Display Time as Implicit Feedback: Understanding Task Effects

Diane Kelly
SILS
University of North Carolina
Chapel Hill, NC 27599-3360 USA
kelly@ils.unc.edu

Nicholas J. Belkin
SCILS
Rutgers University
New Brunswick, NJ 08901 USA
nick@belkin.rutgers.edu

ABSTRACT
Recent research has had some success using the length of time a user displays a document in their web browser as implicit feedback for document preference. However, most studies have been confined to specific search domains, such as news, and have not considered the effects of task on display time, and the potential impact of this relationship on the effectiveness of display time as implicit feedback. We describe the results of an intensive naturalistic study of the online information-seeking behaviors of seven subjects during a fourteen-week period. Throughout the study, subjects' online information-seeking activities were monitored with various pieces of logging and evaluation software. Subjects were asked to identify the tasks with which they were working, classify the documents that they viewed according to these tasks, and evaluate the usefulness of the documents. Results of a user-centered analysis demonstrate no general, direct relationship between display time and usefulness, and that display times differ significantly according to specific task, and according to specific user.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – relevance feedback.

General Terms
Measurement, Human Factors

Keywords
Implicit feedback, display time, task, user profiling, information-seeking context, user modeling, personalization, tailored retrieval.

1. INTRODUCTION
Tailoring retrieval to individuals is becoming an important area of research in interactive information retrieval (IR). User modeling offers the potential of tailoring retrieval to individuals by creating a representation of the user's interests and using this representation to do such things as select retrieval techniques and documents. However, a significant challenge is identifying efficient and reliable techniques for creating and maintaining user models. Specifically, determining how to get information about the user into the model is a difficult problem.

One approach to this problem is to take advantage of the user's previous information-seeking behaviors to identify documents that have been of interest to that person in the past, and use these documents as sources of information for the model. The model could be constructed, for instance, by identifying and recording the documents that the person has looked at and found useful, and automatically classifying those documents according to topic models, derived from the language of the documents. A new search by the user could be associated with one or a few such models, thereby effectively disambiguating the search topic and providing a basis for tailored retrieval.

Recent research has had some success using information-seeking behaviors as implicit feedback for document preference [4, 5, 10, 12, 13, 15, 20]. See [4, 9, 14] for a review and classification of this research. A general finding of this research is that users display documents that they find useful longer than those that they do not [c.f. 4, 5]. However, such studies have been limited because most assume that information-seeking behavior is not influenced by contextual factors such as task, topic and collection.

In the interactive IR literature, it is generally believed that information-seeking behavior is affected by task in a variety of ways [c.f. 19]. Task has been used to explain differences in relevance assessments of information objects and differences in approaches to system use, such as use of search tactics and terms. Empirical evidence demonstrating that users exhibit a range of information-seeking behaviors that can change with respect to task is mounting [1, 2, 3, 17, 18].

Research on implicit feedback has paid little or no attention to task. Most studies have only investigated a single task, such as news or job searching. For instance, Morita and Shinoda [13] and Miller, et al [12] considered the behavior of users interacting with online news services like Netnews and Usenet. Rafter and Smyth [15] considered the behavior of users as they interacted with an online job bank. Kim, Oard, and Romanik [10] studied behavior in a more traditional information-seeking task, finding sources for a research paper, and Cooper and Chen [5] investigated how behavior could be used as implicit feedback in an online library card catalog. Studies that place no limits on the types of information-seeking activities investigated like Claypool, et al [4], make no attempt to measure task, and instead, construe the task as finding useful or interesting information. With little exception, studies of implicit feedback have not characterized information-seeking tasks, or conducted systematic investigations of their impact on the use of observable behaviors as implicit feedback.

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.

Kelly, D. & Belkin, N.J. (2004). Display time as implicit feedback: Understanding task effects. In *Proceedings of the 27th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '04)*, Sheffield, UK, 377-384.

The exception is Kelly and Belkin [8] who found that factors like topic familiarity and task type confound the relationship between display time and relevance in complex ways.

The current study investigates the relationship between information-seeking task and display time, and the potential impact of this relationship on the effectiveness of display time as an implicit measure of document preference.

2. METHOD

Many studies of online information-seeking behavior have relied solely on proxy data, which usually divorces users and their information needs from their information-seeking behaviors. In our study, we were interested in reuniting users with their information needs and behaviors, and therefore selected a naturalistic, case-study approach.

The naturalistic approach allowed us to optimize the ecological validity of the study by providing subjects with an opportunity to engage in multiple information-seeking episodes with tasks that were germane to their personal interests, in familiar environments. Furthermore, this study lasted for fourteen weeks, which allowed us to more fully understand subjects' natural searching behaviors. The case study approach allowed for intensive data collection, which focused on collecting a large, detailed quantity of data about a small number of subjects. Most notably, this approach allowed us to attempt to understand subjects' goals and intentions, and take various measurements and ask specific questions of our subjects throughout the course of the study. We do not claim our seven subjects to be a sample, nor do we claim that our results generalize reliably to larger populations. However, we believe that the quality and comprehensiveness of the data collected in this study can provide much insight into personalization research.

