Integrating Bottom-Up and Top-Down Analysis

For Intelligent Hypertext

James E. Pitkow & Mimi Recker

Graphics, Visualization & Usability Center
College of Computing, Georgia Institute of Technology
Atlanta, GA 30332-0280
E-mail {pitkow,mimi}@cc.gatech.edu

ABSTRACT

The range of hypertext systems continues to expand, from custom-tailored closed systems to dynamic, distributed, and open systems like the World-Wide Web (WWW). The shift from closed to open systems results in a corresponding decrease in the effectiveness of metrics and techniques for providing intelligent hypertext to users. Essentially, the locus of control shifts away from developers towards users. Stated differently, the central question becomes chaotic, loosely constrained environments, like the World-Wide Web, can provide intelligent hypertext. This paper argues that viable answers are derivable from both bottom-up and top-down analyses. Furthermore, intelligent hypertext within open, client-server systems may profit by combining these two approaches. Using the WWW as a case study, a method of analysis for each approach is presented, accompanied by a discussion of the implications for implementations in open hypertext systems.

INTRODUCTION

The proliferation of network and Internet[1]-based resources has dramatically increased the amount of information available. Not surprisingly, it has also increased the number of protocols and native browser software necessary to access the information. As the amount of information available increases, so does the complexity of discovering, retrieving, and filtering documents and data. Aimed at simplifying this task, the World-Wide Web (WWW) [Berner-Lee, et. al. 1992] provides hypermedia access to the widely distributed and heterogenous collection of existing information resources via the Hypertext Transfer Protocol (HTTP) and the Hypertext Markup Language (HTML). Currently, WWW enables seamless access to such Internet-based information resources such as Gopher, File Transfer Protocol (FTP), Wide Area Information System (WAIS), and Network News Transfer Protocol (NNTP), to name just a few. Although a precise growth rate is impossible to determine, the dramatic increase in the use of WWW technologies since inception in 1991 is evidenced by the expansion of HTTP servers (611 in 1992 to over 7,300 in early August of 1994) and the number of users accessing each site. In addition, NSFNET backbone byte counts by port number shows WWW as a top ten information service as of August of 1994 [Merit NIC 1994]

Nonetheless, technical difficulties remain with WWW, with the foremost issue possibly being global organization. While a complete discussion of the problem is beyond the scope of this paper, the most salient ramification is that without a global structure, the ability for users to perform global querying, (i.e., searches that access all resources in all countries and domains), and global browsing (i.e., exploration based upon subject, location, etc.) are not implicitly supported by the WWW protocol. In short, not only does this make intelligent information retrieval very difficult, but it also does not support an interface to the query's results.

Essentially, WWW provides the underlying structure for *global hypermedia* browsers that provide the interface for accessing Internet-based information resources. One of the first global hypermedia graphical user interface (GUI) browsers to be developed was the National Center for Supercomputing Applications' (NCSA) Mosaic for the X Window System (X Mosaic). Officially released in June of 1993, initial empirical research estimate[2] [Koster 1994] that X Mosaic currently accounts for roughly 53% of all WWW related accesses to HTTP servers[3]. While accounting for over half of the client access is impressive, NCSA's X Mosaic has also had a large effect on WWW browser design and functionality. This is evident in the new browsers released by other developers and companies. In essence, X Mosaic has become the current de-facto standard for design and functionality of WWW browsers. Despite this, the interface developed does not incorporate intelligent hypertext interface technology and did not evolve from user-centered design methodology.

With the goal of understanding the design of intelligent hypertext for WWW browsers, we took both a bottom-up and top-down approach in our analysis. Specifically, in the bottom-up approach, we investigated individual user hyperlink usage patterns, This was accomplished by modi-

1. For the purposes of this paper, *network-based* and *Internet* will be used interchangeably, though network-based information systems are not limited to the Internet, e.g., DEC-Net, LAN's, etc.

2. NCSA's Mosaic for the Machintosh and Mosaic for Windows almost evenly account for the another 30% and 14% respectively, resulting in NCSA's browsers being used for 83% of current WWW accesses for the HTTP server being studied.

3. Note that the above usage data support the selection of X Mosaic as the key browser for monitoring user activity.

fying X Mosaic to log the time and the hyperlink that each user traversed. Hyperlinks are uniquely identified by a Uniform Resource Locator (URL) [Berner's-Lee 1994] in WWW. This analysis thus provides us with a view of how individual users choose to search and access documents in the WWW space.

