Chapter 1  
Introduction

As the amount and accessibility of online information increases every day, tailoring system responses to individual interests is becoming an important problem in information systems research. As more information is delivered electronically, a user’s information-seeking activities are no longer bound geographically or temporally. Users can engage in information-seeking from their offices, homes or hotel rooms at any time of the day, on any day of the week. Furthermore, both users and online information-seeking activities are much more heterogeneous than they were twenty years ago. Because of this, information systems are experiencing greater use now than at any other time in their history, and more importantly, by more people than just librarians or scientists.

Increased access to information is not problem-free access to information. There is an overabundance of sources from which users must choose. Collections are no longer homogeneous sets of documents with conventional structure and a standardized vocabulary, but are instead, heterogeneous sets of documents with varying structure and undifferentiated vocabulary. To compound this problem, when seeking information users typically pose their information problems as short queries, often using ambiguous terms. For these reasons, users face more complex information-seeking environments and are equipped with less effective retrieval tools. Thus, it is critical for systems to obtain a more accurate representation of users’ information needs, tasks and document preferences, maintain these representations over time, and use these to tailor retrieval to the individual.

One possible approach to addressing these issues is to take advantage of the user’s previous information-seeking and use behaviors to identify items which have been of interest to that person in the past. This could be done, for instance, by identifying and recording the documents that the person has looked at and found useful as a result of information-seeking, 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 personalized basis for searching for new documents.

User modeling offers the potential of personalizing interactions and tailoring retrieval by individuating users and tracking their information-seeking behaviors and needs over time. Generally speaking, a user model is a description of the user, created or selected by the system, that facilitates interaction between the two (Allen, 1990). User modeling has been used in information retrieval (IR) to personalize retrieval to the individual by characterizing the user on various dimensions, identifying the various goals and tasks of the user, and selecting specific retrieval strategies and documents based on these characterizations (c.f. Brajnik, Guida, & Tasso, 1987; Croft & Thompson, 1987; Oddy, 1977; Rich, 1983). For instance, THOMAS (Oddy, 1977) personalized information-seeking interactions by creating characterizations of an individual’s preference for documents. This was accomplished by presenting documents to a user, monitoring the user’s reaction to each document and suggesting documents based upon these reactions. 1^3R (Croft & Thompson, 1987) assisted users in generating better queries by using a domain model expert to construct personalized thesauri for users based on traditional relevance feedback techniques and user-specified knowledge structures.

Relevance feedback in IR has been construed as a type of user modeling for single session interactions (Croft, Cronen-Townsend, & Lavrenko, 2001; Ruthven, Lalmas, & van Rijsbergen, 2002). A typical relevance feedback interaction consists of the user marking documents that are found relevant during a search. Keywords are extracted from these documents and added to the user’s query or used to re-weight existing query terms. The query that evolves during this interaction can be considered as a representation or model of the user’s interests. However, this type of modeling is limited in that it only considers interactions for a single search session and for a single information need. Moreover, these techniques usually only allow for binary relevance judgments.

The modeling of long-term information needs across multiple search sessions can be found in the work on information filtering (IF) and recommendation systems (c.f. Luhn, 1958; Oard & Kim, 1998). Current implementations of user modeling in typical IF systems consist of a list of keywords provided by the user or a list of keywords extracted from documents marked as relevant by the user. The resulting model is used to filter incoming documents or to recommend objects over time. For instance, SIFTER (Mostafa, Mukhopadhyay, Lam, & Palakal, 1997) filters computer and information science documents collected from the Internet and commercial database services based on profiles of users generated with traditional relevance feedback techniques.

For user modeling to work in any of these types of systems, the modeling function must somehow obtain information from the user about their interests and preferences. Traditional approaches to accomplishing this have required the user to explicitly provide feedback by, for example, specifying keywords, selecting and marking documents, or answering questions about their interests. Such techniques force users to engage in

additional activities beyond their normal information-seeking behaviors. Since the cost to the user is high and the benefits are not always apparent, it can be difficult to collect the necessary data and thus, the effectiveness of explicit techniques can be limited. Additionally, users may not be able to provide enough terms to accurately or exhaustively cover the scope of their information needs since one of the reasons they are seeking information is presumably because they want to learn about something that they do not already know (Belkin, 2000).

