Predicting Information Searchers' Topic Knowledge at Different Search Stages

Jingjing Liu
School of Library and Information Science, University of South Carolina, Columbia, SC 29208, USA. E-mail: jingjing@sc.edu

Chang Liu*
Department of Information Management, Peking University, Beijing 100871, China. E-mail: imliuc@pku.edu.cn

Nicholas J. Belkin
School of Communication and Information, Rutgers, The State University of New Jersey, 4 Huntington Street, New Brunswick, NJ 08901-1071, USA. E-mail: belkin@rutgers.edu

As a significant contextual factor in information search, topic knowledge has been gaining increased research attention. We report on a study of the relationship between information searchers' topic knowledge and their search behaviors, and on an attempt to predict searchers' topic knowledge from their behaviors during the search. Data were collected in a controlled laboratory experiment with 32 undergraduate journalism student participants, each searching on 4 tasks of different types. In general, behavioral variables were not found to have significant differences between users with high and low levels of topic knowledge, except the mean first dwell time on search result pages. Several models were built to predict topic knowledge using behavioral variables calculated at 3 different stages of search episodes: the first-query-round, the middle point of the search, and the end point. It was found that a model using some search behaviors observed in the first query round led to satisfactory prediction results. The results suggest that early-session search behaviors can be used to predict users' topic knowledge levels, allowing personalization of search for users with different levels of topic knowledge, especially in order to assist users with low topic knowledge.

Introduction

As an optimal way to improve search engine performance and users' search experience, personalization has been increasingly attracting research attention. Personalization tailors search results to particular users or user groups, which can be characterized according to various contextual aspects, including users' knowledge level, motivating task type, location, etc. (Belkin, 2008). To perform personalization, information retrieval systems need to learn about these contextual factors. Systems that do this usually try to do such learning, or *prediction*, through implicit feedback, because of its advantage of not interrupting searchers. User search behavior is one of the main sources according to which such predictions can be made.

User knowledge has been well studied and found to have significant effects on searchers' behaviors. The literature has identified different types of knowledge, including subject domain knowledge and search task topic knowledge. The majority of previous studies on knowledge as it affects search behavior (e.g., White, Dumais, & Teevan, 2009; Wildemuth, 2004) were from the domain knowledge perspective, that is, user knowledge of a subject domain, for example, medicine, law, or computer science. However, some studies (e.g., Liu, Belkin, Zhang, & Yuan, 2013; Liu, Liu, & Belkin, 2013; Zhang, Liu, & Cole, 2013) considered users' familiarity with search task topics (the terms topic knowledge and topic familiarity are used interchangeably in this article), which depicted not the users' knowledge of a subject domain but that of a specific search task topic, for example, their familiarity with the topic of "What genetic loci, such as Mental Health Wellness 1 (MWH1) are implicated in mental health?" (Zhang et al., 2013, pp. 184-185). Zhang et al. (2013) found that topic knowledge affected user behaviors in ways other than did domain knowledge. This suggests that personalization according to user knowledge should consider topic knowledge as well as domain knowledge.

*Corresponding author.

Received February 23, 2015; revised May 8, 2015; accepted June 24, 2015

 DOI: 10.1002/asi.23606

As one can imagine, personalizing search results for users with different levels of knowledge may help them learn more efficiently in the search process and finish the search task more effectively. Various methods such as logistic regression, support vector machines, and decision trees have been used in previous studies to predict user satisfaction, frustration, system switch, etc. (e.g., Feild, Allan, & Jones, 2010; Fox, Karnawat, Mydland, Dumais, & White, 2005; White & Dumais, 2009). An important issue in building prediction models is to ensure their applicability in real system settings, in particular to conduct prediction during the course of a search session. This requires using behaviors that can be monitored by the systems during the search process, instead of after the search has been performed. In recent years, research efforts have investigated this capability, for example, White, Ruthven, and Jose (2005) analyzed the usefulness of explicit and implicit relevance feedback for search systems at three stages: start, middle, and end, Liu, Gwizdka, Liu, & Belkin (2010), and Liu, Liu, Cole, Belkin, and Zhang (2012) predicted search task difficulty using within-session behavior, and Liu et al. (2014) predicted task difficulty at three search stages: beginning, mid-point, and end.

Given this background information, we explore predicting topic knowledge based on behavioral measures that can be measured during the search. The purpose of this study is to compare the differences in the predictions among three different points during search process, that is, the first-round, the middle-point, and the end point. Specifically, we examined the following research questions:

- Does users' knowledge of search task topics affect their search behaviors differently at three different search stages? If so, how?
- Can users' knowledge of search task topics be predicted, during the search session, from their search behaviors? If so, which stage would achieve the best prediction performance?
- What are the significant behavioral factors in predicting users' topic knowledge at different stages?

To explore these questions, a laboratory user experiment was conducted to collect client-side user-system interaction log data. Ratings of topic knowledge were elicited from users after they read the task description and before they started searching for the task. Several highlights of our study are: first, we made an extensive examination of user search behaviors between users with low and high topic knowledge; second, logistic regression models were conducted to generate predictive models of topic knowledge at three search stages: the first-round (early in a search, from the beginning of search until issuing the second query); middle-point (calculated in the middle of the search process); and, end-point (calculated when the search task is finished). Third, our results demonstrate that users' search behaviors at the very beginning had the best prediction accuracy of users' topic knowledge. These all shed light on personalization of information retrieval.

Literature Review

User Knowledge and Information Retrieval (IR)

Much research effort has been devoted to examining the effects of user knowledge on information search. As discussed in the Introduction, there are at least two types of knowledge in previous IR studies: (a) domain knowledge, and (b) task topic knowledge. This section reviews previous studies on these two types of knowledge.

The majority of previous studies on knowledge have focused on domain knowledge. They examined users' search tactics and performance and found they were influenced by the level of users' domain knowledge. For example, Hsieh-Yee (1993) found that domain knowledge affected search tactics. As compared to a familiar subject area, when users worked with a search task outside of their field, they used the thesaurus more for term suggestion, made more effort in preparing for the search, monitored the search more closely, included more synonyms, and tried more term combinations. Vakkari, Pennanen, and Serola (2003) found that medical students began to use a wider and more specific vocabulary in their development of research proposals at the end of a 3-month seminar compared with in the beginning. Sihvonen and Vakkari (2004) found that the number and type of terms selected from the thesaurus for expansion by domain experts in the medical area improved search effectiveness, whereas the use of the thesaurus by domain novices had no impact. Wildemuth (2004) found that low domain knowledge, before the users took a microbiology course, was associated with less efficient selection of concepts to include in the search and with more errors in the reformulation of search tactics. These studies demonstrate that users with high levels of domain knowledge used a thesaurus more frequently, issued more specific query terms, and employed better search tactics. Such differences in their search tactics often led to better search performance or search success. Some studies have explicitly examined the relationship between users' domain knowledge and their search performance. Zhang, Anghelescu, and Yuan (2005) found that a high-level knowledge group of users in the heat and thermodynamics engineering domain were found to have better performance (retrieved slightly more relevant documents), issued longer queries, and had more queries per task. Duggan and Payne (2008) found that knowledge of the music domain had little effect on search performance, but that of the football domain had much effect on search performance: it was positively correlated with search accuracy, and negatively correlated with time spent on web pages and mean query length. White et al. (2009) found that within their domain of expertise, experts search differently than nonexperts in terms of the sites they visit, the query vocabulary they use, their patterns of search behavior, and their search success.

Another line of studies on user knowledge focused on task topic knowledge. Compared with domain knowledge, task topic knowledge is a more specific and narrower concept. Allen (1991) defined "topical knowledge" as

DOI: 10.1002/asi

specific factual knowledge of a topic. In his study, participants were given a multiple-choice test to measure their knowledge level of the search topic, and were divided into different levels of topical knowledge groups by their correctness rate in the test. The study found that people with high topical knowledge used more search expressions than those with low knowledge. Hembrooke, Granka, Gay, and Liddy (2005) investigated the effects of knowledge on users' search term selection and reformulation strategies for web searches. From a list of topic areas, participants were asked to choose two topics where they had some expertise (knowledge) and two in which they had none. Experts were found to issue longer and more complex queries than novices. Experts also used elaborations as a reformulation strategy more often as compared to simple stemming and backtracking modifications used by novices.