2.1 Subjects

Seven subjects, who were Ph.D. students at a large, research university on the East coast of the United States, completed the fourteen-week study. A fourteen-week time period was selected since it corresponded to a single university semester and provided a useful temporal division for our subjects. Subjects were from seven different programs: communication, comparative literature, electrical engineering, geography, history, mechanical engineering and political science. Of the seven, five were male and two were female. All subjects were between the ages of 26 and 37, and had a mean of 6.4 years of searching experience. All seven claimed to have at least some computer and World Wide Web experience.

2.2 Instruments

As participants, each subject received a new laptop and printer. Upon completion of the study, subjects were allowed to retain the laptop and printer as compensation for their participation.

2.2.1 Logging Software

Each laptop was equipped with the WinWhatWhere Investigator[1], which was launched automatically each time the subject's laptop was started, and executed in stealth mode while the laptop was in operation. The software unobtrusively monitored and recorded subjects' interactions with all applications including the operating system, web browsers, and word processors. Information such as applications used, URLs visited, start, finish and elapsed times for interactions and all keystrokes, including queries, were recorded and stored in a protected data file located on the laptop.

Of primary interest to the study reported here was the length of time a document was displayed in the subject's active web browser window. Elapsed time was used to measure the display time of documents. The logger automatically computed elapsed time by subtracting the difference in start times for successive entries, where an entry is equivalent to a page request, or any other user-initiated action. A "start" time is an indication of when the user performed some action with the computer, whether requesting a web page, copying a file or opening and editing a document. For example, if a user went to the Google home page (D1), entered a query and clicked search, reviewed the search results page (D2) and displayed one of these results by following one of the links (D3), the elapsed time for D1 would be the difference in start times from D1 to D2, while the elapsed time for D2 would be the difference in start times from D2 to D3.

For identical web pages viewed at different times (for instance, if the user went back to the search results page described above, D2), elapsed time entries were cumulated to arrive at the total elapsed time. Cumulations for identical pages were made on a weekly basis, corresponding to each week of the study.

Subjects' web browsers were further directed through a proxy logger; this direction did not disrupt subjects' activities or cause any noticeable lag times. The proxy logger was a custom built logging application that resided on a local proxy server, and saved a local copy of each page request made by subjects.

2.2.2 Consent Form

Subjects were informed during recruitment and at the beginning of the study, both orally and through a Participant Consent Form (IRB#02-269M), that all of the activities that they performed while using the laptop would be logged and that they would not have access to this software.

2.2.3 Entry Questionnaire

Subjects completed an Entry Questionnaire at the start of the study, which gathered background and demographic information, and previous computer and searching experiences. The information obtained from the Entry Questionnaire was used to characterize subjects, but not in subsequent analysis.

2.2.4 Task Questionnaires

The Task Questionnaires elicited the tasks that were of current interest, or that were expected to be of interest, to the user during the study. Task was defined for this study as the goal of information-seeking behavior. Subjects were asked to think about their online information-seeking activities in terms of tasks, and to create personal labels for each task. They were provided with some example tasks such as "writing a research paper," "travel," and "shopping," but in no other way were they directed, influenced or biased in their choice of tasks. Indeed, anything that a user believed was a task was permitted. While it was not always easy for subjects to identify and create labels for their tasks, feedback from exit interviews conducted at the close of the study indicated that subjects could do this consistently, and with few problems.

