Chapter 6  
Discussion

This chapter discusses the major findings of the study. This chapter begins with a discussion how information-seeking context was measured and what the results of this measurement were. Each research question is then presented along with the major findings that addressed the question. Potential explanations for the findings are discussed in each section, along with their implications. This chapter concludes with a discussion of the relationship of the findings to the theoretical model proposed in Chapter 3.

6.1 Measuring Information-seeking Context

This study was novel in that it was one of the first naturalistic studies of the online information-seeking behavior of users that attempted to capture aspects of the user’s information-seeking context, including specific task and topic, and to relate these aspects to the pages that the user viewed and the user’s evaluation of those pages. Because of this, little guidance for the design of the method for this study was available. This section discusses the method employed in this study for identifying and measuring information-seeking context. This method is a novel and potentially useful one for the study of information-seeking context in online environments.

During this study, subjects were asked to self-identify specific tasks in which they would be working and specific topics in which they were interested. Subjects did this at weekly intervals throughout the course of the study and, in general, indicated that they had few problems doing so. Subjects identified a range of tasks; the number and kinds varied considerably from subject to subject. Some tasks, such as academic or dissertation

research, checking the news, shopping and entertainment were identified by almost all subjects. Other tasks, such as “get insurance quotes,” “learn python language,” “political activism,” and “legal trouble/conflict” were subject-specific. Thus, while it was possible to identify some tasks which all subjects did during the course of the study, a number of tasks were user and situation specific.

The tasks that subjects identified were at many different levels of abstraction. Many subjects identified a general interests task to classify documents which were not related to any specific task. Some tasks were identified on a very specific level, with various temporal instances of the task counting as separate unique tasks or various topic-related instances counting as separate unique tasks. Results from the Exit Interviews confirmed the difficulty subjects had with determining how broad or specific to be in creating tasks. Further problems that subjects had were distinguishing between various sub-tasks of a particular task and relating these to one another. These results indicate that perhaps subjects would benefit from more guidance on how to construct their task classes. For instance, one could provide subjects with generic task classes, such as those identified by almost all of the subjects in this study, and also allow for the identification of subject-specific tasks.

The topics identified by subjects were more user-specific and esoteric than the tasks, and, for most subjects, out-numbered tasks. An examination of the number and kinds of topics identified by subjects and comments from the Exit Interview revealed that subjects had an even more difficult time determining an appropriate level of specificity with which to identify topics. One subject identified eighty unique topics during the course of the study! While the variety and number of topics identified by subjects made

for rich and interesting data about daily information-seeking activities, it made quantitative analysis difficult and laborious and severely limited the amount of post-hoc analysis that could be performed. Because only a single document was associated with some topics, these topics had to be excluded from the post-hoc analysis. Analysis with such small cell sizes is precarious at best.

These results have methodological implications for how studies of users’ real tasks and topics ought to be conducted. An effective methodological decision was allowing subjects to identify tasks and topics while they evaluated documents, since the majority of new tasks and topics were identified during these evaluations. Thus, previously seen documents reminded subjects of their previous activities and helped them to identify new tasks and topics. One subject even recalled the time of day and what he was eating for dinner when he viewed a particular document. These memory triggers effectively increased the numbers and kinds of tasks and topics subjects identified; without them, fewer tasks and topics would have likely been identified by subjects, which might have decreased the ecological validity of the study. Moreover, subjects commented that over time they learned more about how to better identify a task and topic, indicating studies which investigate a user’s real tasks and topics should be conducted over time. Finally, when studying users’ real tasks and topics, some balance between subject-specific and generic methods of data collection needs to be achieved to allow for meaningful analysis.

For each week of the study, subjects were asked to characterize each task that they identified according to endurance, frequency and stage, and each topic that they identified according to persistence and familiarity. Overall, the endurance of tasks

remained relatively constant over time and most tasks were identified as having an endurance of several years. Tasks that were common to almost all subjects, like shopping, reading the news and academic research, often had high endurance ratings. Tasks that had low endurance ratings were most often those that were subject-specific. Tasks that had endurance ratings that changed over time were often those that had a fixed and identifiable completion date; these types of tasks were also usually associated with low endurance ratings.