Elicitation of users' topic knowledge has been performed in previous studies by asking users to self-report on a Likert scale with regard to their familiarity with the search task topic. Kelly and Cool (2002) asked users to rate their familiarity with a search topic on a 5-point Likert scale in a pretask questionnaire. Through examining the correlation between users' search behaviors and their topic knowledge, the study found that increasing topic knowledge resulted in decreased web page reading time and increased search efficacy (the ratio of the number of saved documents to the total number of viewed documents). This suggests the possibility of inferring topic knowledge from searching behavior. Kelly (2006) examined the effect of context on user behaviors in online searching through a longitudinal and naturalistic study that logged users' everyday laptop activities. Users' topic knowledge was elicited using a questionnaire on a 7-point Likert scale. Statistical significance was found between topic familiarity and document display time for five out of seven participants. Liu and Belkin (2010) also employed a 7-point Likert scale to elicit topic knowledge. They found the total dwell time that users spent on a document during a search session was positively correlated with document usefulness regardless of the users' topic knowledge level, however, the first dwell time (the unobtrusive duration from opening a document to the user leaving it for the first time) was not significantly correlated with document usefulness, but that topic knowledge did affect the relationship. This again suggests that topic knowledge is an important contextual factor that affects users' behavior and judgment of document usefulness. Liu, Belkin, et al. (2013) found that users' self-reported general task knowledge levels increased according to the different search stages in multisession tasks.

Zhang et al. (2013) explicitly compared users' domain knowledge and topic knowledge and their differences in affecting users' search performance. They found that self-reported topic knowledge was significantly correlated with performance measures for individual tasks, and domain knowledge was significantly correlated with performance at more general levels, over multiple task sessions. These studies have shown that topic knowledge has a direct effect on users' search behaviors (e.g., how users examined each page), search efficacy, and search success with respect to the current search tasks.

With regard to a comparison of the topic knowledge assessment methods, assessment through tests of some sort is generally time and effort consuming for each topic (unlike that for a domain). However, questionnaire-based, self-rated topic knowledge elicitation is simple, and studies have confirmed its validity. Allen (1991) had all participants work on a knowledge test about the topic, and also asked them how familiar they were with the topic. There was a high correlation between the test score and self-reported familiarity level of the topic. Cole et al. (2010) also looked at the relationship between self-reported knowledge levels and test-based scores. In their study, one way of determining topic knowledge was based on participants' ratings of familiarity with terms in a thesaurus. The other way was a direct question asking users to self-rate their familiarity level with the search task topics. Results showed that these two measures were highly correlated, supporting the validity of the questionnaire-based self-rated topic knowledge elicitation method. According to previous studies, in this study, we adopted the method to measure users' topic knowledge by their self-rated familiarity level with the search task topics.

Prediction in IR

Prediction is an important strategy in order for IR systems to provide better search experience and results that meet users' needs. Researchers have devoted substantial effort on predicting a number of aspects relevant to search experience, including user knowledge, user satisfaction, frustration, search task difficulty, and search system switch, etc.

Kumaran, Jones, and Madani (2005) attempted to differentiate documents that match different levels of topic knowledge: introductory web pages for low knowledge users, and advanced web pages for users with sufficient background knowledge and knowledge with the key technical or important terms in the topic. Their study indicated that certain document features could be predictive of the document being introductory or advanced, and also predictive of a user who read an advanced or introductory document having high or low knowledge with the topic.

Using multiple regression analysis, Zhang, Liu, Cole, and Belkin (2015) built models predicting users' domain knowledge. Four successful prediction models were identified, each involving a slightly different set of behavioral variables. The models were compared for the best model fit, significance of the model, and contributions of individual predictors in each model. The final model highlights three behavioral variables as domain knowledge level predictors: the number of documents saved, the average query length, and the average ranking position of the documents opened.

Fox et al. (2005) examined implicit behaviors for user satisfaction prediction using Bayesian modeling, decision trees, and a new usage behavior pattern analysis "gene analysis." They found an association between implicit

DOI: 10.1002/asi

measures of user activity and the user's explicit satisfaction ratings. The best models for individual pages included the behaviors: clickthrough, time spent on the search result page (SERP), and how a user exited a result or ended a search session. Guo, Yuan, and Agichtein (2011) used machine learning techniques to predict smart phone users' search success and satisfaction. They investigated client-side interaction signals, including the number of browsed pages, and touch screen-specific actions such as zooming and sliding. Their method resulted in nearly 80% accuracy for predicting searcher success, which significantly outperformed previous models.

Feild et al. (2010) extracted features from query logs and physical sensors in a controlled laboratory user study to build models of searcher frustration prediction using logistic regression. They found that the behavioral measures that were most useful for detecting frustration were: the most recent query's length in characters, the average token length of the most recent query, the duration of the task in seconds, the number of user actions in the task, and the average number of URLs visited per task for the current user. Liu et al. (2012) developed models using logistic regression to predict task difficulty at three levels: (a) first-round level at the beginning of the search, (b) accumulated level during the search, and (c) whole-session level at the end of the search. Their results showed that a model incorporating within-session behaviors (those that can be calculated in the search process while the search is going on) had fairly good prediction performance (accuracy 79%; precision 88%), which is comparable with a model using the whole-session level behaviors that are computed only after a search is completed (accuracy 75%; precision 92%). White and Dumais (2009) examined search engine switching behaviors, and developed and evaluated predictive models of switching behavior using logistic regression. Their study demonstrated a relationship between search engine switching and factors such as dissatisfaction with the quality of the results, the desire for broader topic coverage or verification of encountered information, and user preferences.

Task Stage in Information Search and Personalization

Search stage has been found to be a significant factor affecting search behaviors, and also a helpful factor that can be used in personalization. Kuhlthau's (1991) information seeking process (ISP) model found that information seekers' feelings, thoughts, and actions vary along six stages in search process: initiation, selection, exploration, formulation, collection, and presentation. Vakkari and Hakala (2000) and Vakkari (2001) found that, users' querying behavior changed, users were more likely to enter only a fraction of the terms at the beginning of the search, and tended to use more synonyms and parallel terms as the task stage progressed; and the query vocabulary usually changed from broader to narrower terms. White et al. (2005) found that implicit relevance feedback in information retrieval was used more in the middle of the search than at the beginning or end, whereas explicit relevance feedback was used more towards the end. Liu and Belkin (2010) found that for multisession tasks, task stage helped interpret document usefulness from the first dwell time, that is, the first duration that a document was viewed.

Method

To explore our research questions, a controlled laboratory experiment was conducted with four different types of motivating search tasks. Participants performed searches for these tasks in our usability lab. Questionnaires were used to elicit users' topic knowledge level before each search. A variety of user search behaviors were collected, including the number of queries, number of viewed documents, and dwell time on web pages. The following subsections explain the experiment in more detail.

Tasks

We followed Li and Belkin's (2008) classification scheme to design our tasks. This is one of the very few examples of task classification in the IR literature that attempts to identify and integrate the various facets of task in a single scheme. Li and Belkin's (2008) classification scheme has 15 facets of work or search task. Work task is identified as the task that leads one to engage in information-seeking behavior, and search tasks as the specific information-seeking activities themselves. The classification itself is meant to apply to both types of tasks, and in our study, we focused on values associated with search tasks.

It is important to note that the goal of this study is not to examine these four specific tasks or task topics, but rather task types. The tasks we designed in this study have different features or facets, representing a variety of task types in real life searches that are general to many search tasks and topics; thus controlling task facets can increase the generality of our study. Indeed, Li and Belkin's (2008) faceted classification was based on empirical data gathered from a quite different context than journalism, the variety of work tasks performed in a university environment, by staff, students, and faculty.

Table 1 is an overview of the facets of Li and Belkin's (2008) classification scheme that we manipulated. We added one facet, "Level," which we found to be a significant aspect of tasks in the work environment we studied. We held constant the values of the following facets (not in Table 1): Source of task, Task doer, Time (length), Process, Goal (quantity), Interdependence, and Urgency. The choice of facets to be varied was based on related studies' results (e.g., Li & Belkin, 2010), and on characteristics of typical work tasks in the journalism domain.

