

## Chapter 7

### Conclusions

The most significant finding of this dissertation research was that information-seeking behavior changed as a result of information-seeking context, and that these changes differed in degree and kind for different users. Specifically, this research demonstrated a statistically significant difference in the length of time users displayed documents according to which task, topic and familiarity ratings users associated with documents. The users of this study did not represent a sample, so the generalizability of the results is limited. However, it appears that individual differences do matter, and that personal approaches to modeling would be more effective than general-purpose approaches. Finally, the most surprising and interesting finding of this study was the number and kinds of tasks and topics that users identified throughout the course of the study. The methodological approach of this study allowed one to understand, from users' perspectives, what activities they were doing online, and how they construed these activities. These results suggest that personalized methodological approaches can yield insightful, rich and usable data.

The overall goal of this dissertation was to understand how online information-seeking behaviors could be used as implicit feedback for document preference and how information-seeking context affects these behaviors. This dissertation was further concerned with determining how personalized user models might be created based on the behavior of a single individual rather than the behavior of a sample of individuals, and how these models might differ from user-to-user. Finally, this dissertation was concerned

with collecting and measuring information-seeking behaviors and context using a valid and reliable method that optimized ecological validity.

The goals of this dissertation were accomplished by conducting a naturalistic, longitudinal study of the online information-seeking and use behaviors of seven subjects during a fourteen-week period. Subjects were provided with laptops and printers, and their online activities were monitored throughout the course of the study with logging software and paper instruments. Subjects were asked to evaluate the documents that they viewed while online at weekly intervals and to classify these documents into personal task and topic classes. At the end of the study, subjects participated in an Exit Interview to provide feedback about their experiences and various aspects of the methodology. Information-seeking behaviors that were measured included display time and retention. Information-seeking context was measured by task and topic, and several attributes of each of these including task endurance, frequency of activities, stage, topic persistence and topic familiarity. Document preference was measured by the explicit usefulness judgments that subjects assigned to the documents that they viewed.

The results of the study demonstrated that neither display time nor retention, by themselves, were good indicators of document preference. Although mean display time was found to differ significantly for usefulness for one of the seven subjects, this relationship was weak as evidenced by post-hoc tests. Thus, contrary to previous research, display time was not found to be a good indicator of usefulness. Results further demonstrate that there is no correlation between display times collected from a proxy and display times collected from a client, and that Type I and Type II errors are likely to result from using proxy-generated display times alone, or in combination with client-

generated display times. Results also showed that retention behaviors did not occur that often, which suggests that their value as implicit evidence of document preference is limited. In addition, no significant differences were found between the usefulness of documents that were retained and those that were not.

Display time was found to be significantly related to information-seeking context in different ways, for different subjects. For task, display time was found to significantly differ for five of the seven subjects and for topic, it was found to significantly differ for all subjects but one. Tasks that had mean display times that significantly differed were those related to reading the news and shopping. Results of the post-hoc tests for topic were less conclusive since topics varied widely for subjects and many topics had only a single document associated with them. Although statistical tests were not performed to evaluate the relationship between retention and task and topic, an examination of the distributions of retention across tasks and topics indicated that subjects appeared to be more likely to retain documents for tasks that included those related to academic research and shopping, and for topics that were associated with the academic research and shopping tasks.

Overall, the results demonstrated for some subjects significant differences in mean display time according to task, topic, endurance, frequency, stage, persistence and familiarity. With the exception of two subjects, mean display time was found to significantly differ for at least one of the task and topic attributes. Familiarity was the strongest attribute with mean display times significantly differing for all subjects except one. All of the post-hoc tests for familiarity were statistically significant and for all subjects but one, lower levels of familiarity were found to be associated with higher

display times. 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. For endurance, mean display time differed significantly for two subjects. Results of the post-hoc tests were mixed and did not indicate a clear direction for this relationship, although differences were detected in the higher endurance scores rather than the lower scores. Finally, mean display time significantly differed according to frequency for two subjects and results of the post-hoc tests indicated that higher frequencies had lower display times.

