REPORT

Capturing Users' Behavior in the National Science Digital Library (NSDL)

Bing Pan

Human-Computer Interaction Group

Information Science Program

Department of Communication

Cornell University

May, 2003

Capturing Users' Behavior in the National Science Digital Library (NSDL)

In order to design better information technology to fit users' needs, it is essential to capture user's online behavior, e.g. explore how people use different technologies, and what the usability problems are. The development of digital libraries (DL) is facing many challenges nowadays. In the United States, multi-million dollars of funding has been invested in the development of DL every year from National Science Foundation (NSF), Defense Advanced Research Project Agency (DARPA), National Aeronautics and Space Administration (NASA) and other federal and private agencies. However, how these digital libraries were actually used and their usefulness is still under investigation (Thong, Hong & Tam, 2002). One important question to ask is, do users actually use digital libraries as the developers intended? One way to answer this question is to use captured web log data from digital library web sites and explore their usage patterns.

This report firstly gives an introduction to NSDL; then the complexity of capturing user's behavior on the Internet is discussed. The report then discussed the details of this project, including the web log analysis tools used, data cleaning process, the results of data analysis and its interpretation. Finally a general conclusion was given and its implication for digital library design is provided.

National Science Digital Library (NSDL)

According to National Science Foundation (NSF), NSDL program aims to develop, create and maintain a digital library in a national level which supports technology, science, engineering, and mathematical education (NSF, 2003). NSDL intends to provide a rich environment for education for various groups including K-12, college, graduate, and professional students and educators. The NSDL research program includes six major projects in Phase I, and the major portal http://www.nsdl.org publicly released online on December 5th, 2002. From the date of its publication, NSDL attracted many visitors around

the world. The web log recorded on the NSDL web server provides us with valuable information on how users use the web sites.

Capturing Behavior in Networked World

However, against the hyped view of great potential of network technology to capture user's behavior on the Internet, capturing the behavior of networked users is not as simple as it sounds. When the Internet first came into being, researchers around the world were amazed at its potential to capture everything which is happening in the networked world (Nicholas, Huntington, Lievesley & Withey, 1999), which is not possible in a physical world. On the other hand, the Internet today has underwent 20 years of technology evolution. Various technologies on the Internet and their adoptions made behavioral data collection online a tricky business. Take the web as an example, the information the user perceived can hardly be captured using one sets of data. Considering the following figure (Figure 1):

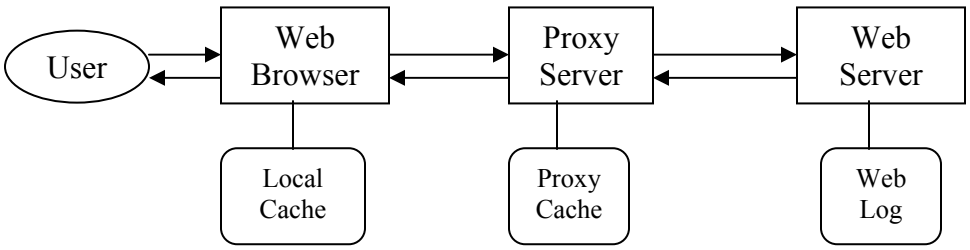

Figure 1. User's Interaction when Retrieving a Web Page

When a user either type in a URL or click on a link, the request will be sent out across the Internet through the web browser, maybe one or more proxy server, and will possibly reach the web server. However, the web browser may contain cached web page which has not expired. The web browser will return cached web page even without passing to next network node; the proxy server may also contain not-expired web page and return to the web browser as a result and will not pass the request to next networked node; if all of the browser's local cache or proxy server's cache do not contain the web page, the web server will return the web page and leave a record in the web log file. From this description, we can see that trying to capture web behavior is always difficult since there are maybe many variance involved. For example, it can not capture the clicks of Back and Forward button on

the web browser, since when the user clicks on these buttons, most of the time the web browser will retrieve the web page from its local cache. Also, if the users' browser was set up to point to a web proxy, the web server can not capture the activities when the web proxy provides the requested web pages from its local cache instead of requesting document from the remote web server. The web log data contains valuable information but it's very hard to translate into actual user's behavior. In this sense, statistical analysis is important (Nicholas, Huntington, Lievesley & Withey, 1999). For example, by studying the real data, we may be able to identify in a statistical sense what percentage of user's request will reach the web server and based on the web log data we can predict the real user's volume.