[1] http://www.winwhatwhere.com

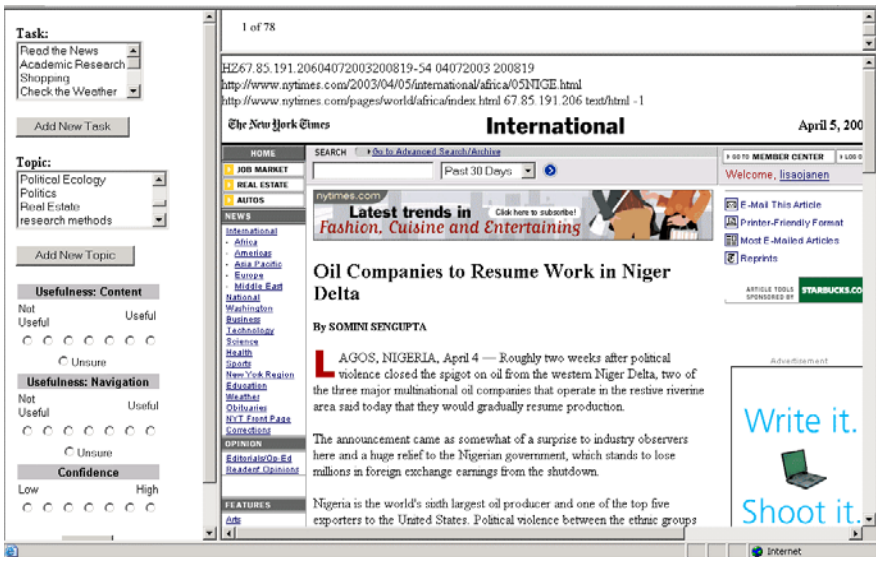

Figure 1. Evaluation interface

Subjects were given a Task Questionnaire at the beginning of the study, and then for each subsequent week of the study they were presented with a list of their previously identified tasks and asked to update the list through additions and/or deletions. Additions were made with an online version of the questionnaire so that new tasks could be used during the evaluation sessions.

2.2.5 Evaluation Interface

For each week of the study, following the presentation of previous tasks, subjects used custom built evaluation software to: (1) classify each document that they viewed according to their self-identified tasks; (2) indicate the usefulness of the document as it related to the task; and (3) indicate their confidence in the usefulness rating that they assigned to the document. Evaluations took place in the office of the investigator.

Our study investigated, in addition to subjects' self-identified tasks, their self-identified topics. In this paper, we do not report on the latter, but mention it here since it appears in Figure 1, which displays the evaluation interface. The evaluation interface presented subjects with the following features:

- A personalized list of the subject's tasks and topics in two separate scrollable windows.
- Buttons that allowed subjects to add new tasks or topics to their lists.
- A scale to indicate the usefulness of the document with respect to the document's content.
- A scale to indicate the usefulness of the document with respect to its support for navigation.
- A scale to indicate confidence in the usefulness judgment.
- A small frame displaying the number of evaluations made during the session and the total number of documents that remained to be evaluated.
- A frame displaying the document.

Usefulness with respect to document content was defined as how useful subjects believed documents were in helping them to complete and/or understand the particular task in which they classified the document. The decision to collect data about the usefulness of a particular item as opposed to its relevance was based on the concept of utility [6] and situational relevance [21], which allows for a user-centered, subjective assessment. Usefulness was measured on a seven-point scale where the scale anchors were "not useful" and "useful." Numeric values were not provided on the scale to give the scale points a continuous, rather than discrete, appearance. Data for each point was coded for analysis with numeric values from 1 to 7, where "1" indicated "not useful" and "7" indicated "useful."

In addition to content usefulness, subjects were asked to indicate the navigational usefulness of the documents that they viewed. The distinction between navigation and content pages is described in [11] where "navigation pages" are defined as "pages that serve to guide the user through a Web site to the information the site was created to provide," and "content pages" are defined as "pages providing that information." This distinction was necessary for this study after interviews from a pilot study [8] revealed that subjects were marking documents as useful because they helped them get to a good page. Because this study was concerned with unobtrusively identifying documents that subjects found useful from a content perspective, navigational usefulness was not of interest and was not considered during analysis.

Confidence was how certain subjects were of the usefulness rating they assigned to a document. Confidence was measured on a seven-point scale, where the scale anchors were "low" and "high." Responses were coded with numeric values from 1 to 7, with "1" representing "low" and "7" representing "high."

For both practical and theoretical reasons, subjects were shown only a selection of the documents that they had requested during the previous week. For instance, some subjects requested over a thousand documents in one week; asking them to evaluate all of these documents during a single session was simply impractical. From a theoretical perspective, not all documents are equally as

useful for establishing the subject's interests. For instance, many subjects viewed the Google homepage. Unless customized, this page is identical for every user and provides no content that could potentially contribute to understanding the subject's interests.

To address these issues, a content-based classification of web page types was created based on a manual examination of 2,000 web documents collected during a pilot study [8] and based on a previous classification of web documents [7]. The purpose of this classification scheme was to systematically identify and eliminate pages such as ads, search pages, email pages, pornography, etc. In addition, documents of all types not written in the English language were excluded and subjects were shown documents that they viewed on multiple occasions during the week only once if the content of that document did not change. In cases where more than 150 documents remained after the application of these rules, 150 documents were selected randomly and shown to the user.

2.3 Procedures

The initial meeting with subjects, as well as all additional meetings, occurred at the office of the investigator. At the start of the study, subjects were informed of the conditions of participation and signed a consent form. Subjects then completed the Entry Questionnaire and the first Task Questionnaire. They were then issued a new laptop and printer. A regularly scheduled time and day for the weekly evaluation session was identified at the end of the meeting, or via email on a following day. Subjects were instructed to bring their laptops to each weekly session.

