or "parkinson." The study participants also searched with generic terms as they were instructed to find clinical trials on the disorders, indicating they may not be representative of the Web site users overall.

TLA and user studies may be complementary, but to integrate the approaches, strengths and weaknesses of

each method must be assessed. TLA provides high-level trends over many user sessions, while user studies provide rich data about individual user actions. Thus, further research is needed to understand how to integrate methods for data at two levels of granularity (i.e., population and individual) and ways to correlate these data when the generality of the user group and the demographics of the population are unknown.

In this exploratory study, we used existing techniques of TLA and user studies to investigate the complex phenomenon of consumer information seeking on a health-related Web site. The pilot user study was intended to increase the understanding on consumer behavior online by providing possible correlations among data from different sources. Future research using both the TLA and user study methods, explored in this preliminary study, will be needed to validate this approach.

Limitations: Sessions were estimated using cookies and time constraints as boundaries; system performance may impact this through load balancing and reboots. Data were analyzed for a three-month period; different time spans may change findings. Finally, the pilot study is based on one consumer health Web site and the data may not be representative of navigation behavior due to design, content or audience.

CONCLUSION

Using TLA techniques, we investigated user navigation on a consumer health Web site, ClinicalTrials.gov. We found that a significant number of users were referred to low-level pages from external Web sites, mainly search engines. The behavior observed through TLA and the pilot user study has implications for Web site design and usability. While this paper investigated user behavior on a consumer health Web site, the approach is generalizable. For example, increased search engine usage is not limited to consumers. Professional medical journals report large numbers of referrals to their online sites by search engines: "Web-based search engines are transforming our use of the medical literature" (p. 4 [14]).

We plan extend our research first by addressing limitations mentioned above and continue to improve TLA techniques and path clustering. Further we plan to include semantic information such as user queries to better understand what information users are seeking on the Web site.

ACKNOWLEDGMENTS

This work was supported by the Intramural Research Program of the NIH, NLM. We thank Nick Ide for his assistance with the Web logs and Graciela Rosemblat for reviewing the manuscript.

REFERENCES

[1] Pew Research Center. Trends 2005: Information for the Public Interest. Washington, DC: Pew Research Center; 2005.

[2] Greenberg L, D'Andrea G, Lorence D. Setting the public agenda for online public health: a white paper and action agenda. Washington, DC: URAC, 2003.

[3] Spink A and Jansen BJ. Web Search: Public Searching of the Web. 2004. Dordrecht: The Netherlands: Kluwer Academic Publishers; 2004.

[4] McCray AT, Tse T. Understanding search failures in consumer health information systems. Proc AMIA Symp 2003: 430-4.

[5] Chen ES, Cimino, JJ. Automated discovery of patient-specific clinician information needs using clinical information system log files. Proc AMIA. 2003:145-9.

[6] Jansen BJ, McNeese MD. Evaluation the effectiveness of and patterns of interactions with automated searching assistance. JASIST. 2005;56(14):1480-1503.

[7] Catledge LD, Pitkow JE. Characterizing browsing strategies in the World-Wide Web. Proc Third WWW Conf. 1994.

[8] Srivastava, J, Cooley R, Deshpande M, and Tan PN. Web usage mining: discovery and applications of usage patterns from Web data. ACM SIGKDD Explorations. 2000;1(2):12-23.

[9] Eysenbach G, Köhler C. How do consumers search for and appraise health information on the world wide web? BMJ. 2002 Mar 9;324(7337):573-7.

[10] Zeng QT, Kogan S, Plovnick, RM, Crowell J, Lacroix EM, Greenes RA. Positive attitudes and failed queries: an exploration of the conundrums of consumer health information retrieval. Int J Med Inform. 2004 Feb;73(1):45-55.

[11] McCray AT, Dorfman E, Ripple A, Ide NC, Jha M, Katz DG et al. Usability issues in developing a Web-based consumer health site. Proc AMIA Symp 2000:556-60.

[12] Nielson J. Designing Web Usability: The Practice of Simplicity. Indianapolis, IN: New Riders Publishing. 2000.

[13] Danielson, DR. Web navigation and the behavioral effects of constantly visible site maps. Interacting with Computers. 2002; 14(5):601-18.

[14] Steinbrook R. Searching for the right search--reaching the medical literature. N Engl J Med. 2006 Jan 5;354(1):4-7.