Even though in terms of capturing web visitor's activities, the web server log has many limitations., web logs are still valuable information source for investigating web visitor's behavior if the web proxy and local caching follow similar pattern and the visiting to web server can still represent analogous model but in a smaller scale. Most web servers log their activities using similar format: when one remote client requests a file on the server, the web server record the client's IP address, the referral source, the file being requested, timestamp, the results of the request.

Analog, ReportMagic, and QuickDNS

There are many log analysis tools available on the market, from free ones to commercial products. According to the author, Analog may be the most widely used log analysis tool available online (Turner, 2003). Different from other web log analysis tools, Analog can be customized according to user's needs with different options (see attached configuration file for an example). For the purpose of analysis the researchers will have more flexibility to look at different aspects of users' behavior.

ReportMagic is a program created by Wadsack-Allen Digital Group which can work together with Analog data to perform more aesthetically appealing results. It parses the data generated by Analog and produced web-based user-interface, including different types of graphics.

QuickDNS is another tool working with Analog. Analog is in general a fast program which can process large amount of log data in a short period of time. The bottleneck of Analog is

its DNS lookup, since DNS lookup needs to communicate with the DNS server, which is largely depend on the DNS server's speed and network speed. QuickDNS can resolve most of IPs in the log file and put them into a text file. When analog is running, it will read resolved DNS file and use the parsed host names instead of IP addresses.

Log Analysis of NSDL

NSDL.org used Tomcat (a type of Java servlet technology) to parse web request and fetch different web pages. Tomcat will automatically insert a session ID into user's query string. Since one web page can have different URL across different sessions, some web analysis tools are not able to handle the comparison between different sessions. Therefore, some results are biased. Developers in NSDL groups have already been doing log analysis in a limited fashion. The analysis is monthly based, and there is no way distinguish log before and after the NSDL.org went public. The analysis used AWStats toolkit (http://awstats.sourceforge.net/), which has limited functionalities.

Data Cleaning

The log data produced by NSDL apache web server was not clean. There are several problems involved with it which was not addressed properly in the web log analysis performed previously:

1. Before February 3rd, the log file was captured in Common Log Format style and after that date the log format was changed to NCSA log format style with additional information about referral and client's web browser type (see Figure 2 as an comparison between these two types of log format). Therefore, a Perl program was written to transform the first format into the second one using pattern matching and replacing in Perl.

Common Log Format:
Example: 12.15.169.15 - - [17/Dec/2002:12:01:13 -0500] "GET / HTTP/1.0" 302 647
Format: IP – Date –Request – Response Code – Size of Response

NCSA Log Format:
Example: 68.45.45.69 - - [03/Feb/2003:22:32:46 -0500] "GET / HTTP/1.1" 301 301 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)"
Format: IP – Date – Request – Response Code – Size of Response – Referral – Browser Type

Figure 2. Comparison of Two Types of Log File Format

2. Furthermore, the Analog software was not able to parse some records in the web log. For example, the Request field in some records contain double quote (") which will confuse Analog program since double quote (") was used to enclose request, referral, and browser type fields. Therefore, a Perl program was written in order to delete all these double quote (") inside GET request. Some Request field in some records contains Java applet file.

Results

The log data used in this project was captured during December 5, 2002 (the date of its public release) to April 16, 2003, in about six months' period. In total there are 2,186,711 records in the aggregate log file, which is 454 MB of disk space. However, with the help of QuickDNS, the running time for Analog is only about 15 minutes on a Pentium III machine. The following paragraphs detail the results of the analysis and their interpretations:

General Pattern

There are in total 14.48 GB data was transferred from the NSDL web server in the last around 6 months; the web server served 21,229 distinct files with its 14,229 distinct files. There are 1,925,168 successful requests and 101, 379 failed requests.