At the start of the weekly evaluation sessions, subjects were presented with a list of their previously identified tasks, and asked to update the list. They were given the opportunity to add new tasks, and then began evaluating the documents that they viewed during the previous week. During the evaluation sessions, subjects were asked to evaluate as many documents as possible. Weekly evaluation sessions were designed not to exceed one hour.

3. RESULTS

An overall description of the data collected during this study is displayed in Table 1. For each subject, this table displays the number of documents requested (**Req**.), number of documents evaluated (**Eval**.), number of tasks identified (**Tasks**), and the means (and standard deviations) for usefulness (**Useful**.) and confidence (**Confid**.). The difference in the number of documents requested and the number evaluated is primarily the result of the classification scheme that was used to identify candidate documents for evaluation. Subjects were, in general, highly confident in the ratings that they assigned to documents.

Table 1. Overall description of the data collected

		Req.	Eval.	Tasks	Useful.	Confid.
Subject	**1**	15,499	843	6	4.88 (1.65)	5.67 (0.66)
	2	5,319	771	11	5.99 (2.00)	6.84 (0.54)
	3	3,157	381	19	5.54 (2.20)	6.78 (0.59)
	4	3,205	351	25	6.05 (0.80)	5.78 (0.79)
	5	3,404	193	12	5.17 (2.40)	6.71 (0.94)
	6	14,586	1,181	21	4.62 (0.80)	5.21 (0.41)
	7	11,657	1,125	33	5.14 (2.40)	6.89 (0.42)

3.1 Tasks

Subjects identified a range of tasks during the fourteen weeks that they participated in this study. The tasks identified by each subject are displayed in Appendix A. Next to each task is a number that is used in a subsequent figure (Figure 5) to refer to the specific task. Tasks are presented in the chronological order in which subjects identified them; thus, subjects identified tasks with smaller numbers before tasks with larger numbers.

3.2 Display Time

One goal of this study was to compute display time using data collected at the client, rather than the proxy. However, despite an extensive, six-week pilot testing effort [8], the client-side logger, nevertheless, malfunctioned on several occasions, which resulted in its automatically shutting-down. One drawback of naturalistic studies is an inability to control and anticipate all possible events; this study was no exception. Because of logger malfunction, display time often had to be computed from the proxy. The extent of this problem is illustrated in Figure 2. This figure shows for each subject, how much of the overall display time data came from the client and how much came from the proxy. For some subjects, such as Subject 1, Subject 5, and Subject 7, over half of the display time data had to be computed from the proxy logs.

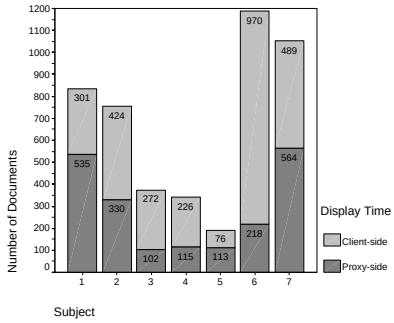

Figure 2. Origin of display time

A Pearson's correlation was computed between the proxy- and client-generated display times in cases where both types of data were available (n=2,776). The results of this analysis found no statistically significant relationship between the two types of display times, r =.013, p=.497. This suggests that proxy- and client-generated display times should not be pooled, and that proxy-generated display times, at least as computed in this study, are not valid substitutes for client-generated display times. Thus, all analyses reported in this paper involve the use of client display times, which comprised 58% of the all display time data.

The distribution of client-side display times for all subjects combined was skewed, with a majority of points being described by less than ten seconds and the distance between points increasing as display time increased. The shape of this distribution is similar to that observed in other studies of display time [12, 13, 15]. To adjust this distribution, a logarithm transformation was performed using the natural log. The log distribution of display time is displayed in Figure 3.

The distribution of display time according to usefulness score is displayed in Figure 4. While no general relationship between display time and usefulness is apparent from the Figure, it is clear that display times differ for subjects. For instance, Subject 7

consistently displayed documents for a longer period of time than other subjects, while Subject 6 consistently displayed documents for a shorter period of time than other subjects. While there was some crossover amongst display times for the other five subjects, the figure demonstrates no clear relationship between display time and usefulness. One-way ANOVAs were conducted to explore the relationship between display time and usefulness for each subject. A statistically significant relationship between mean display time and usefulness was only found for Subject 7, $F(6,485)=6.35^*$; $p<0.01$. A Scheffe post-hoc test was conducted to evaluate pair-wise differences between display time and usefulness for Subject 7. The Scheffe post-hoc test is a conservative test that allows for pair-wise comparisons when there are unequal cell sizes. The result of this test indicated two statistically significant pair-wise differences between usefulness 1 and 6, and usefulness 1 and 7.