For reasons of both validity and convenience, the work domain of journalism was chosen in our study. Although journalism can be associated with any topic, it has a relatively small number of work task types. This means that we were able to have a range of topics for our tasks, while maintaining a good measure of control over realistic tasks,

DOI: 10.1002/asi

TABLE 1. Task classification scheme facets (after Li & Belkin, 2008).

Facets	Values	Operational definitions/rules
Product	Physical	A task that produces a physical product
	Intellectual	A task that produces new ideas or findings
	Decision (Solution)	A task that makes a decision or solves a problem
	Factual information	A task locating facts, data, or other similar items in information systems
	Image	A task locating image(s) in information systems
	Mixed product	A task locating different types of items in information systems
Goal (Quality)	Specific goal	A task with a goal that is explicit and measurable
	Amorphous goal	A task with a goal that cannot be measurable
	Combined goal	A task with both concrete and amorphous goals
Objective task complexity	High complexity	A work task involving at least five activities during engaging in the task; a search task involving searching at least three types of information sources
	Moderate	A work task involving three or four activities during engaging in the task; a search task involving searching two types of information sources
	Low complexity	A work task involving one or two activities during engaging in the task; a search task involving searching one type of information source
Level	Document	A task for which a document as a whole is judged
	Segment	A task for which a part or parts of a document are judged

TABLE 2. Variable facet values for the search tasks.

Task	Product	Level	Goal (Quality)	Objective complexity
BIC	Mixed	Document	Specific	High
CPE	Factual	Segment	Specific	Low
INT	Mixed	Document	Mixed	Low
OBI	Factual	Document	Amorphous	High

thus enhancing validity. We also had ready access to a university journalism department, which meant both that we had experts to help us define the work tasks and access to participants trained for such professional journalism tasks.

Journalism task identification was conducted by interviewing journalism faculty and practicing journalists about typical journalism work and searching tasks for which professional journalists receive training. Task descriptions were formalized from those interviews. We then identified a set of four work or search tasks along several task classification facets, which we believed could affect users' search behaviors.

The four work tasks and associated search tasks that we identified are presented here. These tasks follow the simulated task environment as proposed by Borlund (2003), and are couched in journalism terms: Journalists are typically given an assignment, and an associated task to complete.

Background Information Collection (BIC): Your assignment: You are a journalist at the *New York Times*, working with several others on a story about "whether and how changes in US visa laws after 9/11 have reduced enrollment of international students at universities in the US." You are supposed to gather background information on the topic, specifically, to find what has already been written on this topic. **Your Task**: Please find and save all the stories and related materials that have already been published in the last 2 years in the *New York Times* on this topic, and also in five other important newspapers, either US or foreign.

Interview Preparation (INT): Your assignment: Your assignment editor asks you to write a news story about "whether state budget cuts in New Jersey are affecting financial aid for college and university students." **Your Task**: Please find the names of two people with appropriate expertise that you are going to interview for this story and save just the pages or sources that describe their expertise and how to contact them.

Advance Obituary (OBI): Your assignment: Many newspapers commonly write obituaries of important people years in advance, before they die, and in this assignment, you are asked to write an advance obituary for a famous person. **Your Task**: Please collect and save all the information you will need to write an advance obituary of the artist Trevor Malcolm Weeks.

Copy Editing (CPE): Your assignment: You are a copy editor at a newspaper and you have only 20 minutes to check the accuracy of the three underlined statements in the excerpt of a piece of news story below. New South Korean President Lee Myung-bak takes office. Lee Myung-bak is the *tenth* man to serve as South Korea's president and the first to come from a business background. He won a landslide victory in last December's election. He pledged to make the economy his top priority during the campaign. Lee promised to achieve 7% annual economic growth, double the country's per capita income to *US$4,000* over a decade and lift the country to one of the top seven economies in the world. Lee, *66*, also called for a stronger alliance with its principal ally Washington and implored North Korea to forgo its nuclear ambitions and open up to the outside world, promising a better future for the impoverished nation. Lee said he would launch massive investment and aid projects in the North to increase its per capita income to US$3,000 within a decade "once North Korea abandons its nuclear program and chooses the path to openness." **Your Task**: Please find and save an authoritative page that either confirms or disconfirms each statement.

Classification of Tasks

Table 2 shows the values of the varied facets for each of the four search tasks that we gave to the participants. These

DOI: 10.1002/asi

values constitute the independent variables in our study, which are related to the dependent behavioral search variables.

BIC was a Mixed Product, because identifying "important" newspapers is intellectual, and finding documents on the topic is factual. It is at the Document Level because whole stories are judged; it has the Specific Goal of finding documents on a well-defined topic; it has High Objective Complexity because of the number of sources and activities that need to be consulted or performed.

CPE was a Factual Product, because facts have to be identified; it is at the Segment Level, because items within a document need to be found; it has the Specific Goal of confirming facts; it has Low Objective Complexity because only three facts need to be confirmed.

INT was a Mixed Product, because defining expertise is intellectual, and contact information is a fact; it is at the Document Level, because expertise is determined by a whole page; Goal Quality is Mixed, because determining expertise is amorphous but contact information is specific; it has Low Objective Complexity because only two people need to be found.

OBI was a Factual Product, because facts about the person are needed; it is at the Document Level because entire documents need to be examined; Goal Quality is Amorphous because "all the information" is undefined; it has High Objective Complexity because many facts need to be found.

We acknowledge that these tasks each had specific requirements, for example, the BIC task has a 2-year limitation requirement, the CPE task has an obvious credibility requirement, the INT task has a requirement of looking for the specific web pages, and the OBI task has a requirement of identifying the right person. These may affect users' search behaviors in various ways, but we also think that specific requirements related to the facet values exist in all the tasks and that the users will have taken these into consideration while they read the task requirements and made assessment of their knowledge with the topics, and conducted the search. The effects of task type on users' interactions were analyzed in Liu, Cole, et al. (2010), and the current paper focuses on predicting users' topic knowledge at different search stages using their search behaviors.

Participants

A convenience sampling method was used in the study by recruiting students from Rutgers University undergraduate Journalism and Media Studies program to mimic journalists. To ensure that the participants had appropriate journalism skills, only upper-division undergraduates who had completed at least one journalism writing or reporting class were selected. There were a total of 32 participants, 26 female and six male, aged between 18 and 27 years old. Most students were native English speakers (73%) with the remainder of the population claiming a high degree of English knowledge. Participants rated their computing skills high with an average search experience of 8.5 years. Students rated their search experience generally high but claimed more experience with WWW search than online library catalog search. They were generally positive about their average success during online search.

Participants were recruited from relevant writing and reporting classes, using flyers and via targeted e-mails. They were informed in advance that their payment for participation in the experiment would be $20.00, and that the eight who saved the best set of pages for all four tasks, as judged by an external expert, would receive an additional $20.00. The rationale for the extra payment was to encourage participants to treat their assigned tasks seriously.

System and Data Collection

The experiment was conducted using a system that was designed for interactive information retrieval (IIR) experiments that logs users' multidimensional interactive search behavior (Bierig et al., 2010). The system has a client-server architecture where researchers configure IIR experiments from a range of extensible tasks. The current experiment l configuration applied assigned tasks and questionnaires, which included (a) a background questionnaire, (b) pre- and post task questionnaires to gather users' perceptions on their familiarity with search task topics, task difficulty, success, and satisfaction, etc., and (c) a usefulness questionnaire that elicits users' judgments of the saved documents' usefulness with respect to their task. Users accessed the experiment through an interface that presented them with their tasks, provided them with additional instructions, and administered the various questionnaires. The system is able to log a wide range of user behaviors with an array of heterogeneous logging tools. The data used for the current article were captured using logging software Morae Recorder 3.0 (http://www.techsmith.com).

The search interface in the experiment system has two frames: on the right side is the regular Internet Explorer (IE) window, with a blank starting page; on the left side is a panel that allows the users to save desired pages and also to delete them in case they change their mind. Figure 1 depicts the search interface with two saved web pages.