The top-down approach is a complement of the first approach. Here, we analyzed the access patterns of Internet users to our local WWW database. In this second approach, we applied a model from psychology research on human memory to perform post-hoc analysis on the assesses made to the Georgia Institute of Technology WWW server during a three month period from January 1994 through March 1994.

While the bottom-up approach reveals individual user strategies for accessing information in distributed databases, the top-down approach reveals overall patterns of access in one database. From a network point of view, the former is a client concern and the latter is a server issue. Moreover, we believe that these approaches need to be considered in tandem in order to provide the basis for designing open hypertext systems that adapt to individual user search strategies against a backdrop of global access patterns in specific databases.

BOTTOM-UP: MONITORING INDIVIDUAL USERS

Given the anticipated diversity in users' browsing strategies and hyperlink traversals, we sought to monitor as many users as possible. To this end, we gained approval to provide all of the faculty, staff, and students at the Georgia Institute of Technology's College of Computing who use X Mosaic on Sun OS 4.1.3 machines, with the opportunity to participate in the study. The study was conducted for three weeks during August 1994. Participation was solicited via a window that appeared the first time each user executed X Mosaic.

Approximately 63% of the available user pool volunteered to participate, resulting in data from over 100 subjects. The computing environment consisted of 250 machines with the syslog configuration files remapped on all Sun OS 4.1.3 machines to point to a dedicated file server that accepted all relevant logging requests. Thus, if the dedicated machine suffered a system failure, logging of events would not fail, but rather be delayed until the system and or network returned to operation. As would be expected after taking such precautions, neither the dedicated file server nor the network failed during the period of observation.

While the system issues related to logging user actions on a medium scale network are important, the representation of the tasks associated with the application, i.e. Mosaic, facilitated the semantic interpretation of the collected data. That is, the data were recorded in a UIDE [Sukaviriya, et. al, 1993] representation. By this we mean that application actions, interface actions, and interface techniques were the vehicle used in the encapsulation of user event data. In the study, we captured all user events. Table 1 shows the relevant task analysis representation for X Mosaic, per our observations and our recording scheme.The following example shows a logged entry. It corresponds to a user clicking on a hyperlink in the document window that points to http://www.somehwhere/. The user is identified as participant number 5555, and the event was generated from machine foo on August 3rd, 1994 at 12:21:10 a.m.

```
Aug 3 00:21:10 foo.cc.gatech.edu uel: 775887872 5555 1 Mouse Document
Anchor:: http://www.somewhere/
```

While this method effectively monitored the behaviors of Sun OS users, it did not include other users of other platforms. This needs to be emphasized in subsequent generalizations and inferences made from the data. Furthermore, an indeterminate subset of users computed on Sun OS and other operating systems during their normal course of work. As such. there exists users whose interactions with Mosaic were only partially recorded via our study. This remains a confound in our study. Replicating the logging software on all operating systems in the College of Computing presents itself as the most viable, though time-consuming solution to this problem.

RESULTS

Analyzing the user log files first required partitioning the data into usage sessions. Determining session boundaries was ambiguous given the application in question. As such, session boundaries were derived by computing 1 1/2 standard deviations (25.5 minutes) from the mean time between events (555.6 seconds) across all users. Other analysis on the processed log file revealed that users averaged one session per day, with a range of 1 - 84.

Application Action	Source of Action	Interface Technique	Category of Action	Description of Action
Anchor	Document	M	Navigate	Selection of Hyperlink in Document
Hotlist GoTo	MenuItem	M	Navigate	Go to Document via Hotlist
Open Local	MenuItem	M K	File	Open Local File
Open URL	MenuItem, Icon	M K I	File	Open File via a URL
Window History GoTo	MenuItem	M	Navigate	Go to Document via Window History

Table 1. Mapping of X Mosaic user events to UIDE- like representation., where M = mouse; K = keyboard; I = Icon.

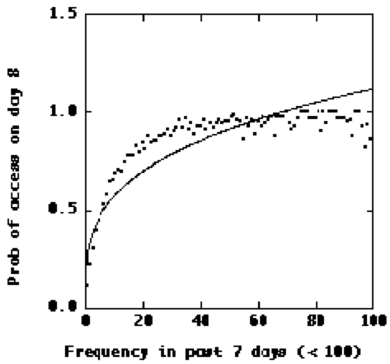

Figure 1.Probability of access against frequency.