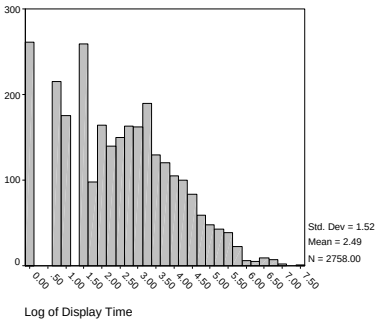

Figure 3. Distribution of display time

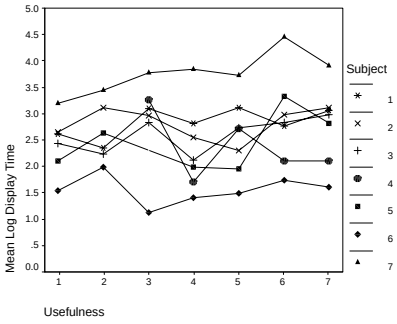

Figure 4. Mean log display time according to usefulness score

3.3 Display Time and Task

The mean log display time according to task for each subject is shown in Figure 5. For cases with extreme standard deviations, and thus, extreme variance bars, the axis is truncated to better highlight the differences in display time (e.g. Subject 3). In the majority of cases, these large standard deviations occurred when a subject associated only two documents with a particular task. For tasks with a greater number of documents, the variances were much smaller, and thus, the bars are shorter. A few tasks had a single document associated with them. For these cases, no variance bars appear on the figure (e.g. Subject 2).

One-way ANOVAs were conducted to explore the relationship between display time and task for each subject. The results are displayed in Table 2. They demonstrate a statistically significant difference in display time according to task for five of the seven subjects.

Scheffe follow-up tests were conducted to evaluate pair-wise differences between the significant relationships. Since pair-wise comparisons cannot be performed with the Scheffe post-hoc test if one or more cells has only a single data point, tasks with only one document were excluded from the post-hoc analysis. The Scheffe tests identified only two statistically significant differences. Both statistically significant differences demonstrated that documents associated with "reading news" were displayed significantly longer than those associated with "shopping" (Subject 2) or those associated with "job searching" (Subject 6). In many cases, the Scheffe test was not significant, indicating that while the F demonstrated at least one mean difference, there was not enough data or strong enough evidence to identify pair-wise differences. This is likely due to the uneven distribution of documents across tasks, and the conservatism of the Scheffe test.

Table 2. Display time and task ANOVA results (*p<.01, NS: Not Significant)

Subject	**1**	$F(5,309) = 2.98^*$
	2	$F(9,424) = 5.20^*$
	3	NS: $F(15,272) = 1.19$
	4	$F(17,226) = 2.28^*$
	5	NS: $F(7,76) = 1.16$
	6	$F(17,969) = 4.14^*$
	7	$F(28,489) = 3.54^*$

Finally, analyses were performed to understand the potential interaction between task, usefulness and display time. The previous analysis demonstrated little direct relationship between display time and usefulness. Results of the ANOVA revealed no statistically significant interaction effects between task, usefulness and display time for any subject.

4. DISCUSSION

Overall, there was great variation between subjects in the relationship of display time and usefulness rating. This, combined with the lack of any significant relationship between these two factors, indicates that using display times averaged over a group of users as a measure of usefulness is unlikely to work. In addition, using mean display time for a single user without taking account of contextual factors is also unlikely to work well.

Within subjects there were large differences in display time according to task, which suggests that task might be one such factor. Statistically significant differences were found between mean display times according to task for five subjects. Although many of the post-hoc tests were inconclusive, documents associated with "reading news" were displayed significantly longer than those associated with "shopping" (Subject 2) or those associated with "job searching" (Subject 6). An examination of mean display times for various tasks (Figure 5) clearly distinguishes some tasks from others, most notably that display times were usually longer for academic-related tasks. However, these pair-wise differences were not statistically significant as evidenced in the post-hoc tests. Additionally, it is difficult to make any generalizations about which types of tasks have higher mean display times because of the low number of subjects in this study and the subject-specific method for identifying tasks.

Kelly, D. & Belkin, N.J. (2004). Display time as implicit feedback: Understanding task effects. In *Proceedings of the 27th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '04)*, Sheffield, UK, 377-384.