Procedure

Each participant was invited individually to an interactive lab to conduct the experiment. After signing the consent form, the participants were given a warm-up task as a tutorial. They then performed the four web search tasks. The order of the four tasks was systematically rotated for each participant following a latin square design for a total of 32 participants. Before each task, the participants were given a pretask questionnaire that asked about their self-assessed familiarity with the search task type, search task topic, and estimated difficulty of the task. After each task, they were given a post task questionnaire that asked about the usefulness of each page they saved, their experienced difficulty of the task and their satisfaction with the results of the search.

DOI: 10.1002/asi

FIG. 1. The search system interface. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Although the experiment setting was controlled, participants were allowed to search freely on the web using IE 6.0 to access any search engines or websites in their search for information. Participants were asked to continue the search until they had gathered enough information to accomplish the task, although they were reminded that when they had been searching for 15 minutes. All of the participants' interactions with the computer system were logged. The entire search process was stored via the Morae screen-capture program. In each task, when participants decided that they had found and saved enough information objects for purposes of the task, they were then asked to evaluate the usefulness of the information objects they saved, or saved and then deleted, through replaying the search using the screen capture program. An online questionnaire was then administered to ask about their searching experience, including their subjective evaluation of their performance, and reasons for that evaluation.

After completing four different tasks, an exit questionnaire was administered, asking about subjects' perceptions of their search experiences, the extent to which they found differences in the tasks, their ability to perform the tasks, and their overall search experiences in the tasks.

Behavioral Variables and Definitions

The following behavioral variables were examined, including those about querying behaviors and web page reading behaviors:

- *Number of queries*: the number of queries issued
- *Query length*: the length of the query measured in the number of words
- *Average query length:* the average number of words or all queries
- *Query interval (seconds)*: the total elapsed time after a query is issued and before the next query (if any) is issued
- *Average query interval (seconds)*: the average of all query intervals, from the time point after one query is issued and before the subsequent query is issued
- *Number of viewed documents*: the number of documents that the user viewed (sometimes a user viewed the same documents repeatedly)
- *Number of viewed unique documents*: the number of unique documents that the user viewed
- *Number of documents per query*: average number of viewed documents per query
- *Number of unique documents per query*: average number of viewed unique documents per query
- *Number of viewed SERPs*: the number of SERPs that the user viewed. Different pages in a search result are different SERPs (sometimes a user viewed the same SERPs repeatedly)
- *Number of viewed unique SERPs*: the number of unique SERPs that the user viewed
- *Number of SERPs per query*: average number of viewed SERPs per query
- *Number of unique SERPs per query*: average number of viewed unique SERPs per query
- *Dwell time on a document (seconds)*: the elapsed time between the time when the user viewed a document and the time he/she left this document
- *First dwell time on a document (seconds)*: the elapsed time between when a user first opened a document and when the user first left the document

DOI: 10.1002/asi

- *Mean first dwell time on documents (seconds)*: the average of durations between when a user opened a document and when the user first left the document
- *Total time spent on documents (seconds):* the sum of time users spent on all viewed documents
- *Mean dwell time for all documents (seconds)*: the average of the total dwell time for all documents viewed
- *Mean dwell time for unique documents (seconds)*: the average of the total dwell time for all unique documents viewed
- *Dwell time on a SERP (seconds)*: the duration between the time when the user viewed a SERP to the time when he/she left this SERP
- *First dwell time on a SERP (seconds)*: the duration between when a user first opened a SERP and when the user first left the SERP
- *Mean first dwell time on SERPs (seconds)*: the average of durations between when a user opened a SERP and when the user first left the SERP
- *Total time spent on SERPs (seconds): the aggregate time users spent on all SERPs*
- *Mean dwell time of all SERPs (seconds)*: the average of the total dwell time of all SERPs viewed
- *Mean dwell time of unique SERPs (seconds)*: the average of the total dwell time of all unique SERPs viewed
- *Task completion time (min.)*: time users spent on each task in minutes

Results

Users' Knowledge Level in General

Because self-reported knowledge levels on search task topics have been found to be highly correlated with test-based elicitation of topic knowledge in previous research (e.g., Cole et al., 2010), as well as the fact that self-reported knowledge elicitation is manageable with multiple task topics in a study, questionnaire-based, self-ratings of topic familiarity were used to represent users' topic knowledge levels in this study.

The 32 participants each completed four search tasks, coming up with a total of 128 experimental sessions. Before each task, the participants were given a pretask questionnaire asking about their self-assessed familiarity with search task topics on a 7-point scale as the evaluation of their topic knowledge. Although 7-points are good for respondents, this scale contains distinctions too fine for a future personalization system to differentiate. We therefore collapsed the rating scores into two groups based on the distribution of scores: scores 1–3 into a *low* knowledge (*Novice*) group, and scores 4–7 into a *high* knowledge (*Knowledgeable*) group. Table 3 shows the number of sessions in the two groups in each task.

TABLE 3. Topic knowledge group distribution by tasks.

	Knowledge groups		
Tasks	Number of novice subjects	Number of knowledgeable subjects	Total
BIC	20	12	32
CPE	19	13	32
INT	12	20	32
OBI	23	9	32
Total	74	54	128

Search Behaviors by Topic Knowledge Groups

This section reports the comparison of behavioral measures by the two topic knowledge groups when all four tasks are considered. We aim in this article for "in-session" prediction of topic knowledge, and so it is also our goal to examine novice and knowledgeable users' behavioral differences at different time phases during the search session.

Given our data set with the limited number of 128 sessions, and the fact that the task completion time and number of queries varied by session, one reasonable and appropriate method of analysis was to conduct the prediction by task stage, or time phase, which was operationalized as making the prediction at three points:

First round (FR): the behavioral measures were calculated right before users issued the second query in the search session.
Middle point (MID)[1]: the behavioral measures were calculated after users issued their middle queries and before they issued subsequent queries. The middle query was defined as the query whose sequence number equals half of the total number of queries in that search session (the number was rounded up if needed). For example, if a search session contains three or four queries in total, then the middle point means after the second query was issued and right before the third query was issued.
End point (END): the behavioral measures were calculated after the whole search session was finished.

Because three different points of search stage were considered in this study, search sessions that contained fewer than three queries were excluded from the current examination. In total, there were nine sessions excluded, leaving 119 sessions in the analysis.

An exploration of the data found that most of the variables were not normally distributed, so the nonparametric Mann-Whitney U tests were used to explore if there were differences in the examined behavioral measures between the two groups of users. As can be seen from Table 4, no single variable at the FR phase showed significant differences between groups, and only one single variable (mean first dwell time on SERPs) did at the MID and the END phases. Specifically, in the MID phase, users with low topic knowledge had longer mean first dwell time on SERPs (Median = 10.75 seconds) than high topic knowledge users (Median = 8.70 seconds), U = 1295.5, $p = 0.03$; and in the END phase, low topic knowledge users again had longer

[1]We recognize that in an operational environment, MID cannot be determined during search, and that prediction would take place query-by-query. For the purposes of this article, we used the MID measurement to have one common point that is comparable across sessions, for demonstration purposes.

DOI: 10.1002/asi

TABLE 4. Search behaviors by topic knowledge groups (median of low and high groups, and Mann-Whitney U(p) values) (bold indicated those with significant differences at 95% confidence level).