In general, subjects did not have a difficult time determining task endurance. However, for some tasks, subjects were unable to accurately predict how long they expected to be working on the task. In addition, some problems with assigning endurance ratings were introduced by the level of specificity with which subjects identified tasks. For instance, confusion resulted over the task of teaching, as it related to a specific instance (i.e. the class the subject was currently teaching) and as it related to a general, on-going task that would be repeated numerous times over the course of many years. Even though subjects identified some problems assessing endurance, overall, the results of this study indicate that the length of time a person expects to be engaged with a task is a measurable entity of information-seeking context.

Subjects used a range of frequencies to describe their task-specific online information-seeking activities. However, the rate with which these frequencies changed over time differed. Frequency ratings did not change that often for three subjects, but for four subjects frequency ratings changed multiple times during the course of the study for many different tasks. Results from the Exit Interviews indicated that frequency fluctuated depending on the immediacy of the current task and other tasks. Subjects also

noted that their perceptions and expectations of how often they conducted information-seeking activities for certain tasks often did not match reality. In general, subjects had a tendency to overestimate how often they did tasks like academic research and underestimate how often they did tasks like shopping and entertainment, which should come as no surprise to anyone who has ever done academic research. Subjects further described many of their activities as opportunistic and intense, rather than planned and continuous. These results indicate that frequency as an aspect of information-seeking context, should, at a minimum, be measured once a week and that perhaps a different type of scale would be useful for more accurately capturing frequency of activities. These results further indicate that to get the most accurate measure of frequency one should rely not only on self-report data, but also on objective measurements, such as those that can be collected from logs.

Stage was assessed by subjects only for those tasks for which they believed it was applicable. Subjects indicated that stage was not applicable for most of their tasks. Tasks for which stage was rated were usually those tasks which had a specific date of completion, such as “get course materials,” “studying quals,” and “project in operating systems.” Subjects used a range of stage levels to describe these tasks, but stage ratings did not change all that much over time for any given task and, interestingly, when they did, it was not always in a forward direction. Subjects considered stage to be “not applicable” for most of the tasks that they worked on during this study. For tasks which cannot be described by stages, it might be a potentially useful idea to consider the length of time the subject has been working on the task instead, to better understand the subject’s progress with the task.

Overall, there were relatively few tasks that were completed during the course of this study. Given that subjects indicated that most of their tasks had a high endurance, this is unsurprising. As with endurance and frequency, some problems were caused by the level of specificity with which subjects originally identified tasks. The results of this study indicate that stage, as a measure of information-seeking context, is only appropriate for a small number of those tasks which are conducted online. It might be more appropriate to consider “phases of activity” for online tasks rather than stages, since phases do not imply sequential order or that the task will necessarily ever be completed. However, measuring phases would require more than a generic numerical scale, such as the kind used in this study.

Persistence, like endurance, did not change all that often. Subjects described most topics as persisting for several years. Similar to endurance and frequency, subjects’ estimated topic persistence did not always correspond to the actual number of weeks that the topic was active. Topics whose persistence did change over time were usually topics associated with a specific deadline. However, in the Exit Interviews, one subject stated that topic persistence was more fleeting and volatile than task endurance, especially for those short-term topics that were associated with tasks which did not have a specific date of completion. Overall, this study demonstrates that topic persistence is a measurable entity, even though it did not vary that much during the course of the study.

The final aspect of information-seeking context which this study measured was topic familiarity. Subjects used a wide range of scores to describe their familiarity with topics, although for two subjects familiarity was somewhat constant throughout the course of this study. For the remaining five subjects, both increases and decreases in

familiarity were observed. Subjects explained that these changes were the result of learning more about a topic. Learning more about a topic was also given as a reason for stable familiarity ratings with topics whose content changed rapidly, such as news, politics and sports. In other words, subjects felt that they needed to be constantly learning more about some tasks just to maintain a certain level of familiarity. These results demonstrate the complexity of familiarity and suggest that more work needs to be performed on developing techniques for measuring and capturing this complexity. Perhaps familiarity measures should be topic specific; for example, familiarity with topics whose content rapidly changes should be measured differently from familiarity with other types of topics. Alternatively, one might measure interest in a topic instead of familiarity. Overall, the results with respect to familiarity are encouraging, since of the five attributes, familiarity changed the most and subjects used the largest range of responses to describe this attribute.