Monthly Report

When we look at Figure 3, we can see that there is an increasing trend of requests in the monthly reports, especially in the March of 2003, the volume of requests was drastically

increased to more than three times on February of 2003. After talking with NSDL staff, it is found out that the increase was caused by a program requesting NSDL.org from inside Cornell University. Therefore, the monthly report was biased by non-human log records.

Hourly Report

From Figure 4, we can see during all time period, the working hours, from 8AM to 5PM, is the peak time of visits. During working hours, the request volume peaks at 12:00PM to 1:00PM, indicating that users usually access NSDL at working time.

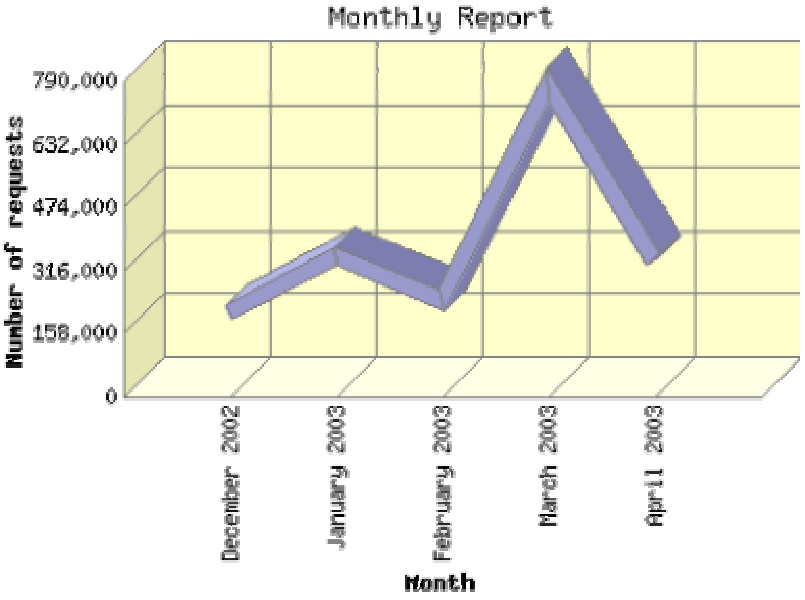

Figure 3. Monthly Report of Volume of Requests

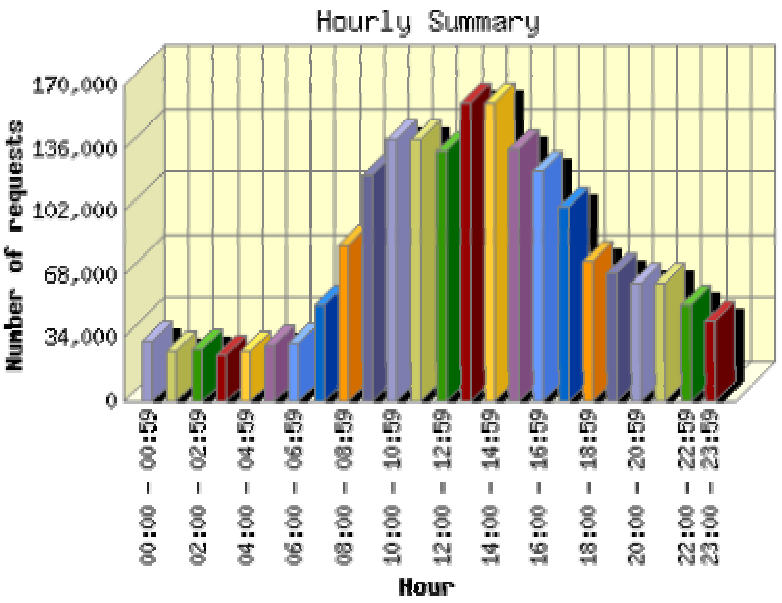

Figure 4. Hourly Report of Volume of Requests

Domain Report

Generated domain report shows that in total 90 countries access NSDL.org besides United States. The number one domain is .edu, e.g. American educational institutions, which constitutes 30.4% user requests. 23.4% of user requests came from .com, e.g. commercial web hosts. Most of them are from Internet providers like Road Runner and At&T broadband, which will be shown in another report shortly.