	Behavioral measures	Novice (median)	Knowledgeable (median)	Mann-Whitney U(p)
FR	First dwell time on first SERP	7.55	10.40	1,971.0 (.13)
	First dwell time on first viewed document	14.10	15.40	1,622.0 (.70)
	First query length	3	3	1,702.5 (.95)
	First query interval	30.05	23.90	1,589.5 (.58)
	Number of viewed documents at first query	1	1	1,751.5 (.73)
	Number of viewed unique documents at first query	1	1	1,739.0 (.79)
	Number of viewed SERPs at first query	1	1	1,872.0 (.27)
	Number of viewed unique SERPs at first query	1	1	1,764.0 (.08)
	Mean dwell time of all documents at first query	7.29	6.15	1,543.0 (.40)
	Mean dwell time of unique documents at first query	7.78	6.15	1,536.5 (.38)
	Mean total dwell time of all documents at first query	11.85	8.30	1,596.0 (.59)
	Mean dwell time of all SERPs at first query	8.98	7.30	1,461.0 (.21)
	Mean dwell time of unique SERPs at first query	12.80	11.20	1,552.5 (.41)
	Mean total dwell time of all SERPs at first query	12.80	11.20	1,583.5 (.56)
MID	Mean dwell time of all documents	13.84	14.34	1,633.0 (.75)
	Mean dwell time of unique documents	18.34	17.84	1,537.5 (.40)
	Mean first dwell time on documents	15.03	14.26	1,649.5 (.82)
	Mean dwell time of all SERPs	10.14	8.41	1,355.5 (.07)
	Mean dwell time on unique SERPs	16.71	14.68	1,481.0 (.25)
	Mean first dwell time on SERPs	**10.75**	**8.70**	**1,295.5 (.03)**
	Number of documents per query	1.50	1.70	1,831.5 (.45)
	Number of unique documents per query	1.25	1.29	1,831.5 (.45)
	Number of SERPs per query	1.75	2	1,910.0 (.24)
	Number of unique SERPs per query	1	1	1,761.5 (.69)
	Average query length	4	3.90	1,582.5 (.55)
	Average query interval	43.26	48.38	1,756.0 (.73)
	Number of all documents	11	12	1,737.0 (.81)
	Number of unique documents	8.5	10	1,747.0 (.77)
	Number of SERPs	12	13	1,757.5 (.72)
	Number of unique SERPs	7.5	8	1,699.5 (.97)
	Number of queries	7	6	1,694.0 (.99)
	Total time spent	313.85	364.80	1,658.0 (.85)
	Total time spent on documents	164.85	157.10	1,633.0 (.75)
	Total time spent on SERPs	132.55	110.50	1,607.5 (.65)
END	Mean dwell time of all documents	12.34	11.76	1,556.0 (.46)
	Mean dwell time of unique documents	18.51	16.21	1,602.5 (.63)
	Mean first dwell time on documents	13.73	12.51	1,627.0 (.72)
	Mean dwell time of all SERPs	9.60	8.79	1,391.0 (.10)
	Mean dwell time on unique SERPs	17.34	14.93	1,418.5 (.14)
	Mean first dwell time on SERPs	**10.00**	**8.30**	**1,199.0 (.01)**
	Number of documents per query	1.81	2.14	1,915.5 (.22)
	Number of unique documents per query	1.30	1.56	1,995.5 (.10)
	Number of SERPs per query	1.98	2.03	1,685.5 (.97)
	Number of unique SERPs per query	1.13	1.11	1,620.0 (.69)
	Average query length	4.15	4.20	1,592.0 (.59)
	Average query interval	50.18	52.82	1,739.0 (.80)
	Numbers of all documents	27	31	1,808.0 (.53)
	Numbers of unique documents	20.5	23	1,814.5 (.51)
	Number of SERPs	25.5	29	1,643.0 (.79)
	Number of unique SERPs	15.5	14	1,610.5 (.66)
	Number of queries	14	11	1,692.5 (.99)
	Task completion time (minutes)	14.31	11.78	1,576.5 (.53)
	Total time spent on documents	394.70	353.80	1,600.5 (.62)
	Total time spent on SERPs	267.20	252.30	1,510.5 (.32)

Note. All time variables are measured in seconds except for those specified as minutes.

DOI: 10.1002/asi

mean first dwell time on SERPs (Median = 10.00 seconds) than High topic knowledge users (Median = 8.30 seconds), U = 1199.0, $p = 0.01$.

Logistic Regression Models and Cross Validation Preparation

Because single searcher behaviors were not strongly associated with topic knowledge, we investigated the predictive power of combinations of behaviors. Binary logistic regression was chosen to predict searchers' knowledge levels. This is a frequently used method for predicting dichotomous dependent variables.

In the analysis, six models were built using variables at different time points: FR; MID; END; FR + MID; FR + END; and, ALL (FR + MID + END). Among these six models, FR, MID, and FR + MID can be viewed as during-session prediction models, which could learn users' knowledge levels while the they are searching; and END, FR + END, and ALL are end-of-session models, which cannot be applied until the end of a session and therefore cannot be used to make prediction during a search session. The latter three models were examined to compare their performance with in-session prediction, on the assumption that having the END information would lead to more accurate prediction.

In order to validate the models, a five-fold cross validation method was used. The models' ability to generalize was tested by evaluating their performance on a set of data not used for training. The 119 data sessions were randomly divided into five groups, with a selection criterion that the same user's sessions were put in the same fold, the purpose of which was to avoid the models learning specific users' behaviors. A total of five runs were conducted, each time using one group as the test set and the other four groups as the training set. The performance measures were then averaged across the five runs to obtain the final values.

Because the range of raw data values varied widely because of the different measuring nature of different variables, a normalization process was performed to make each variable's value range between [0, 1]. The formula used for this normalization is as follows:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

The major assumptions of logistic regression are: (a) the predictors (independent variables) do not have high intercorrelations (multicollinearity), (b) large sample size, and (c) there are no outliers (Tabachnick & Fidell, 2012). The variables should be selected for the predictors used in the model for them to meet the assumption requirements. For the first assumption, the coefficients between the behavioral variables were generated and examined. As O'Brien (2007) noted, variables with variance inflation factor (VIF) values higher than 10 suggests a multicollinearity problem. For each model, in each of the five folds, variables were examined for their VIF values based on the training set, and those with VIF values higher than 10 were removed for the logistic regression analysis, leaving those variables not highly correlated with each other. Then these variables were applied on the test set to generate predictions. For the second assumption about sample size, Hosmer, Lemeshow, and Sturdivant (2013) suggests a minimum of 10 cases for each predictor or independent variable, meaning that the number of independent variables in the current research should not exceed 13 to ensure adequate statistical power. The number of variables selected for the first assumption met the second assumption. The third assumption was also checked, and the current study data set did not have outliers (the general rule of 3*standard deviation was used).

The following variables in Table 5 were selected to meet the aforementioned assumptions, illustrated by the different folds in the six models.

Cross Validation Results

A number of measures were used in this study to evaluate the models' prediction performance. In this context, accuracy is the fraction of correctly predicted novices (i.e., predicted novices that actually were novices) and correctly predicted knowledgeable users (i.e., predicted knowledgeable participants that actually were knowledgeable participants) divided by the total number of participants for that search topic. Precision is defined as the number of correctly predicted novices divided by the total number of sessions predicted as novices. F(β) score measures a test's accuracy by considering both precision and recall (Van Rijsbergen, 1979). Because the experiment tasks required users to find and save as many relevant documents as possible, F (β = 0.5) was selected. F (β = 0.5) was calculated using the following equation:

$$F_\beta = \left(1 + \beta^2\right) \cdot \frac{\text{precision} \cdot \text{recall}}{\left(\beta^2 \cdot \text{precision}\right) + \text{recall}}$$

We selected the baseline as the case when all the users were predicted as novice users, and then calculated the accuracy, precision and F (β = 0.5) measure. Table 6 shows the performance of the six models as well as the baseline model. The values were all averaged across five runs.

As can be seen in Table 6, all six models received higher performance than the baseline model in all three measures. FA showed the best performance in all three measures, and had quite large improvements over the baseline model: 11.03% in accuracy, 13.46% in precision, and 9.31% in F(β = 0.5). The FR + END model followed FR, with a 11.03% improvement over baseline in accuracy, a 13.07% improvement in precision, and a 9.27% improvement in F(β = 0.5).

Conveying the same information as Table 6, Figure 2 more intuitively demonstrates the prediction performance of the six predictive models. It is clear that the FR + END and FR models had better performance than the others, and that adding END to FR resulted in slightly decreased

DOI: 10.1002/asi

DOI: 10.1002/asi

TABLE 5. Variables selected for each fold in each model.