There have been a number of studies that have investigated the potential application of information-seeking and use behaviors as implicit indicators of interest for user modeling and profile acquisition (Claypool, et. al, 2001; Cooper & Chen, 2001; Morita & Shinoda, 1994; Konstan, et. al, 1997; Nichols, 1997; Oard & Kim, 2001; Seo & Zhang, 2000; White, et. al, 2002). These techniques unobtrusively obtain information about users by watching their natural interactions with the system. Behavior-based metrics that have been most extensively investigated as sources of implicit feedback include reading time, saving, printing, selecting and referencing. The primary advantage to using implicit techniques is that such techniques remove the cost to the user of providing feedback and thus provide the system with the potential for creating more robust, dynamic user models without disrupting the user’s information-seeking and use activities. The results of the research on implicit feedback techniques have been encouraging. Morita and Shinoda (1994) found that users spent longer reading those news documents that they found interesting as opposed to those that they did not find interesting. Claypool, et. al (2001) found a similar result for general web documents and also found that the amount of scrolling on a document had a strong positive correlation

with interest for that document. Kim, Oard, and Romanik (2000) found that documents which were saved or printed were more likely to be relevant to users.

The preceding work has contributed much to what is known about using behavior as implicit sources of evidence, but such studies are limited in many ways. Studies based solely on the analysis of log data are limited because often little is known about individual users, or their needs and intentions. Furthermore, techniques used to measure behavior in these studies usually are plagued with validity and reliability issues. While laboratory studies attempt to understand and measure attributes of users, needs and intentions, they are limited by design since they typically make one or more aspects of the user’s information-seeking interactions artificial. Searching tasks or information problems are usually assigned to users, systems are often modified in one or more ways, and interactions occur during an unnatural, and often compressed time frame. Studies of implicit feedback have been further limited because they often assume that all users exhibit the same or similar behaviors, and that this behavior is not subject to change with respect to contextual factors.

Information-seeking research has demonstrated that users exhibit a range of information-seeking behaviors that change with respect to task (Bystrom & Jarvelin, 1995; Vakkari, 1999), topic (Pennanen & Vakkari, 2002), familiarity (Kelly & Cool, 2002) and problem solving stage (Kuhlthau, 1993). Although some previous research (Kelly & Belkin, 2001a; Quiroga & Mostafa, 2002) has found that several contextual factors, such as specific task and topic, affect the relationship between information-seeking behaviors and document preference, no major attempt has been made to integrate implicit feedback and information-seeking research. Moreover, it has been demonstrated

that information-seeking occurs across multiple information-seeking episodes and that information-seeking behavior can change across episode (Lin & Belkin, 2000; Spink, 1996; Wilson, et. al, 2002). No work has attempted to describe systematically how a behavior like reading time changes over time and how this impacts the effectiveness of this behavior as an implicit indicator of document preference.

The purposes of this dissertation research are to: (1) collect and measure information-seeking behaviors using a valid, reliable method that optimizes ecological validity; (2) evaluate the predictive power of information-seeking behaviors as implicit sources of document preference; and (3) determine how information-seeking context affects the occurrences of these behaviors. Using a longitudinal, naturalistic approach, the work reported in this thesis seeks to understand how an individual’s online information-seeking behavior can be used as implicit feedback for the construction and maintenance of a personalized user model for that individual, and how this behavior changes with respect to contextual factors such as task and topic, and characteristics of each of these factors, such as task endurance and topic familiarity.

The work of this thesis contributes to user modeling and information retrieval research by attempting to determine empirically, what characteristics of information-seeking behavior and context are significant for the development and maintenance of user models in the domain of online information-seeking. The results from this study can provide evidence of how behavior changes over time and with respect to contextual factors, such as specific task and topic. This work further contributes to the research by seeking to determine 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 work establishes a novel methodology for studying online information-seeking behavior, information-seeking context, and document preferences over time, in a naturalistic setting. This method is a novel and potentially useful one for studying information-seeking context and behavior in online environments.

The remainder of this thesis is organized as follows: Chapter 2 provides a review of the literature with a special emphasis on user modeling and implicit feedback techniques for model acquisition. Chapter 3 presents the theoretical framework for the study, including the conceptual model, research questions and general study design. Chapter 4 presents details of the study design, including subjects, instruments and procedures. Results are described in Chapter 5. Chapter 6 discusses the results of the study and how these results addressed each of the research questions and the theoretical model presented in Chapter 3. Chapter 7 concludes this thesis by presenting the research contributions, implications, limitations and future research directions.