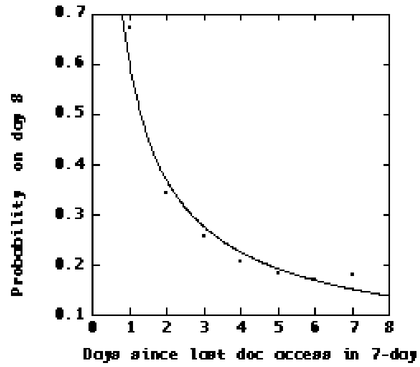

Figure 2. Probability of access against recency.

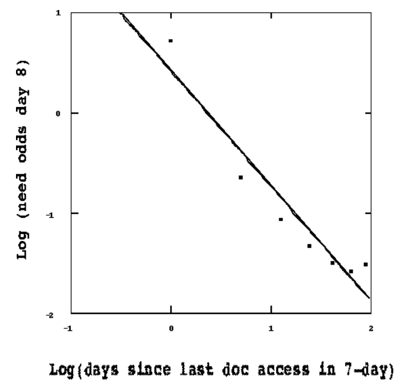

Figure 3. Log -Log linear relation of recency.

These analyses facilitated session boundary construction and began to characterize frequency of use. However, we specifically sought to determine patterns of accesses to different WWW sites on a per user and per session basis. Specifically, the analysis program identified the frequency of the longest repeating sequences of site accesses by using a modified version of the PDM algorithm [Crow & Smith 1991]. From this analysis, we discarded accesses to Georgia Tech machines.

For example, consider a user who visited ftp://sunsite.unc.edu, http://info.cern.ch, and then http://cbl.leeds.ac.uk. This represents a sequence of length three, with the user traversing hyperlinks from Sunsite, to CERN, to Leeds. The analysis program then summed the occurrences of the access patterns. We intentionally discarded sequences of access to only one site, i.e., repeated traversals within one database, thus only identifying between site path analysis. However, we plan to examine within site access patterns in the future, though this might be more effectively examined from the server access logs.

To our surprise, we found long sequences of between site access patterns on a per session and a per users basis. By "per session", we refer to patterns within a session for a single user. Likewise, by "per user", we refer to all sessions by a user, thus allowing between session pattern identification. For the per session analysis, paths of up to seven different sites occurred with a frequency of five times. Stated differently, in one session a user visited seven different sites in consecutive order five times. on a per user basis, the PDM algorithm identified sequences of length eight with a frequency of nine occurrences. Furthermore, numerous shorter sequences were discovered with higher frequencies, with a maximum of 17 times.

TOP-DOWN: MODELING USER ACCESSES

The previous section revealed patterns in user search and document access strategies. This section describes an approach for analyzing and predicting overall patterns of access in a database.

Our analysis of access patterns is based upon a model from psychological research on human memory, which has long studied retrieval of memory items based on frequency and recency rates of past item occurrences (Ebbinghaus, 1885/1964). It was our expectation that human memory, where many items are recalled on a daily basis from a large available store of memories, forms a useful starting point for understanding document access in large, distributed, heterogeneous information spaces, or what we have termed *dynamic information ecologies*.

In particular, we employ a model from Anderson & Schooler (1991) to estimate the probability of future document access using frequency and recency rates of prior document accesses. The model is applied to the log file of accesses made to the Georgia Institute of Technology WWW repository during a three-month period in 1994. At the time of our analysis, the repository contained more than 2000 multimedia documents, and averaged over 3300 document requests per day.

We used their algorithm in order to determine the relationship between the number of document requests during a period (called the window) and the probability of access on a subsequent day (called the pane). This analysis can be viewed as a parallel to the practice function in human memory research. In this case, given the frequency of past document requests, we are interested in determining the probability of new requests. Following the algorithm described in Anderson and Schooler (1991), we computed the frequency of all document accesses during each 7-day window in an access log file, and measured the probability of access during the next day (i.e., day 8). We selected a window of 7 days because we intuitively felt that this window would encompass the typical fluctuations inherent in the calendar week.

Similarly, we applied their algorithm in order to determine the relationship between how recently documents are requested during a period and their probability of access on a subsequent day. This analysis parallels the retention function in human memory research. In this case, we are looking at the probability of document access on the eighth day (the pane) based on how many days have elapsed since the document was last requested in the window (still 7 days).