6.2 Types of Behaviors that Can Be Gathered

The first research question of this dissertation was concerned with discovering what types of information about a user’s document preference could be gathered implicitly through monitoring the user’s online information-seeking and use behaviors and examining the results of that behavior over time. This question was addressed through the literature review, which identified and assessed the various types of behaviors that had been used in previous research as implicit indicators of interest, and identified the various techniques that have been employed for collecting and using these behaviors as implicit indicators of interest. This question was further addressed by the

results of this study which described the distributions and occurrences of display time and retention, and the differences between client-generated display times and proxy-generated display times.

A review of the literature demonstrated that display time, selection and linking have been the most frequently used behaviors for implicit feedback. Likely reasons for this are that these behaviors are available for every page requested, that they can be collected in relatively unobtrusive ways, and that they are available in large quantities. Linking behavior, in the form of web link analysis, is perhaps the most successful large-scale use of implicit feedback, as seen in the commercial system Google. Selection has also been used with some success for certain types of tasks, as evidenced in its success as a model for generating online advertising revenues. However, for personalizing information-seeking activities, selection can be considered as rather weak evidence of interest since every page that a user selects is considered equally useful. Research examining the effectiveness of display time as an implicit measure of interest has had mixed results, although generally favorable. However, display time as evidence faces severe measurement problems, especially when gathered by a proxy logger, since these measurements are not necessarily valid or reliable. Overall, implicit feedback research has been further limited because it has not take into account the potential effects of information-seeking context on the behaviors that it studies.

Specific results that addressed this research question included those that described the differences in client- and proxy-generated display times, and those that described the distributions and occurrences of display time and retention. Recall that no correlation was found between client- and proxy-generated display times. These results suggest that

proxy-generated display times are neither valid nor reliable substitutes for client-generated display times. Typical studies of implicit feedback have spent little, if any, time exploring the validity and reliability of the metrics that they employee, and have instead assumed that (1) the metrics represent the actual behavior of interest (e.g. the length of time a document is displayed in the browser window is equivalent to how long the user reads the document; and that (2) the measurement technique is valid and reliable (e.g. display time collected at the proxy represents the actual time the document is displayed in the user’s browser window, and this measurement is consistent regardless of document type or network speed). Moreover, while these studies often report where the metrics were collected (proxy or client), they do not report how the metrics were computed. For instance, since display time is not directly available at the proxy, it is usually computed from other available information, such as the difference between successive URL requests. In this study, this technique proved problematic because many documents contained multiple URLs, such as advertisements or frames, and resulting display times, as computed, were only one or two seconds long. Depending on how others compute display times from proxy data, this could pose a significant measurement problem. Furthermore, in this study, display times for documents that were displayed multiple times within a single week were cumulated to arrive at a week-based display time. In cases where proxy time had to be used instead of client time, display times for documents displayed more than once could not be cumulated.

The distributions of client- and proxy-generated display times indicate further differences between the two types of times. Both distributions were similar to those observed in other studies of display time as implicit feedback in that they were skewed,

with a majority of documents being displayed for a few seconds and the distance between points increasing as display time increased. However, the distribution of proxy-generated display times had a more extreme skew, with a large number of documents described by a few seconds, and then a huge drop in display times. Conversely, the distribution of client-generated display times exhibited a more steady decline as display times increased. Another difference between these distributions and those observed in other studies that use proxy-generated display times was the maximum display times observed. When display times are collected from proxies, there are typically numerous display times that exceed sixty minutes, and many even exceed several hours. However, in this study, the maximum display times for both client- and proxy-generated data were a little over 30 minutes. Overall, these results suggest that not only are extreme maximum display times a problem with data collected via proxy-loggers, but that minimum display times also introduce errors.