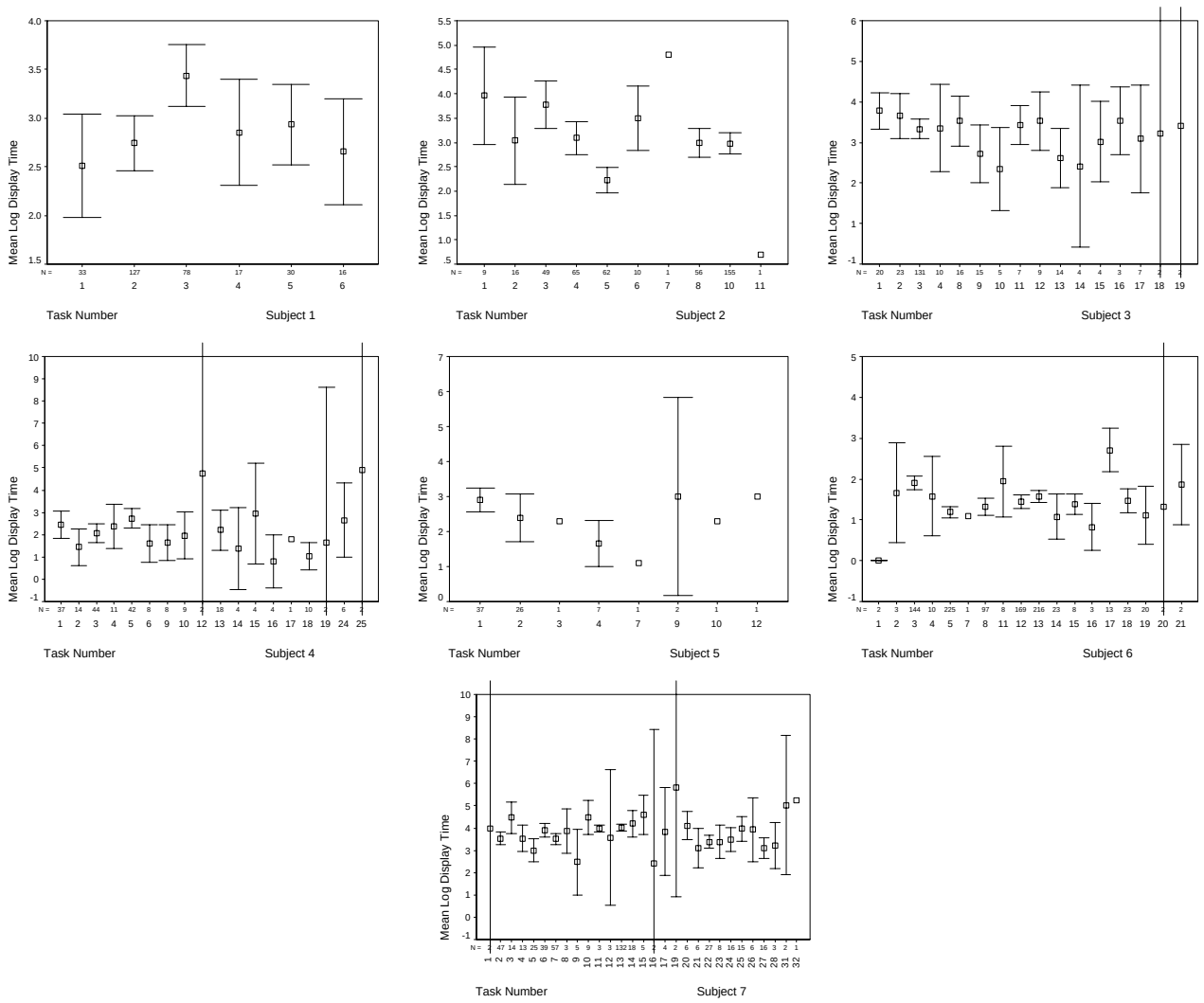

Figure 5. Mean display time according to task

Subjects exhibited varying amounts of online information-seeking behavior. For instance, Subject 6 requested and evaluated almost six times as many documents as Subject 5, and Subject 7 displayed documents longer than any other subject. Furthermore, the number and kinds of tasks identified by each subject differed. Subject 1 identified a low number of very broadly defined tasks (6), while Subject 7 identified a high number of more specifically defined tasks (33). Some tasks, such as reading the news, shopping, travel and working on academic projects, were identified by almost all subjects, while other tasks were only identified by a single subject. It is clear that when studying online information-seeking behavior in natural environments, it is important to normalize and consider the large individual differences that are likely to occur. These differences indicate that a user modeling system might not be equally useful for all people, since some people are heavier users than others, and that such systems should base techniques for modeling on the behaviors of individual users rather than average users.

No statistically significant interaction effects between task, usefulness and display time were found for any subject. This lack of significance is likely the result of the lack of a main effect for usefulness. This result may also be due to the number of possible combinations of levels of each variable in the two-way ANOVAs, and the distribution of data points across these combinations. For instance, Subject 7 identified 33 tasks, when viewed in combination with usefulness, the result is a 33 X 7 matrix. While many of these cells contained display time data, many cells were empty or had fewer than five cases. Grouping usefulness responses, or using a different technique to measure usefulness might lead to different results. Grouping tasks, which we plan to do in the future, might also lead to more generalizable results, and a better understanding of the range of tasks that users are trying to accomplish in online environments.