When combined with usefulness, no significant interaction effects were found between display time and information-seeking context. While there may not be any relationship between these various variables, the results might also be the result of the measurement techniques used in this study. All of the information-seeking context variables, as well as usefulness, were measured by at least seven levels. This type of detailed measurement might prevent significant relationships from being detected because of the distributions of data points across these levels. Alternative methods of measuring and analyzing information-seeking context and usefulness might reveal interactions between these variables. However, the analyses of information-seeking context and display time did reveal several significant relationships, and these results should be viewed as evidence for the complexity of relationship between display time and usefulness. Clearly, more work needs to be conducted to understand the complexity of this relationship, and the potential of using behaviors to infer document preference.

In sum, this research contributes several major findings to the research on implicit feedback. First, proxy-generated display times are not good substitutes for client-

generated display times. Second, behavior by itself is not a good indicator of document preference. Third, behavior can change as a result of information-seeking context. Fourth, relationships between behaviors and information-seeking context differ in degree and kind for different users. Fifth, although tentative, knowledge of information-seeking context is necessary to effectively use behaviors as implicit evidence of document preference.

This research establishes a novel and innovative methodology for studying online information-seeking behavior, information-seeking context, and document preferences over time, in a naturalistic setting. The data collected using this method was incredibly detailed, user-focused and in large quantity. Subjects generally had few problems identifying task and topic classes and classifying documents into these classes, except for determining at what level of abstraction to specify tasks and topics. Task classification seems to be the best place to start with future studies, since there were many overlapping tasks between subjects and overall, there were fewer tasks than topics. Future studies might also explore various techniques for eliciting users' tasks and topics. The techniques used for identifying attributes of tasks and topics were also effective in this study. While subjects identified some problems using these attributes and scales to characterize specific tasks and topics, the results of this study established a starting point from which to begin exploration of techniques for measuring these attributes in online information-seeking environments. Future studies might also investigate the relationship between these aspects of context and how they can be assessed using log-based metrics. Finally, because of the goals of this study, analyses were not performed on task and topic

combinations. However, it is likely that using the task and topic combination as the unit of analysis would yield additional insights.

Because of the methodology, care must be taken when generalizing the results of this study because study subjects represented a rather homogenous group and were not in large enough quantity to be considered a sample. Because all subjects were Ph.D. students, many of their tasks were related to their membership in this group and when present, many of their deadlines for completing tasks corresponded to a single university semester. However, these subjects conducted a variety of tasks, many that were not academic-related. The decision to collect data on a case-by-case basis was made because of the goals of the study. In many respects, the results of this approach to data collection indicate a distinct trade-off between the quantity and the quality of data that one can collect. However, overall, the amount and detail of data collected for each individual is quite large compared to what other studies have managed to collect. Future research might use this study method to collect data on a more heterogeneous sample of users and in greater numbers.

The results of this study have several implications for the theoretical model proposed in Chapter 3. Most notably, the results demonstrated the importance of valid and reliable measurement techniques, and provided support for the notion that a general, all-purpose model of how behavior can be used to infer document preference is likely to be inadequate, and that modeling should occur on an individual basis. The results further demonstrated that behaviors are affected by the context in which the user seeks information, and, in particular, that display times differ significantly according to task, topic and familiarity. Finally, the results showed no direct relationship between

behaviors and document preference, which suggests that behavior may be necessary, but not sufficient evidence for inferring a user's document preference.