While display times of one second dominated the proxy-generated data, they were also prevalent in the client-generated data, describing approximately 10% of all display times. Display times of one second are interesting cases because it was unclear what subjects are doing during these single second interactions. While this might suggest the potential for using these documents as negative feedback, many of these documents were rated as useful by subjects. Although a more detailed study of these times needs to be conducted to identify what is occurring during these one second interactions, there are several possible explanations for these display times. It is possible to imagine potential one second interactions, for example, if the subject is checking to see if new information has been posted to a page. Some studies have assumed that display times of less than one

second are not meaningful and have eliminated these from data analysis, but the subjects of this study had few problems identifying the pages that were shown to them for evaluation, including those with display times of one second. This suggests that these display times are meaningful and warrant a separate study to understand what is occurring during these very quick interactions.

The retention behaviors that were of interest in this study included saving, printing and bookmarking. Retention behaviors occurred less frequently than expected, which suggests that their value as implicit evidence of document preference is limited. Bookmarking occurred the most frequently of all retention measures; subjects bookmarked anywhere from 0.3% to 4.5% of the pages that they evaluated. Saving occurred less frequently, with subjects saving anywhere from 0.3% of the documents that they evaluated to 1.4%. There were no instances of subjects printing documents that were evaluated, although there were some instances of subjects printing documents from word processing applications. Subjects indicated that they typically printed materials from other locations since it was more economical for them. The low observation of printing behavior is in contrast to studies of the past which have found that online users often print items before they review them in their entirety. Although very tentative, this result may indicate a general trend in the reduction of printing of online materials and an increase in the online reading of documents.

It should be noted that there were many items that were bookmarked but not included in the results of this study because these items were not selected for evaluation. These types of pages included home pages of online news sources like CNN and the New York Times, and search services like Google and Yahoo. It should be further noted that

the two subjects who saved the most documents were both engineers. Most of the pages that were saved locally were postscript or portable document format (pdf). Because these types of pages usually have to be saved locally to be viewed, the relationship between saving and usefulness is tenuous, since saving had to occur before usefulness could be determined.

6.3 Relationship between Behaviors and a User’s Document Preference

The second research question of this dissertation was concerned with the relationship between the observable behaviors of display time and retention and a user’s document preference. Recall that document preference was measured by subject-assigned usefulness scores. This section begins with a discussion of the distribution of usefulness ratings assigned to documents by subjects. This is followed by a discussion of the relationship between display time and usefulness, and retention and usefulness.

6.3.1 Usefulness Ratings

Overall, usefulness scores were quite high for all subjects. Subject 2 used seven to describe almost three-fourths of the documents that she evaluated, while Subject 7 used seven to describe almost half of the documents that he evaluated. Subject 6 used five to describe a large proportion of the documents that he viewed, reserving six and seven for a small percentage of documents. Given that subjects’ mean confidence in the scores that they assigned to documents was high, this result appears to be in contrast with what many other studies report regarding satisfaction with search results on the web. However, it should be noted that subjects were not evaluating documents with respect to

how relevant they were to a particular query issued during a particular search, but rather, were evaluating how useful they found the document for a particular task and topic. Tasks included those that required both searching and browsing, with the majority of cases being characterized by browsing rather than direct searching. Because the tasks that subjects identified went beyond just basic searching tasks, the concept of usefulness as applied by subjects in this study was very liberal and subjective. For instance, many documents viewed by subjects were done so for entertainment purposes and it is difficult to imagine that a great deal of these types of pages were not rated as useful. This is encouraging as it suggests that the web fulfills a number of needs in a number of different ways. This further suggests that multiple measures of document preference, as opposed to single measures, might be more appropriate for web studies.

6.3.2 Display Time and Usefulness

Overall, the results of this study demonstrate that display time, by itself, is not a good indicator of usefulness, and that mean display times differ for each subject. Mean display time was only found to differ significantly according to usefulness for one subject. Post-hoc tests detected a significant difference between mean display time for those documents rated as 1 and those rated as 6 or 7. The finding from this post-hoc test was similar to the findings of previous research which have found that documents rated as more useful have higher display times than those documents rated as less useful. However, contrary to the findings of others, display time was, in general, not a good indicator of usefulness.