			Run 1	Run 2	Run 3	Run 4	Run 5
1	**FR**	1	firstDwellTimeFirstSerp	firstDwellTimeFirstSerp	firstDwellTimeFirstSerp	firstDwellTimeFirstSerp	firstDwellTimeFirstSerp
		2	firstDwellTimeFirstDoc	firstDwellTimeFirstDoc	firstDwellTimeFirstDoc	firstDwellTimeFirstDoc	firstDwellTimeFirstDoc
		3	firstQueryLength	firstQueryLength	firstQueryLength	firstQueryLength	firstQueryLength
		4	fr_num_serp	fr_num_serp	fr_num_serp	fr_num_serp	fr_num_serp
		5	fr_num_serp_unique	fr_num_serp_unique		fr_num_serp_unique	
		6				fr_mean_dwell_serp	
2	**MID**	1	middle_mean_firstdwell_serp	middle_mean_firstdwell_serp	middle_ave_total_dwell_content_unique	middle_ave_querylength	middle_ave_total_dwell_content_unique
		2	middle_num_unique_serp_per_query	middle_ave_querylength	middle_mean_firstdwell_serp		middle_mean_firstdwell_serp
		3	middle_ave_querylength		middle_ave_querylength		middle_ave_querylength
3	**FR + MID**	1	firstDwellTimeFirstSerp	firstDwellTimeFirstSerp	firstDwellTimeFirstSerp	firstDwellTimeFirstSerp	firstDwellTimeFirstSerp
		2	firstDwellTimeFirstDoc	firstDwellTimeFirstDoc	firstDwellTimeFirstDoc	firstDwellTimeFirstDoc	firstDwellTimeFirstDoc
		3	firstQueryLength	firstQueryLength	firstQueryLength	firstQueryLength	firstQueryLength
		4	FR_number of serp	FR_number of serp	FR_number of serp	FR_number of serp_unique	FR_number of serp
		5	FR_number of serp unique	middle_q.y	middle_mean_firstdwell_serp	middle_ave_querylength	middle_ave_querylength
		6	middle_ave_querylength		middle_ave_querylength		mid_mean_query_interval
4	**END**	1	end_first_dwell_serp_unique	end_first_dwell_serp_unique	end_first_dwell_serp_unique	averageQueryLength	end_first_dwell_serp_unique
		2	averageQueryLength	averageQueryLength	averageQueryLength		averageQueryLength
5	**FR + END**	1	firstDwellTimeFirstSerp	firstDwellTimeFirstSerp	firstDwellTimeFirstSerp	firstDwellTimeFirstSerp	firstDwellTimeFirstSerp
		2	firstDwellTimeFirstDoc	firstDwellTimeFirstDoc	firstDwellTimeFirstDoc	firstDwellTimeFirstDoc	firstDwellTimeFirstDoc
		3	firstQueryLength	firstQueryLength	firstQueryLength	firstQueryLength	firstQueryLength
		4	fr_num_serp	fr_num_serp	fr_num_serp	Fr_num_serp_unique	fr_num_serp
		5	fr_num_serp_unique	averageQueryLength	averageQueryLength	averageQueryLength	averageQueryLength
		6	averageQueryLength				
6	**ALL**	1	firstDwellTimeFirstSerp	firstDwellTimeFirstSerp	firstDwellTimeFirstSerp	firstDwellTimeFirstSerp	firstDwellTimeFirstSerp
		2	firstDwellTimeFirstDoc	firstDwellTimeFirstDoc	firstDwellTimeFirstDoc	firstDwellTimeFirstDoc	firstDwellTimeFirstDoc
		3	firstQueryLength	firstQueryLength	firstQueryLength	firstQueryLength	firstQueryLength
		4	fr_num_serp_unique	fr_num_serp		fr_num_serp_unique	

Note.

- firstDwellTimeFirstSerp: first dwell time on the first SERP in the first query round.
- firstDwellTimeFirstDoc: first dwell time on the first document in the first query round.
- firstQueryLength: first query length.
- fr_num_serp: number of SERPs in the first query round.
- fr_num_serp_unique: number of unique SERPs in the first query round.
- fr_mean_dwell_serp: mean dwell time on SERPs in the first query round.
- middle_mean_firstdwell_serp: mean first dwell time on SERPs at the middle point.
- middle_num_unique_serp_per_query: number of unique SERPs per query at the middle point.
- middle_ave_querylength: average query length at the middle point.
- middle_ave_total_dwell_content_unique: average total dwelling time on unique documents at the middle point.
- middle_q.y: number of queries at the middle point.
- middle_mean_firstdwell_serp: mean first dwell time on SERPs at the middle point.
- middle_ave_querylength: average query length at the middle point.
- mid_mean_query_interval: mean query interval at the middle point.
- end_first_dwell_serp_unique: first dwell time on unique SERPs at the end point.
- averageQueryLength: average query length at the end point.

TABLE 6. Performance measures in the six models: values and percentage over the baseline model.

Models	Accuracy (percentage over baseline)	Precision (percentage over baseline)	$F(\beta = 0.5)$ (percentage over baseline)
Baseline	57.62%	57.62%	62.87%
FR	**63.98% (+11.03%)**	**65.37% (+13.46%)**	**68.72% (+9.31%)**
MID	58.20% (+1.01%)	59.98% (+4.09%)	64.40% (+2.44%)
END	62.98% (+9.30%)	64.07% (+11.19%)	67.90% (+8.01%)
FR + MID	60.43% (+4.87%)	63.89% (+10.88%)	66.74% (+6.17%)
FR + END	**63.98% (+11.03%)**	65.15% (13.07%)	68.69% (+9.27%)
ALL	61.64% (+6.98%)	62.44% (+8.37%)	66.75% (+6.19%)

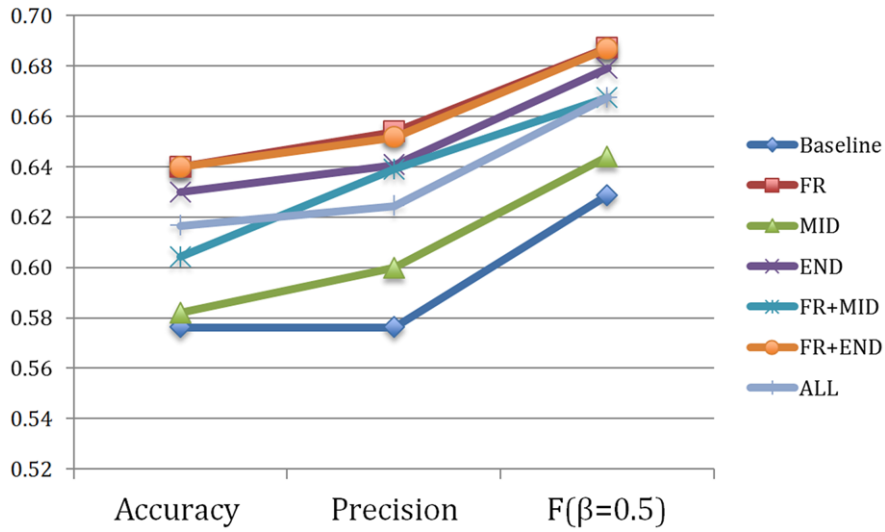

FIG. 2. Performance measures in six models. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

performance. As an in-session, and indeed, early session prediction model, FR led to what could be called "decent performance."

Figure 3 is the recall-precision graph of the six models. Table 7 shows the AUC (area under the ROC curve) of these models, with the greater value corresponding to better performance. As can be seen, the END model had the best performance (0.62) overall when the recall value varies from 0 to 1, followed by FR. Again, as an early session prediction model, FR did a decent job.

Although the primary goal of the current research is to explore the predictive performance of the models at various search stages, it is still informative to determine the variables with their weights in the early-session FR model that shows practically good performance. Table 8 shows the coefficient B (as an indicator of the weight) and *p* values in the FA model built using the whole data set. The variables were selected before doing this. Among the four variables, two had negative B values, and two had positive B values. None of the four variables showed statistical significance on their own; First dwell time on first SERP appeared to have the most weight (B = 4.936). The equation of this model for calculating the predicted topic knowledge would be:

$$\text{Log}(p/(1-p)) = -0.572 + 4.936*(\text{First dwell time on first SERP}) - 0.975*(\text{First dwell time on first viewed document}) - 0.936*(\text{First query length}) + 0.929*(\text{Number of SERPs at first round})$$

(p is the probability of a user is predicted as low topic knowledge).