The results of this study have several practical implications for user modeling systems using information-seeking behaviors as a technique for constructing user models. First, extensive work was conducted creating a classification of page types to identify which pages should be shown to subjects for evaluation. If a system that makes use of a user's web browsing history is to be created, then the system needs to know when it should consider a document for inclusion in modeling. Whether judged useful or not useful, not all documents are equally good for constructing models. A document that only contains a search box is not as good as one which contains the text of a conference paper. This also applies to using behavior as implicit feedback: observing a high display time at a document containing a search box and little text most likely indicates something different than observing a high display time at a document containing a conference paper. Clearly, the system needs some assistance in identifying candidate documents for inclusion in modeling and as sources of implicit feedback.

A second result which has practical implications for the design of user modeling systems that support information-seeking was that the number of documents viewed by subjects varied considerably. It is unclear if there is an optimal number of observations that needs to be made before a system can infer a rule for using behavior as evidence of document preference. A more focused study on this might indicate that a user modeling system for online information-seeking is not for everybody, especially those users who do not use the Internet that often.

Results of this research suggest additional research agendas involving both new data collection efforts and re-use of the existing data set. Because the current data set consists of a large quantity of online documents that have been classified into personalized task and topic classes by users, its potential for understanding how tasks and topics can be identified with automatic techniques is great. The current data set could be mined using a bottom-up approach to investigate how tasks might be automatically learned and detected for different users over time based on access patterns. For instance, most users who viewed the news did so at the same source, and often at the same time each day. Being able to identify tasks from a bottom-up approach would be useful since it would allow for the automatic identification of some aspects of context. The topic classes that users created and the documents that they associated with these classes could be used to evaluate the effectiveness of automatic techniques for constructing topic clusters. Further, the effectiveness of using behavioral evidence in the construction of these clusters could be directly examined. Preliminary work investigating the effectiveness of two language modeling approaches to topic clustering and the effects of using behavioral evidence during this construction has already been done using pilot data from this study (Kelly, Diaz, Belkin and Allan, 2003).

More work needs to be done on discovering effective formalizations of modeling and retrieval techniques, in terms of how behaviors and context will be collected and used, how documents will be selected and modeled, and how retrieval will be modified as a result of this. The inaugural HARD track (Allan, in press) at this year's TREC has begun to address some of these issues. However, the track lacks an appropriate and richly enough defined data set. In particular, it is difficult to determine what aspects of

context might be useful and how this might be collected and distributed for a TREC-style evaluation. The results of this study not only suggest which aspects of context might be important, but also how one might go about collecting the type of data that the HARD track is interested in exploring.

More research needs to be conducted on understanding what display time represents. Are users actually reading the displayed documents or are they simply skimming the documents? It seems likely that other cognitive activities are occurring aside from reading. Recall that a full 10% of the documents were displayed for one second. Considering that subjects had no problems remembering these documents during their evaluation activities, future research is needed to better understand what is occurring during these one second interactions. Because display time is available for every document that a user requests, its potential as a standard evaluation metric in interactive information retrieval should also be explored further. In addition to the one second interactions, consider the finding that few documents were retained, and in particular, few were printed. This result may indicate a general trend in the reduction of printing of online materials and an increase in the online viewing of documents. Again, future research would need to determine if this is the case and if these one second interactions are related to a user's decision to retain a document.

Finally, more work needs to be conducted on understanding the relationship between various metrics and the behaviors they are meant to measure, and the techniques employed for collecting and computing these metrics. While it is nice to be able to collect a large amount of data anonymously on a proxy, if the metrics are invalid and unreliable, then the results are meaningless. Furthermore, because of the level of interest

in the research community in collecting and using data about users' online interactions, more effort should be made towards developing and sharing tools that collect valid and reliable data, and store it in a format that is easily interpretable and useable by researchers interested in studying information-seeking behavior and context.

In conclusion, this research has contributed to a better understanding of how information-seeking behaviors can be used as implicit evidence of document preference by determining empirically, what characteristics of information-seeking context significantly affect this relationship and how this varies for different users. The research findings have practical implications for how behaviors can be used in the development and maintenance of personalized user models for context- and user-centered information retrieval, and suggest future research agendas that can further address these issues.