Several explanations can be offered for this finding. First, the methodological approach of most studies has been to average results over a number of users; this averaging may result in conflating the significance of the findings, especially if the finding is significant for only a small number of users, who also happen to be the users who viewed the most pages. Consider the results of this experiment, if results were averaged across all subjects and the relationship between display time and usefulness was significant for Subjects 6 and 7, who evaluated almost five times as many documents as one of the other subjects did, then it is likely that display time and usefulness would appear to be significantly related for all subjects. Indeed, in an examination of the relationship between display time and usefulness for thirty-six subjects, Kelly and Belkin (2001a) found this to be the case; the relationship between display time and usefulness was significant when averaged over all subjects, but when examined on an individual basis, it was only significant for one-third of the subjects. Miller, Riedl, and Konstan (2003) also found similar results. The current study provides continued support for the findings of these previous studies, in that individual differences seem to be important, and that a user modeling system that uses behavior to infer document preference should do so on an individual basis. Furthermore, the distribution of mean display times according to usefulness score in the present study demonstrated a number of differences in display times between subjects.

Another difference in this study and previous studies concerned the types of tasks and searching environments that were examined, and the technique for measuring usefulness. Many previous studies have limited their investigations to a single task, such as news or job searching, and searching environments have varied from Usenet news to

closed, off-line collections of documents, such as Dialog. The exception to this is Claypool, et. al (2000), whose users were instructed to do unstructured browsing on the web for twenty minutes. None of these studies collected as much detailed, long-term and user-specific data as this study did. Further, this study collected data about numerous tasks and topics from a variety of domains. These differences may have contributed to the difference in results between this study and others. However, recall that previous studies of display time as an implicit measure have had mixed results. Thus, the results of this study should be interpreted as evidence for the complexity of the relationship between display time and usefulness rather than a refutation of its potential. As will be discussed later, information-seeking context appears to have a significant impact on this relationship.

The technique for measuring a user’s document preference might explain why the results of this study differed from previous studies of display time and usefulness. The scales used to measure preference in these studies have been purported to measure a number of different things including, “relevance,” “usefulness,” and “interest.” Some researchers have even asked users to assign grades to the documents that they view. These scales have further ranged from two- to seven-point scales. The technique for measuring a user’s document preference should not be taken lightly, although few researchers of implicit feedback have spent much, if any, time discussing how the techniques used to measure preference might affect the overall results. There is a great deal of research that discusses how sensitive subjects are to varying types of relevance scales and it would be erroneous to assume that these results do not apply to studies of implicit feedback.

Finally, usefulness scores were, in general, skewed, with a majority of the points described by display times of less than ten seconds. These distributions may have prevented significant results from emerging. Additionally, only a small number of data points were available for Subject 5 after the proxy-generated times were excluded, which may have affected analysis.

Analyses of proxy-generated display times and usefulness, and a combination of proxy- and client-generated displays and usefulness were conducted to understand the potential impact of erroneous data on analysis. Results demonstrated significant relationships between display time and usefulness for two subjects when only the proxy-generated display times were used, and no significant relationships when a combination of proxy- and client-generated display times were used. These results indicate the potential danger of Type I and Type II errors when using proxy-generated data alone, or in combination with client-generated data.

6.3.3 Retention and Usefulness

With the exception of Subjects 6 and 7, the mean usefulness of documents that were retained was high for all subjects, but it did not differ significantly from the mean usefulness of documents that were not retained. Differences may not have been detected because the usefulness ratings were, overall, quite high for most subjects. Further, as described earlier, some documents had to be saved before they could be viewed; thus, their usefulness remained unknown until after they were saved and viewed. These results have some interesting implications for those studies of implicit feedback that use retention behavior as an evaluation metric for usefulness. The results from this study

suggest that retention is neither a reliable, nor valid method for evaluating document usefulness.

6.4 Relationship between Behaviors and Information-seeking Context

The third question of this dissertation was concerned with the relationship between display time, retention and information-seeking context. Recall that information-seeking context was measured by specific task and topic, as well as the task attributes of endurance, frequency and stage, and the topic attributes of persistence and familiarity. This section begins with a discussion of how display time and retention varied according to specific task and topic. The relationships between display time and endurance, frequency, stage, persistence and familiarity are then presented. Analysis of the relationship between task and topic attributes and retention was not conducted because of the dearth of data on retention behaviors. While numerous statistically significant relationships were found with respect to display time and information-seeking context, the reader is cautioned that a rather large number of tests were conducted, and that because of this there is a chance of finding at least some statistically significant results by chance.