Organization Report

From Table 1, we can see the biggest request for NSDL.org came from Cornell.edu, which is the development university of the NSDL portal. Around 20% percent of volume is internal links; another 11.6% of requests are from the search robot of google. The fourth largest request volume is from ucar.edu, which is University Corporation for Atmospheric Research and a major participants of National Science Digital Library project. All these indicate that the major volume of traffic came from the development teams. Even Road Runner, which has the third largest request volume, may come from the researchers of development teams, since in Ithaca, NY area, the major broadband Internet access provider is Road Runner.

Organization		Number of requests	Percentage of the bytes
1.	cornell.edu	241,841	18.95%
2.	googlebot.com	79,065	11.63%
3.	rr.com	77,826	2.84%
4.	ucar.edu	33,825	1.33%
5.	attbi.com	29,260	1.04%
6.	comcast.net	20,627	0.74%
7.	128.253	18,047	0.85%
8.	syr.edu	16,359	0.51%
9.	aol.com	16,117	1.17%
10.	cox.net	15,102	0.50%

Table 1. Top 10 Organizations Accessing NSDL.org

Search Term Report

Table 2 shows the top 20 most used search terms in order to reach NSDL.org. The users type in these keywords in search engines in order to reach NSDL.org. Here we can see "nsdl", "national science digital library" are the top keywords to reach the web site and there is no topic words like "physics", "mathematics", or "technology". It indicates that only users who know NSDL can will search through google for NSDL.org and users look for information on a specific topic won't reach the web sites.

Status Code Report

Table 3 shows status code report. We can see that 90K among 2M requests are failures, which is around 4.5% failure rate. The failure may come from a broken link outside or inside NSDL.org. Interestingly, 510K requests are request for If-Modified-Since, indicating that web caches and web proxy server are widely used in the user's requests so the browser or the web proxy are inquiry the change status of its cached web objects.

	Search Query	Number of requests
1.	nsdl	895
2.	national science digital library	487
3.	digital library	107
4.	nsdl.org	71
5.	www.nsdl.org	62
6.	national science library	53
7.	tomcat java.lang.outofmemoryerror	53
8.	national digital science library	51
9.	the server encountered an internal error that prevented it from fulfilling this request	47
10.	science digital library	42
11.	apache tomcat/4.0.3 http status 500 internal server error	41
12.	tomcat the server encountered an internal error that prevented it from fulfilling this request	35
13.	java.lang.outofmemoryerror tomcat	34
14.	tomcat error	32
15.	national digital library	30
16.	org.apache.catalina.core.applicationfilterchain.internaldofilter applicationfilterchain.java:269	28
17.	at org.apache.catalina.core.applicationfilterchain.internaldofilter applicationfilterchain.java:269	23
18.	digital science library	20
19.	the server encountered an internal error internal server error that prevented it from fulfilling this request	17
20.	digital library science	16

Table 2. Top 20 Query Terms

Status Code		Number of requests
1.	200 OK	1,414,896
2.	206 Partial content	26
3.	301 Document moved permanently	36,768
4.	302 Document found elsewhere	123,395
5.	304 Not modified since last retrieval	510,246
6.	400 Bad request	552
7.	403 Access forbidden	89
8.	404 Document not found	90,236
9.	408 Request timeout	480
10.	416 Requested range not valid	1
11.	500 Internal server error	10,009
12.	501 Request type not supported	12

Table 3. Status Code Report

Request Report

Request report shows that 50% of requests are for the home page of NSDL.org, indicating that NSDL.org is a "shallow" web site and users are mostly access the homepage and maybe one or two more pages.

From Figure 5, we can see that different view ratio of the menu items on the home page. "Exhibits" was viewed more frequently than the other menu items. "Help" is the least viewed menu item on the menu page.