Discussion

Although our study is based on a controlled lab experiment, and is not naturalistic, the use of tasks familiar to the participants, and of scenarios relevant to a task domain, but not to a specific subject domain, suggests that the observed behaviors are at least close to realism, although we cannot say that the results can be generalized without careful consideration of the task format and/or user group. In this section we discuss the major points demonstrated in the results, their implications for IR system design, and future research directions.

The examination of novice and knowledgeable users' behavioral variables showed that only one behavioral variable on its own, at the MID and END points, was

DOI: 10.1002/asi

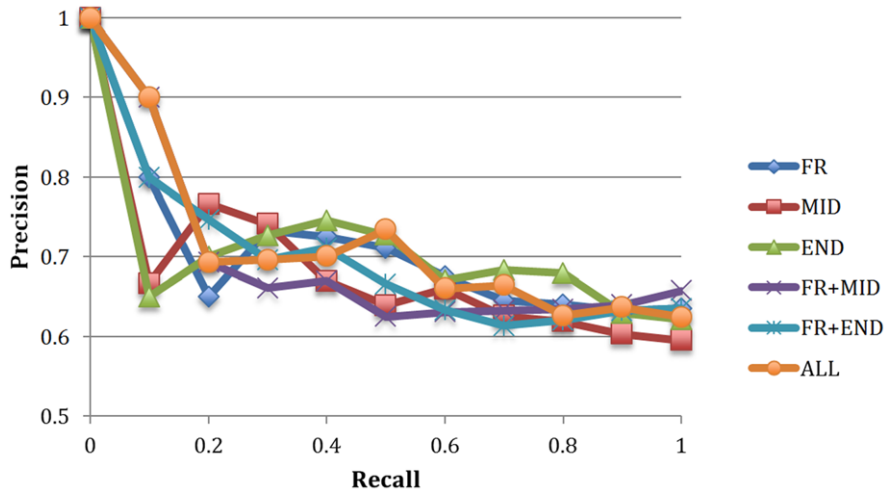

FIG. 3. Recall-precision graph of the six models. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

TABLE 7. AUC (area under curve) of each model.

	Baseline	FR	MID	END	FR + MID	FR + END	ALL
AUC	0.5	0.60	0.57	0.62	0.56	0.58	0.59

TABLE 8. FA model variables and B, p values.

Predictors	B	p
First dwell time on first SERP	4.936	0.074
First dwell time on first viewed document	–0.975	0.337
First query length	–0.936	0.473
Number of SERPs at first round	0.929	0.346
Constant	–0.572	0.236

significantly different between knowledge level groups: mean first dwell time on SERPs. It is reasonable that on average, users with low knowledge spent a longer time when first viewing the SERPs, in that they may have needed a longer time to determine which item(s) in the result list may be useful for them to click in. No other aspects of search behaviors about querying, number of documents, etc. showed significant differences. This indicates that dwell time on SERPs is a more important factor than others when trying to identify users with different levels of topic knowledge.

The prediction model results show that the early session FR model (the model using first-round measures) had decent performance, with the best accuracy, precision, and $F(\beta = 0.5)$ scores of all models, and is comparable with the FR + END model overall when the recall value varies. When looking closely at the variables selected for the combination models, that is, FR + MID, and ALL, we noticed that the FR variables were most often included in the models. Specifically, the three FR variables: first dwell time on the first SERP, first dwell time on the first document, and first query length appeared in all combination models. This seems to tell us that the early session behavioral variables are factors that are good enough to predict users' topic knowledge at a reasonable level.

On the other hand, the MID model performed poorly both on the performance measures (precision, accuracy, and $F(\beta = 0.5)$) and as shown in the recall-precision graph. This demonstrates that, in general, search behaviors in the middle session of search are not as good predictors of users' topic knowledge as in early in the search session. This result makes sense because we could expect that users' topic knowledge, as elicited before they work with the tasks, may influence users' search behaviors more in the beginning of the search than at the later stages of the search, as their topic knowledge may change as users proceed to a later stage of their search. Meanwhile, the whole-task level behaviors are better at representing the total amount of effort that users have devoted to searching. In summary, the prediction performance of the FR model indicates that it is not only feasible, but also promising, to predict searchers' knowledge levels during the search, early in their search session, from their search behaviors.

It is noted that generally speaking, the performance of all prediction models is relatively low, with the best accuracy, precision, and $F(\beta = 0.5)$ scores to be 0.64, 0.65, and 0.69 (in the FR model). This may be because of, we think, the fact that in general, search behaviors between novice and knowledgeable users did not show significant differences, as indicated in Table 4. In addition, the maximum number of variables used in the prediction models was six, with a minimum only one. Nevertheless, the FR model did improve over the baseline model, indicating that predicting users' knowledge level using the method presented in this article is

DOI: 10.1002/asi

beneficial. Future research can be conducted on larger data sets, which may enable the use of more behavioral variables as predictors, which could hopefully obtain better prediction performance.

Because of the limitation of the sample size, the current study did not examine the models' prediction performance on different types of tasks. With the same data set, Liu, Belkin, et al. (2013) and Liu, Liu, et al. (2013) found that on some types of tasks, more behavioral variables showed significant differences, and on others, fewer behavioral variables showed differences between the high and low knowledge users. Considering task types, there is a good chance that the prediction performance might lead to better results on some task types than the results obtained in the current study. Teevan, Dumais, and Liebling (2008) have suggested that applying personalization only on the queries that can benefit from personalization, instead of on all queries, improves search experience. Liu, Belkin, et al. (2013) and Liu, Liu, et al.'s (2013) results, and ours, support that conclusion.

Our research used the MID point to represent the in-session prediction. We think that selecting the middle query in the total query sequence appropriately dealt with our constrained data set, which had varying numbers of queries in different sessions. One limitation of our MID point determination method is that it is not possible to know which is the middle query during a search session. Nevertheless, our results showed that the MID performance is not as good as the FR model, making it not an important issue in this regard. We also realize that another way to handle in-session prediction, would be to conduct the prediction for a sequence of query run numbers: for example, first, second, third, and so on. The size of our data set prevented us from doing this analysis; we intend to address this is in future studies, with larger-size data sets.

One note about topic knowledge is that we elicited the users' base knowledge before they searched for the task. It is acknowledged that users' knowledge will or could change during the search. Because evolving topic knowledge is not easy to elicit, especially in the process of searching, our prediction models still have practical implications on designing systems that can monitor user behaviors and predict users' topic knowledge.

Another issue is that there is likely to be some type of interaction between topic knowledge level and task difficulty level, as shown in Liu et al. (2012), which found an interaction effect between subject domain knowledge and task difficulty on dwell time of content pages and the percentage of dwell time on content pages to the task completion time. This is not the main focus of the current study; nevertheless, it would be an interesting future direction to predict topic knowledge in relation to task difficulty level.

Conclusions

This research examined information search behavioral differences between users with high versus low levels of search task topic knowledge. Logistic regression models were built to predict, using search behavioral variables, which users were novices to the specific search tasks. Although none behavioral variable in the first query round of the search sessions showed significant differences between novice and knowledgeable users, the early-session prediction model FR, that considers multiple behaviors in the first query interval, had good prediction performance. Specifically, the FR model achieved the best performance compared to other models, as well as large improvements over the baseline model in all evaluation measures of precision, accuracy, and $F(\beta = 0.5)$. Our results indicate that using a limited number of early session behavioral variables could predict users' topic knowledge fairly decently. Future research could build prediction models with larger data sets taking account of different task types, which can hopefully receive even better prediction performance. Future research could also attempt to build and compare prediction models during search sessions after each query interval. The method used in the current research could be applied in personalized IR system design, in order to detect novice users and provide them with personalized search assistance accordingly, or to re-rank search results according to predicted knowledge levels. Furthermore, general inference of users' topic knowledge during search would enhance the interpretation of such behaviors as dwell time in predicting document usefulness.

Acknowledgments

Our thanks to IMLS for sponsoring the research experiment under grant number LG#06-07-0105-05. We thank all PooDLE members for their contributions in collecting and preparing the data used in this article.

References

Allen, B.L. (1991). Topic knowledge and online catalog search formulation. The Library Quarterly, 61(2), 188–213.