6.4.1 Display Time and Task and Topic

Display time was found to differ significantly for task for five of the seven subjects and it was found to differ significantly for topic for all but one of the subjects. Post-hoc tests pinpointed some differences between specific tasks, although it was necessary to combine tasks which had only a single document associated with them with

other tasks to do the post-hoc analysis. Tasks that had mean display times that
significantly differed for some subjects were those related to reading the news and
shopping.

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 experiment and the
subject-specific method for identifying and conceptualizing tasks. For two subjects,
results from the post-hoc tests found a statistically significant difference between mean
display time for shopping and reading the news. While an examination of the mean
display times for various tasks clearly distinguishes some tasks from others, most notably
that display times were usually longer for academic-related tasks, the low number of
observations for some tasks made it difficult to find many statistically significant
differences. A controlled study, focusing on a specific set of tasks, would likely identify
more precise differences in mean display time for task.

For most subjects, the relationship between topic and display time was stronger
than for that of task and display time. However, it is more difficult to identify and
discuss meaningful differences in mean display times for topics than for tasks because of
the high number of topics identified by subjects, because many topics were only
associated with a single document, and because topic differences do not make as much
intuitive sense as task differences. Post-hoc tests failed to identify significant pair-wise
differences for most subjects because of a combination of the large number of topics
identified by many subjects and the small cell sizes. Further, topics that had only a single
document associated with them had to be eliminated from the post-hoc analysis. Suffice
to say, display time differs according to topic and it is likely that this confounds the

Kelly, Diane (2004). *Understanding implicit feedback and document preference: A naturalistic user
study*. Ph.D. Dissertation, Rutgers University.
relationship between display time and usefulness. As with tasks, a more focused, controlled study of topic and display time might lead to more detailed results.

Analyses were further conducted to investigate the relationship between proxy-generated display times, and task and topic for each subject. Significant relationships were only identified for a single subject, for both task and topic. Analyses were also conducted with the combination of proxy- and client-generated display times, and task and topic. These results demonstrated significant relationships for four subjects for task, and three subjects for topic. Overall, these results provide more evidence for the potential danger of using unreliable and invalid data during analysis and suggest that care should be taken in understanding, collecting and using behavior-based metrics, especially when they are generated by a machine.

6.4.2 Retention and Task and Topic

Statistical tests exploring the relationship between retention and task and topic were not conducted because there were so few occurrences of retention. Some tasks and topics appeared to have more retained documents associated with them than others. Tasks included those related to academic research and shopping and topics most often included those associated with these two types of tasks. Although subjects were not asked explicitly why they retained documents, it is likely that documents related to each of these types of task were retained more frequently than others because these types of task often require that documents be referred to more than once. In addition, as described earlier, the format of some online documents makes it necessary to download the document before it can be viewed. For the two subjects where this behavior frequently

occurred, the documents were associated with academic-related tasks. While some subjects retained a few news and entertainment documents, overall, these documents were retained less frequently than the academic-related and shopping documents. Clearly, a more focused study on retention behavior should be conducted to find out if the results of this study represent a changing trend in retention frequency, especially printing, or if these results are particular to these seven subjects.

### 6.4.3 Display Time and Task and Topic Attributes

Mean display time was found to significantly differ for all subjects but two, for various task and topic attributes. Stage and persistence were the weakest attributes; mean display time only differed significantly for one subject for each attribute, and these relationships, in general, were not very strong. Familiarity was the strongest attribute; mean display times differed significantly for all subjects but two. Although the variances for these relationships varied from 2% to 9%, the relationship was in a consistent direction for most subjects, with lower levels of familiarity most often associated with higher display times.

Mean display time differed significantly according to endurance for two subjects. Post-hoc tests demonstrated that these differences were in opposite directions, with display time being longer for documents characterized as endurance 7 than 8 for one subject, and display time being longer for documents characterized as endurance 8 than 6 for another subject. Although the direction of these relationships differed, it should be noted that differences were only detected for the higher endurance ratings. Recall that most subjects used the highest endurance score to describe the majority of their tasks, and

that these scores did not change over time. It is likely that these skewed distributions prevented any significant differences from being detected for other subjects, or for other levels of this variable.