Finally, as a result of the client-side logger failure, we found in this study that proxy-generated display times were neither valid nor reliable substitutes for client-generated display times. This result demonstrates that the integrity of behavior-based metrics used in studies of implicit feedback is an important research issue. More work needs to be conducted on understanding the relationship between various metrics used as implicit feedback and the behaviors they are meant to measure, and the techniques used for collecting and computing these metrics.

Kelly, D. & Belkin, N.J. (2004). Display time as implicit feedback: Understanding task effects. In *Proceedings of the 27th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '04)*, Sheffield, UK, 377-384.

5. CONCLUSIONS

Despite the small number of subjects in this study, several conclusions can be drawn from it. Foremost among them is that much more care needs to be taken when attempting to use display time as implicit feedback for document preference than has generally been the case. In particular, the results indicate that users may differ substantially from one another with respect to display time behavior, and that any single user's display time behavior may differ substantially according to the task leading to the information-seeking behavior. Also, the lack of significant relationships between display time and usefulness judgments for the subjects in this study suggests that using the simple, unmediated display time as a measure of document preference is likely to fail. Furthermore, just how display time is determined is likely to influence strongly its utility; simple, easily collected proxy-side data may well give misleading results.

Studies based exclusively on proxy-generated data are attractive since a great deal of data can be collected in a relatively short period of time. However, studies employing intensive data collection efforts with fewer users can contribute much to our understanding of information-seeking activities and how, and when, they can be used as implicit feedback. Future studies will continue to explore how users, tasks and behaviors can be studied in natural environments. A study exploring fewer tasks or task groups, is likely to reveal more precise and reliable display time differences according to task. Future studies will also investigate how tasks can be identified using a bottom-up approach, rather than a top-down approach, and how other contextual variables, such as topic familiarity, impact the effectiveness of behaviors as implicit measures of document preference.

6. ACKNOWLEDGEMENTS

We would like to thank our seven study subjects and Fernando Diaz, Haizheng Zhang, Xiaojun Yuan and Gaurav Sood for their technical assistance. This work was supported by NSF Grant #99-11942 and Beta Phi Mu. Opinions, findings and conclusions are our own and do not necessarily reflect those of our sponsor.

7. REFERENCES

[1] Borlund, P. & Ingwersen, P. (2000). The development of a method for evaluating interactive information retrieval systems. *Journal of Documentation, 53*(3), 225-250.

[2] Bystrom, K. (2002). Information and information sources in tasks of varying complexity. *Journal of the American Society for Information Science and Technology, 53*, 581-591.

[3] Bystrom, K., & Jarvelin, K. (1995). Task complexity affects information seeking and use. *Information Processing & Management, 31*(2), 191-213.

[4] Claypool, M., Le, P., Waseda, M., & Brown, D. (2001). Implicit interest indicators. *Proceedings of 6th International Conference on Intelligent User Interfaces*, 33-40.

[5] Cooper, M. D., & Chen, H.-M. (2001). Predicting the relevance of a library catalog search. *Journal of the American Society for Information Science, 52*(10), 813-827.

[6] Cooper, W. S. (1973). On selecting a measure of retrieval effectiveness, part I: The "subjective" philosophy of evaluation. *Journal of the American Society for Information Science, 24*, 87-100.

[7] Haas, S. W., & Grams, E. S. (2000). Readers, authors and page structure: A discussion of four questions arising from a content analysis of web pages. *Journal of the American Society for Information Science, 51*(2), 181-192.

[8] Kelly, D., & Belkin, N. J. (2002). A longitudinal, naturalistic study of reading behavior as implicit feedback: Preliminary findings. (Technical Report #02-100), Rutgers University, SCILS Information Interaction Lab.

[9] Kelly, D., & Teevan, J. (2003). Implicit feedback for inferring user preference: A bibliography. *SIGIR Forum 37*(2), 18-28.

[10] Kim, J., Oard, D. W., & Romanik, K. (2000). Using implicit feedback for user modeling in internet and intranet searching. University of Maryland CLIS Technical Report.

[11] McDonnell, J., Koehler, W., & Carroll, B. (2000). Cataloging challenges in an area studies virtual library catalog (ASVLC): Results of a case study. *Journal of Internet Cataloging, 2*(2), 15-42.

[12] Miller, B. N., Riedl, J. T., & Konstan, J. A. (2003). GroupLens for Usenet: Experiences in applying collaborative filtering to a social information system. In C. Lueg & D. Fisher (Eds.) *From Usenet to CoWebs: Interacting With Social Information Spaces* (pp. 206-231), London: Springer Press.

[13] Morita, M., & Shinoda, Y. (1994). Information filtering based on user behavior analysis and best match text retrieval. In *Proceedings of SIGIR '94*, 272-281.