Belkin, N.J. (2008). Some(What) grand challenges for information retrieval. SIGIR Forum, 42(1), 47–54.

Bierig, R., Cole, M., Gwizdka, J., Belkin, N.J., Liu, J., Liu, C., Zhang, J., & Zhang, X. (2010). An experiment and analysis system framework for the evaluation of contextual relationships. In Doan, Jose, Melucci & Tamine-Lechani (Eds.), Proc. of the 2nd Workshop on Contextual Information Access, Seeking and Retrieval Evaluation (pp. 5–8). Milton Keynes, UK: Milton Keynes. March 28, 2010.

Borlund, P. (2003). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. Information Research, 8(3), 289–291. paper no. 152.

Cole, M., Zhang, X., Liu, J., Liu, C., Belkin, N.J., Bierig, R., & Gwizdka, J. (2010). Are self-assessments reliable indicators of topic knowledge? In C. Marshall, E. Toms, & A. Grove (Eds.), Proceedings of the Annual Conference of the American Society for Information Science & Technology (ASIS&T) 2010 (pp. 30:1–30:10). Silver Springs, MD: ASIS&T.

Duggan, G.B., & Payne, S.J. (2008). Knowledge in the head and on the web: Using topic expertise to aid search. SIGCHI ′08, 39–48.

Feild, H., Allan, J., & Jones, R. (2010). Predicting searcher frustration. SIGIR ′10, 34–41.

Fox, S., Karnawat, K., Mydland, M., Dumais, S., & White, T. (2005). Evaluating implicit measures to improve Web search. ACM Transactions on Information Systems (TOIS), 23(2), 147–168.

DOI: 10.1002/asi

Guo, Q., Yuan, S., & Agichtein, E. (2011). Detecting success in mobile search from interaction. SIGIR ′11, 1229–1230.

Hembrooke, H.A., Granka, L.A., Gay, G.K., & Liddy, E.D. (2005). The effects of expertise and feedback on search term selection and subsequent learning. Journal of the American Society for Information Science and Technology : JASIST, 56(8), 861–871.

Hosmer, Jr., D.W., Lemeshow, S., & Sturdivant, R.X. (2013). Introduction to the Logistic Regression Model, in Applied Logistic Regression, Third Edition, John Wiley & Sons, Inc., Hoboken, NJ, USA.

Hsieh-Yee, I. (1993). Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. Journal of the American Society for Information Science. American Society for Information Science, 44(3), 161–174.

Kelly, D. (2006). Measuring online information-seeking context, part 2. Findings and discussion. Journal of the American Society for Information Science and Technology : JASIST, 57(14), 1862–1874.

Kelly, D., & Cool, C. (2002). The effects of topic familiarity on information search behavior. JCDL ′02, 74–75.

Kuhlthau, C.C. (1991). Inside the search process: Information seeking from the user's perspective. Journal of the American Society for Information Science and Technology : JASIST, 42(5), 361–371.

Kumaran, G., Jones, R., & Madani, O. (2005). Biasing Web search results for topic familiarity. In Schek, H-J., Fuhr, N., Chowdhury, A., & Teiken, W. (Eds.), Proceedings of CIKM "05 (pp. 271–272). October 31–November 5, 2005, Bremen, Germany.

Li, Y., & Belkin, N.J. (2008). A faceted approach to conceptualizing tasks in information seeking. Information Processing & Management, 44(6), 1822–1837.

Li, Y., & Belkin, N.J. (2010). An exploration of the relationships between work task and interactive information search behavior. Journal of the American Society for Information Science and Technology : JASIST, 61(9), 1771–1789.

Liu, C., Liu, J., Cole, M., Belkin, N.J., & Zhang, X. (2012). Task difficulty and domain knowledge effects on information search behaviors. In Chang, S.L., Fulton, C., Hersberger, J., & Grove, A. (Eds.), Proceedings of the Annual Conference of the American Society for Information Science & Technology (ASIS&T) 2012 (pp. 1–10). Baltimore, MD, USA.

Liu, C., Liu, J., & Belkin, N.J. (2014). Predicting search task difficulty at different search stages. In Li, J., Wang, X. S., Garofalakis, M., Soboroff, I., Suel, T., & Wang, M. (Eds.), Proceedings of CIKM 2014 (pp. 569–578). New York: ACM.

Liu, J., & Belkin, N.J. (2010). Personalizing information retrieval for multi-session tasks: The roles of task stage and task type. SIGIR ′10, 26–33.

Liu, J., Gwizdka, J., Liu, C., & Belkin, N.J. (2010). Predicting task difficulty for different task types. ASIST ′10.

Liu, J., Cole, M., Liu, C., Bierig, R., Gwizdka, J., Belkin, N.J., Zhang, J., Zhang, X. (2010). Search behaviors in different task types. JCDL ′10.

Liu, J., Belkin, N.J., Zhang, X., & Yuan, X. (2013). Knowledge change in multi-session search tasks. Information Processing & Management, 49(5), 1058–1074.

Liu, J., Liu, C., & Belkin, N.J. (2013). Examining the effects of task topic familiarity on searchers" behaviors in different task types. In A. Grove (Ed.), Proceedings of ASIS&T 2013 (pp. 1–10). Silver Springs, MD: ASIS&T.

O'Brien, R.M. (2007). A caution regarding rules of thumb for Variance Inflation Factors. Quality & Quantity, 41(5), 673–690.

Sihvonen, A., & Vakkari, P. (2004). Subject knowledge improves interactive query expansion assisted by a thesaurus. The Journal of Documentation; Devoted to the Recording, Organization and Dissemination of Specialized Knowledge, 60(6), 673–690.

Tabachnick, B.G., & Fidell, L.S. (2012). Using multivariate statistics (6th ed.). Boston, MA: Pearson.

Teevan, J., Dumais, S.T., & Liebling, D.J. (2008). To personalize or not to personalize: Modeling queries with variation in user intent. In Chua, T.-S., Leong, M.-K., Myaeng, S.H., Oard, D.W., Sebastiani, F. (Eds.), Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 163–170). New York: ACM.

Vakkari, P. (2001). A theory of the task-based information retrieval. The Journal of Documentation; Devoted to the Recording, Organization and Dissemination of Specialized Knowledge, 57(1), 44–60.

Vakkari, P., & Hakala, N. (2000). Changes in relevance criteria and problem stages in task performance. The Journal of Documentation; Devoted to the Recording, Organization and Dissemination of Specialized Knowledge, 56(5), 540–562.

Vakkari, P., Pennanen, M., & Serola, S. (2003). Changes of search terms and tactics while writing a research proposal: A longitudinal research. Information Processing & Management, 39(3), 445–463.

Van Rijsbergen, C.J. (1979). Information retrieval (2nd ed.). London: Butterworth.

White, R.W., & Dumais, S.T. (2009). Characterizing and predicting search engine switching behavior. CIKM "09, 87–96.

White, R.W., Ruthven, I., & Jose, J.M. (2005). A study of factors affecting the utility of implicit relevance feedback. In Baeza-Yates, R., Ziviani, N., Marchionini, G., Moffat, A., & Tait, J. (Eds.), Proceedings of 28th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR 2005) (pp. 35–42). New York, NY: ACM.

White, R.W., Dumais, S., & Teevan, J. (2009). Characterizing the influence of domain expertise on Web search behavior. WSDM ′09, 132–141.

Wildemuth, B. (2004). The effects of domain knowledge on search tactic formulation. Journal of the American Society for Information Science and Technology : JASIST, 55(3), 246–258.

Zhang, X., Anghelescu, H.G.B., & Yuan, X. (2005). Domain knowledge, search behavior, and search effectiveness of engineering and science students. Information Research, 10(2), retrieved from http://www.informationr.net/ir/10-2/paper217.html.

Zhang, X., Liu, J., & Cole, M.J. (2013). Task topic knowledge vs. background domain knowledge: Impact of two types of knowledge on user search performance. Advances in Intelligent Systems and Computing, 206, 179–191.

Zhang, X., Liu, J., Cole, M., & Belkin, N.J. (2015). Modeling domain knowledge from searchers" behaviors using multiple regression analysis. Journal of the Association for Information Science and Technology, 66(5), 980–1000.

DOI: 10.1002/asi