Mean display time differed significantly according to frequency for two subjects. Results of the post-hoc tests were in a consistent direction; for both subjects, higher frequencies had lower display times. These results may indicate that differences in display times are likely to occur for tasks that are conducted daily as opposed to those that are conducted with other frequencies. The results seem to support this, although they are somewhat difficult to interpret because of the number of levels of frequency (four) and the low number of observations in some cells. Moreover, most subjects did not use all of the levels of frequency. However, the results of this study demonstrate that frequency of activities affects display time for some subjects, and in a consistent direction.

As previously mentioned, stage was one of the weaker contextual variables, with only a single significant relationship identified for one subject. For this subject, the post-hoc test revealed a variety of differences: display times were significantly less for documents characterized by stage 1, than stage 3 or 5, and display times for documents characterized by stage 3 were significant greater than for those characterized by stage 5. These results seem to suggest, at least for this subject, that documents characterized by mid-range stage scores are displayed longer than documents characterized by scores located near the beginning or end of the scale. However, recall that stage was not used all that much by subjects to characterize their tasks, so in many respects, the overall lack of significant results is unsurprising.

Along with stage, persistence was also one of the weaker contextual variables, with only a single significant relationship identified for one subject. For this subject, the post-hoc test demonstrated that display time was longer for documents characterized as persistence 7 than 8, which is similar to the post-hoc results from endurance and display time. Recall that the distribution of persistence scores were similar to those of endurance, with most subjects using the higher persistence scores to describe the majority of their tasks, and with little change in these scores over time. This lack of change most likely made it difficult to observe any significant differences between persistence scores.

Finally, mean display time according to familiarity was found to differ significantly for all subjects except two. For familiarity, all of the post-hoc tests were statistically significant, and for most subjects, lower levels of familiarity were found to be associated with higher display times. These results indicate that users who are less familiar with a topic are likely to spend more time displaying a document than those users who are more familiar with a topic. These results are encouraging as they demonstrate a significant effect for familiarity across almost all subjects and a general direction for this relationship.

When the results are considered on an individual basis, differences in the strength of the relationships between display time and the various attributes for each subject are apparent, although there is no general trend. For Subject 1, the only significant relationship found was between mean display time and familiarity. For Subject 2, the relationship between mean display time and frequency was significant, as well as mean display time and familiarity. The relationship between mean display time and familiarity was the strongest, even though this relationship was not particularly strong. For Subject

3, no statistically significant relationships were found between mean display time and any of the five attributes. For Subject 4, significant relationships were found between display time and endurance, persistence and familiarity. The strongest of these relationships was that of display time and familiarity. For Subject 5 no statistically significant relationships were found between mean display time and any of the five attributes. Although these results are identical to those of Subject 3, it should be noted that over half of Subject 5’s display time data was excluded from analysis because it came from the proxy. The analyses reported here were only performed with 76 data points, which would make finding significant results difficult given the number of levels of each attribute.

For Subject 6, the relationships between mean display time and all attributes but persistence were statistically significant. In contrast to the number of data points available for Subject 5, a full 969 were available to use in the analyses of Subject 6’s data, which may indicate that a large number of points are necessary to detect significant differences, and that behavior starts to stabilize after a certain number of observations. However, none of the relationships between display time and endurance, frequency, stage and familiarity were particularly strong, with only 2 to 4% of the variance in display time being explained by these factors. Finally, for Subject 7, the only statistically significant relationship found was between display time and familiarity.

In sum, the results of the analysis of the relationship between display time and information-seeking context demonstrate significant differences in mean display time according to task, topic, endurance, frequency, stage, persistence and familiarity. These differences varied in number and strength for each subject, although the differences in display time and topic were statistically significant for six subjects, and display time and

task and familiarity were statistically significant for five subjects. The results further demonstrate that different aspects of information-seeking context are significantly related to display time for different subjects, in different ways. Although many of the results regarding display time and endurance, frequency, stage and persistence were weak, these results should be interpreted as providing support for the potential effects of these contextual variables on display time, rather than as establishing solid relationships. Care should further be taken in interpreting the results, since many data points had to be excluded from analyses. Thus, the data used in these analyses represent an incomplete picture of subjects’ activities during the 14 weeks.