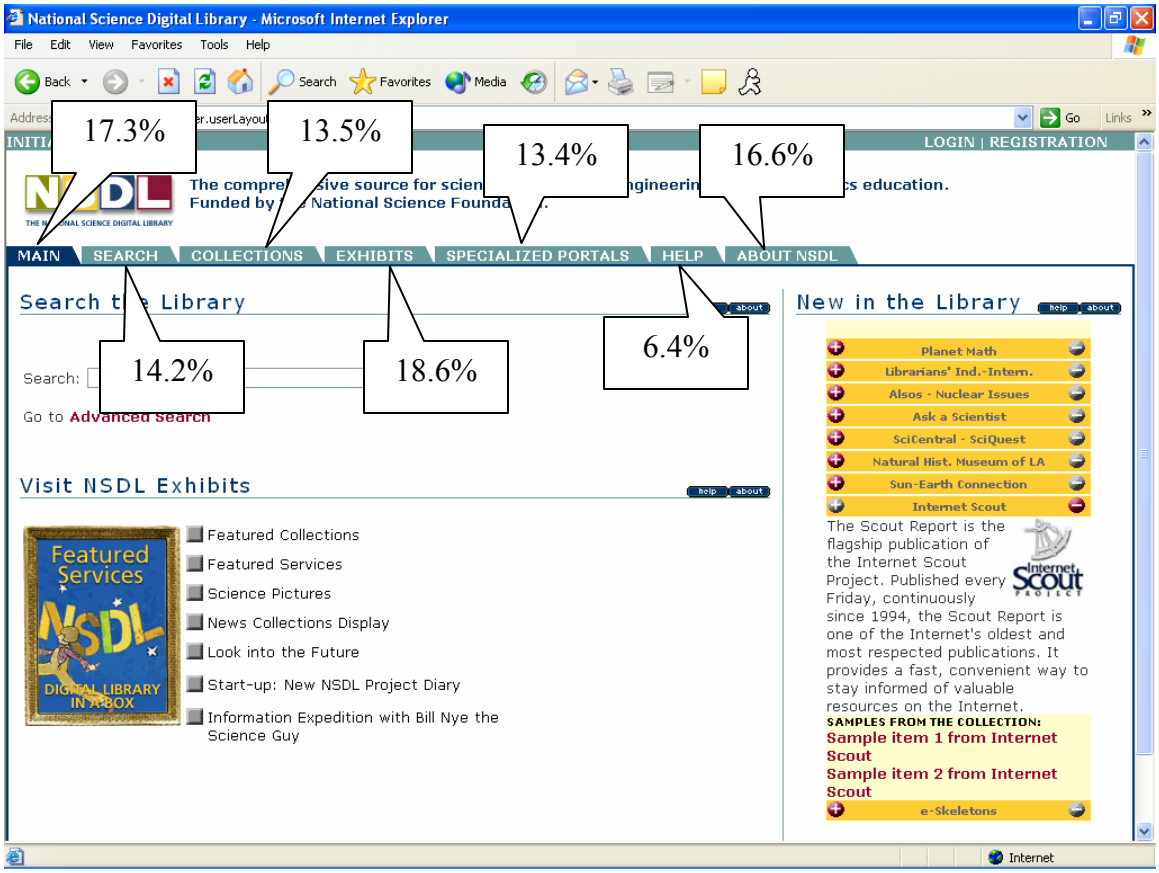

Figure 5. The Request Ratio of Different Menu Items

Conclusion and Future Direction

The previous analysis shows that NSDL.org attracted large amount of volume from all over the world. However, the major volume is from the development teams around Cornell University and collaborative organizations. Because of the nature of web log data, the results are not conclusive. We are not sure about the requests from rr.com are from developers of NSDL.org or from general users. The results showed that the increased volume of NSDL but we are still not sure whether or not these increased volume is from the developers or general users.