[14] Oard, D. W., & Kim, J. (2001). Modeling information content using observable behavior. *Proceedings of the 64th Annual ASIST Conference*, 38-45.

[15] Rafter, R., & Smyth, B. (2001). Passive profiling from server logs in an online recruitment environment. In *Proceedings of the IJCAI Workshop on Intelligent Techniques for Web Personalization*, 35-41.

[16] Tang, R., Shaw, M., & Vevea, J. L. (1999). Towards the identification of the optimal number of relevance categories. *Journal of the American Society for Information Science, 50*(3), 254-264.

[17] Vakkari, P. (1999). Task complexity, problem structure and information actions: Integrating studies on information seeking and retrieval. *Information Processing and Management, 35*, 819-837.

[18] Vakkari, P. (2001). A theory of the task-based information retrieval. *Journal of Documentation, 57(1)*, 44-60.

[19] Vakkari, P. (2003). Task-based information searching. *Annual Review of Information Science and Technology, 37*, 413-464.

[20] White, R. W., Ruthven, I., & Jose, J. M. (2002). Finding relevant documents using top ranking sentences: An evaluation of two alternative schemes. In *Proceedings of SIGIR '02*, 57-64.

[21] Wilson, P. (1973). Situational relevance. *Information Storage and Retrieval, 9*, 457-469.

Kelly, D. & Belkin, N.J. (2004). Display time as implicit feedback: Understanding task effects. In *Proceedings of the 27th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '04)*, Sheffield, UK, 377-384.

APPENDIX A: Tasks

Subject 1	Research Dissertation (1), Shopping (2), Read News (3), Movie Reviews & Schedules (4), Preparing Course Material (5), Entertainment (6)
Subject 2	Get insurance quotes (1), Get course materials (2), Read the news online (3), Academic research (4), Shopping (5), Check weather (6), Hobbies (7), General interests (8), Teaching (9), Get information (10), Check email (11)
Subject 3	Studying Quals (1), Scheduling resources (2), General knowledge (3), Entertainment (4), Research paper – interests groups (5), Research paper – intellectual property (6), News (7), Shopping (8), Person search (9), Translation (10), Weather (11), Amherst Alumni Activities (12), Political activism (13), Funding (14), Teaching stuff (15), Pet search assistance (16), Computer maintenance (17), Student government work (18), Research – general (19)
Subject 4	Searching and browsing for computer interests (1), Searching for project – brain images (2), Online shopping (3), Viewing news (4), Writing an academic paper (5), Checking the weather (6), Looking up directions (7), Searching health information (8), Searching auto repair information (9), Browsing friends' homepages (10), Checking email online (11), Language studying (12), Checking university web pages (13), Sports searching (14), Searching for employment (15), Install software (16), Find books in university library (17), Find companies' information (18), Check what's on tv (19), Check my own homepage (20), Check my lab web page (21), To know holidays (22), Check university network utilization (23), Learn python language (24), Learn to use SIP (25)
Subject 5	Travel (1), Applying for fellowships, grants and awards (2), Submitting papers to conferences (3), Staying in touch with people (4), Shopping for material possessions (5), Writing dissertation (6), Legal trouble/conflict (7), Weather (8), Development as a scholar (9), Teaching instruction (10), Additional teaching gigs (11), Housing options (12)
Subject 6	Dissertation (1), Publication (2), News (3), Reading (4), Job searching (5), Registration (6), Household method (7), Shopping (8), Booking tickets (9), Music – downloads (10), Download software (11), Project in operating systems (12), Research (13), Database systems (14), Compiler (15), Change new I-20 (16), Look up directory (17), Review maps online (18), Recreation (19), English learning (20), Registration (21)
Subject 7	Selling on Ebay (1), Shopping for books for oral exam (2), Writing reviews for amazon (3), Printing chords and tablatures (4), Setting up personnel website (5), Running fantasy soccer league (6), Research for dissertation (7), Download updates for digital recording studio (8), Reviewing for journal (9), Purchase mandolin (10), Research for Exit9 (11), Record music (12), Check news (13), Shopping (14), Correct Greek homework (15), Purchase guitar (16), Check on flights (17), Purchase theater tickets (18), Get address [1] (19), Consult teaching resources (20), Follow conference (21), Compile reading packet (22), Check on jobs (23), Locate a friend (24), Find housing (25), Shop for Profs for exam (26), Rent movies (27), Check on parking (28), Check transcripts (29), Get address [2] (30), Research Napster (31), Check train schedule (32), Purchase concert tickets (33)

Kelly, D. & Belkin, N.J. (2004). Display time as implicit feedback: Understanding task effects. In *Proceedings of the 27th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '04)*, Sheffield, UK, 377-384.