6.5 Predicting Usefulness from Behaviors

The final question of this dissertation was concerned with understanding the relationship between behaviors, information-seeking context and usefulness. The goal in doing so was to identify at least some types of information that would be necessary to effectively use behaviors as implicit evidence of document preference. However, no statistically significant relationships were found between display time, task and usefulness, and display time, topic and usefulness. Nor were statistically significant relationships found between display time, usefulness and each of five task and topic attributes. Analysis of the proxy-generated display times demonstrated significant interaction effects for task and usefulness, and topic and usefulness for many subjects, as well as significant interaction effects for endurance, frequency, stage, persistence and familiarity, and usefulness for a many subjects. These differences in results provide

further support for the possibility of erroneous results, and, in particular, Type I errors, when using proxy-based display times.

While there were significant main effects for task and topic for almost all subjects, and significant main effects for the various attributes for many of the subjects, when combined with usefulness, these contextual variables do not seem to be related to display time. This is likely 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, for endurance and usefulness there were seven levels of each variable, which resulted in 49 possible combinations (7 X 7). While many of these cells contained display time data, the majority of cells were empty or had fewer than five points. Interestingly, while subjects stated in the Exit Interview that they appreciated the large range of responses available to them, confirming the work of Tang, Shaw and Vevea (1999), this large range of responses can make analysis problematic as evidenced in this study. While grouping responses, and various tasks and topics might result in more potent analysis, this will be left for future work.

6.6 Relationship of Findings to the Theoretical Model

This dissertation was concerned with several aspects of the user modeling system proposed in Chapter 3 that include understanding which behaviors could be used as implicit feedback, understanding how information-seeking context affect these behaviors, and understanding how these behaviors can be used to infer document preference. The findings from this research have implications for both the Generalized Behavioral Model and the Personalized Behavioral Model proposed in Section 3.2.

A primary concern for any system that proposes to use behaviors to infer document preference is insuring that the techniques for collecting and computing these behavior-based metrics are valid and reliable. Clearly, a system that makes inferences about which documents users prefer based on erroneous data is unlikely to be very effective at tailoring system responses to users. The results of this dissertation demonstrated that techniques for measuring behavior do affect the types of inferences that one is able to make, and that using invalid and unreliable data can lead to the identification of unsubstantiated relationships, or a failure to detect significant relationships. A secondary concern is determining how many observations are necessary to generate a model of the relationship between behaviors and document preference. In this study, the number of documents viewed by subjects varied considerably, and in many cases, especially for Subject 5, a low number of observations may have prevented significant results from emerging. Moreover, because subjects’ online information-seeking activities consisted of a mixture of various tasks and topics, and the results demonstrated that behaviors differ according to task and topic for most subjects, tracking and computing task- and topic-specific models appears to be an important issue.

The results of this dissertation have several implications for the Generalized Behavioral Model proposed. Recall that the purpose of the Generalized Behavioral Model is to describe how implicit feedback can be used to identify and track a user’s document preferences. These descriptions are general, in that they suggest potential ways that behaviors are likely to vary depending upon on a user’s interests in a document, and depending upon the context in which the user seeks information. The results of this dissertation found no general relationships between behaviors and document preference,
but the results did demonstrate that behaviors differ according to some contextual factors, most notably task and topic, and topic familiarity. These results suggest that perhaps knowledge of some aspects of context is necessary to use behaviors to infer users’ interests, and that a General Behavioral Model should anticipate these relationships. However, the subjects of this study do not represent a sample, so identifying generalizable results with confidence is not possible.

The results of this dissertation have several implications for the Personalized Behavioral Models proposed. The results demonstrated that different contextual variables affected behaviors in different ways, for different subjects, and that the relationship between behaviors and document preference, when examined on a subject-by-subject basis, appears to be a rather tenuous one. Overall, the results provide support for the notion of a Personalized Behavioral Model by demonstrating that differences do exist between subjects’ display times, and that a general, all-purpose model of how behavior can be used to infer document preference is likely to be insufficient.