The analysis of the web log data in this project follows a traditional way, e.g. focusing on simple statistics, for example, number of hits in a certain period of time, or types of response code (web page returned or page not found). These data can hardly reveal the behavior aspect of individual web visitors. However, more meaningful and significant tasks is to

explore web visitor's model through sequential analysis, in other word, the sequence of traversing the digital library web pages. The users may follow a visiting sequence when they click through the digital library web page for interested information. The goal of this project is to answer the following questions:

(1) Can we derive common patterns of web visiting sequence through web log analysis?
(2) Can we distinguish different user groups from the different web page visiting sequence?
(3) How can we identify pitfalls of digital library design from sequential analysis (for example, which web pages were never visited or the users need to traverse a large amount of web pages in order to reach the destination web page)?

A lot of researchers have used different methods to analyze and visualize the data (Benabdeslem, Bennani & Janvier, 2002; Joshi, Joshi & Krishnapuram, 2000; Pei, Han, Mortazavi-Asl & Zhu, 2000; Spiliopoulou, 1999), including analysis on e-journals and online libraries (Ke, Kwakkelaar, Tai & Chen, 2002; Rozic-Hristovski, Hristovski, Todorovski, 2002). To compare similarity of sequence, several methods could be used but haven't been addressed in web log research:

(1) String-editing method: The basic idea is, in order to compare two strings, the degree of differences can be represented by the number of basic operations needed to match two strings. For example, in order to match A B C with B A C D, we need to move one letter B and add one letter D. The basic operations needed are two. This method can be used to generate the similarity matrix of different search patterns;
(2) Genetic algorithm;
(3) Hidden Markov Model: HMM was used extensively in sequence mapping. It can identify the underlying model behind different sequences.

Research in these directions will surely feasible and meaningful. Again, one important question is the translation between data of web log and user behavior. More controlled experiment and large scale statistical analysis are urgently needed.

References

Benabdeslem, K., Bennani, Y., & Janvier, E. (2002). Visualization and Analysis of Web Navigation Data. J.R. Dorronsoro (Ed.): ICANN 2002, LNCS 2415, pp. 486-491.

Joshi, A., Joshi, K., & Krishnapuram, R. (2000). On Mining Web Access Logs. ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery.

Ke, H., Kwakkelaar, R., Tai, Y., & Chen, L. (2002). Exploring Behavior of E-Journal Users in Science and Technology: Transaction Log Analysis of Elsevier's ScienceDirect OnSite in Taiwan. Library & Information Science Research, 24: 265-291.

National Science Foundation. (2003). National Science, Technology, Engineering, and Mathematics Education Digital Library (NSDL): Program Solicitation. Available Online at: http://www.nsf.gov/pubs/2003/nsf03530/nsf03530.htm.

Nicholas, D., Huntington, P., Lievesley, N., & Withey, R. (1999). Cracking the Code: Web Log Analysis. Online & CD-ROM Review, 23(5): 263-267.

Pei, J., Han, J., Mortazavi-Asl, B., & Zhu, H.(2000). Mining Access Patterns efficiently from Web logs. Proc. 2000 Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD'00), Kyoto, Japan, April 2000.

Rozic-Hristovski, A., Hristovski, D., Todorovski, L. (2002). Users' Information-seeking behavior on a Medical Library Website. Journal of Medical Library Association, April, 2002: 210-217.

Spiliopoulou, M. (1999). The Laborious Way from Data Mining to Web Mining. Int. Journal of Comp. Sys., Sci. & Eng. 14 (1999), Special Issue on Semantics of the Web, 113-126.

Thong, J.Y.L., Hong, W., Tam, K. (2002). Understanding User Acceptance of Digital Libraries: What are the Roles of Interface Characteristics, Organizational Context, and Individual Differences?. International Journal of Human-Computer Studies, 57(3): 215-242.

Turner, S. (2003). Analog User's Manual. Available Online at: http://www.analog.cx/docs/Readme.html.

Wadsack-Allen Digital Group. (2003). Documentation for Report Magic 2.21. Available Online at: http://www.reportmagic.org/docs/index.html.

AnalogX. (2003). QuickDNS ReadMe. Available Online at: http://www.analogx.com/contents/download/network/qdns.htm.