The dataset used in our analysis was the log file of accesses to the Georgia Tech WWW repository during a three month period, January 1 through March 31, 1994. From the log file, we removed all accesses made by Georgia Tech machines. We felt that these accesses may not accurately represent the average user to the data because they often represent users testing new documents or default document accesses made by client programs.

The trimmed log file comprised 35 megabytes of data, with a mean record length of 100 bytes and totaling roughly 305,000 requests. The number of requests ranged from 300 to 12,000 document per day, with a mean of 3379 accesses per day over the three month period. Some individual documents in the database were accessed over 4000 times per week.

RESULTS

Results show that the model predicts document access with a remarkable degree of accuracy. Figure 1 plots the probability of access on Day 8 as a function of frequency of access during the previous 7-day window (for frequencies < 100). The plot shows a strong power relationship between frequency and probability of access. Interestingly, this relationship mirrors the power law of practice found in memory research. As expected, the regression analysis of the log-log transform reveals a linear relationship, accounting for 72% of the variability, $F(1, 94) = 246.53$; $p < .001$; $MS_E = 124.32$; $R^2 = .72$.

Similarly, Figure 2 plots the probability of access as a function of the recency of access. The plot shows the steep negative slope typically found in retention plots in memory research. Figure 3, shows the regression analysis of the log-log transform which reveals a near-perfect linear relationship, and accounts for 92% of the variability, $F(1, 5) = 56.17$; $p = .001$; $MS_E = 3.71$; $R^2 = .92$.

These robust relationships were found despite the non-standard, heterogeneous, and inherently chaotic nature that characterizes WWW repositories, i.e. dynamic information ecologies. Moreover, the frequency and recency relationships mirror those found in the human memory literature. In addition, regression analysis suggest that recency proved to be a much better predictor than frequency.

DISCUSSION

There are many benefits in detecting patterns of site accesses as obtained via client logging. For example, patterns based on actual user behavior can be identified, including unanticipated patterns. Documents can be pre-fetched, hence decreasing connection latency, which can be especially helpful for large file transfer. Temporal dependent usage, for example, visit these sites every Monday, can be automated. Finally, guidelines can be provided for intelligent agents to contact frequently visited sites for updates, that may be of interest to the user.

Similarly, advantages of determining document access within a site include: determining the relevance of sets of hypertext documents and suggesting predefined navigational paths for new users to the site. Note that these paths could differ from the hypertext designer's plans and hence expand the utility of the database.

Server side modelling of global access patterns of users fiurther increases the intelligence gained by open hypertext systems. The gains include but are not limited to:

- rank ordering of files on server side
- presentation/layout based upon frequency/recency metric. For example, presentation could be manipulated by different coloring schemes for important files, larger fonts of relevance, and integration with client side user profile.
- alternative views in response to queries. That is, query results can be presented based on the value derived from the model.
- restructure the document space to improve access to under- utilized files. SImilarly, over-accessed fiels can be prominently linked.

As hypertext systems move to more open, distributed environments, the ability to intelligently adapt to users becomes more complex. In this paper, we have argued that intelligent hypertext systems need consider in tandem user search patterns and global server access patterns. We have presented methodologies for performing both the top-down (server) and bottom-up (client) analyses and discussed the implications for intelligent hypertext systems.

REFERENCES

[1] Anderson, J. and Schooler, L. (1991). Reflections of the Environment in Memory. *Psychological Science*, 2(6):192-210.

[2] T. Berners-Lee, (1994) Uniform Resource Locators. Internet Engineering Task Force Working Draft. URL: ftp://ds.internic.net/internet-drafts/draft-ietf-uri-url-03.txt

[3] T. Berners-Lee, R, Cailliau, J-F. Groff, & B. Pollermann, (1992) World-Wide Web: The Information Universe. Electronic Networking: Research, Applications and Policy.

[4] D. Crow, & B. Smith, (1992) in eds. Beale, R. & Finlay, J. Neural Networks and Pattern Recognition in Human Computer Interaction.

[5] Ebbinghaus, H. (1885/1964). *Memory: A contribution to experimental psychology.* Mineola, NY: Dover Publications.

[6] M. Koster, (1994) Personal communication.

[7] Merit NIC Services, (1994) URL: gopher://nic.merit.edu:7043/11/nsfnet/statistics/1994

[8] P. Sukaviriya, J. Foley, &T. Griffith, (1993). A Second Generation User Interface Design Environment: The Model and The Runtime Architecture. In INTERCHI 1993 Conference Proceedings. Ny, NY: ACM Publications.