Personalizing Information Retrieval for Multi-session Tasks: Examining the Roles of Task Stage, Task Type, and Topic Knowledge on the Interpretation of Dwell Time as an Indicator of Document Usefulness

Jingjing Liu
School of Library and Information Science, Davis College, University of South Carolina, 1501 Greene Street, Columbia, SC 29208. E-mail: jingjing@sc.edu

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

Personalization of information retrieval tailors search towards individual users to meet their particular information needs by taking into account information about users and their contexts, often through implicit sources of evidence such as user behaviors. This study looks at users' dwelling behavior on documents and several contextual factors: the stage of users' work tasks, task type, and users' knowledge of task topics, to explore whether or not taking account contextual factors could help infer document usefulness from dwell time. A controlled laboratory experiment was conducted with 24 participants, each coming 3 times to work on 3 subtasks in a general work task. The results show that task stage could help interpret certain types of dwell time as reliable indicators of document usefulness in certain task types, as was topic knowledge, and the latter played a more significant role when both were available. This study contributes to a better understanding of how dwell time can be used as implicit evidence of document usefulness, as well as how contextual factors can help interpret dwell time as an indicator of usefulness. These findings have both theoretical and practical implications for using behaviors and contextual factors in the development of personalization systems.

Introduction

As the amount of information on the web grows dramatically, it becomes increasingly difficult for information searchers to find documents that meet their particular needs. The traditional one-size-fits-all approach that search systems return the same results to everyone who issues the same query has been replaced by "personalizing" results for specific users, which takes into consideration the users' particular information needs and their contexts. For example, many search engines now take into account a person's location or previous search results in retrieving and ranking search results. Nevertheless, personalization is still in its early phase and many other factors can be used for personalization such as task, user background, etc., as Belkin (2008) noted. Continuous exploration and research in this area are needed.

In learning about users' situations, personalization systems often adopt an implicit approach so as to avoid bothering and interrupting users during their search. They usually infer user interests from monitoring their search behaviors and/or search contexts, such as dwell time, click-through, browsing history, and query history. As a salient behavior that can be easily captured, dwell time has gained research attention, but previous studies (e.g., Morita & Shinoda, 1994; Kelly & Belkin, 2004) have produced different findings on whether or not the length of dwell time indicates the usefulness of web pages. In particular, Kelly and Belkin (2004) suggested that contextual factors should be taken into account in understanding the relationship between dwell time and document usefulness. Based on this background, the current study hopes to contribute to personalization research by taking a careful examination of dwell time with regard to whether, and how it can be an indicator for systems to predict document usefulness. The study

Received February 12, 2013; revised October 15, 2013; accepted October 16, 2013

© 2014 ASIS&T • Published online 21 April 2014 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.23160

considers the effects of three contextual factors that have not yet been much studied in personalization but are very likely to help interpret dwell time as an indicator of user interest. They are: (a) the stage in a work task; (b) task type; and (c) the knowledge that the user has of task topic. In other words, this study seeks to explore how these search contextual factors may help implicitly infer document usefulness, in particular examining the possible interaction effects of these different characteristics and the users' dwelling behaviors on performing personalization.

Related Work

Contextual/Situational Factors in Information Retrieval (IR) Studies

Over the past two decades or so, the concepts of context and situation have been brought into the foreground of information science research. However, context is a term that is most "often used," least "often defined," and "when defined so variously" (Dervin, 2003, p. 112), and the concepts of context and situation have often been used interchangeably (Cool, 2001). Based on a thorough review of the concept of situation across six disciplines, as well as the distinction between context and situation, Cool (2001) suggests that "contexts are frameworks of meaning, and situations are the dynamic environments" (p. 8), or more simply, "situation is the dynamic aspect of context" (p. 31). Cool (2001) further concludes that situation has the potential for being an important unit variable and should be the focus of analysis in information science research. Despite this clear description that disambiguates situation from context, the following more than a decade has continued to see the interchangeable use of context and situation, and perhaps a more extensive use of the term context when more and more single studies tend to take into account multiple factors. As Allen (1996) points out, context is not a single thing, but rather is a composite of things comprised of a number of elements or aspects. In this sense, using context as an umbrella term to refer to a variety of factors could simplify things for a certain purpose. Due to these reasons, the following literature review does not attempt to differentiate context and situation, but uses the term context in general, unless specified.

Contextual factors have been addressed by many researchers as important in IR research and system design (e.g., Belkin, 1993; Ingwersen & Järvelin, 2005; Kelly, 2006a; Dumais, 2007). Task as a contextual factor has attracted fairly extensive research efforts with regard to the effect of different task features on information search and use (e.g., Byström & Järvelin, 1995; Byström, 2002; Vakkari, 1999, 2001). User knowledge of a subject domain (e.g., Wildemuth, 2004; White, Dumais, & Teevan, 2009) or of a search task topic (e.g., Kelly & Cool, 2002; Liu, Liu, & Belkin, 2013) has also attracted attention with respect to their effects on information-seeking behaviors. However, it is only in recent years that research findings on all these aspects have been effectively used, or been proposed to be used in the design of operational systems that tailor search results toward individual users. The following section reviews related studies in two main aspects of context: the user's task and the user's topic knowledge. All studies reviewed here belong to the "objectified context" camp (Talja, Keso, & Peitilainen, 1999), treating context as an objective reality which provides a background for the study of an individual's or a group's behaviors.

Task and Personalization

Task and task classification. In a problematic situation, the activities by which one attempts to keep one's work or life moving on are often called "tasks" (Ingwersen & Järvelin, 2005). Various types of tasks routinely examined in the information science field include search tasks and work tasks. A search task usually refers to a user's search for information through his/her interactions with information systems, and a work task is an activity one performs to fulfill the responsibility for one's "work" (Ingwersen & Järvelin, 2005; Li, 2008). The relationship between the two types of tasks is that work tasks are often motivations for search tasks. It should be noted that tasks that drive people to engage in information seeking are not restricted to those which are strictly work-related, but include various sorts of nonwork information-seeking activities in individuals' everyday lives. For instance, everyday life information seeking (ELIS) has been attracting increasing research attention. Previous studies in this area have investigated aspects of seeking orienting information from media (Savolainen, 1995, 2007), planning for a vacation trip (Lin, 2001), and others like shopping, weather, transportation, etc. (Agosto & Hughes-Hassell, 2005). Such a phenomenologically informed approach provides novel ideas for IR research. It helps clarify the preference and relevance criteria for information seeking by extending the evaluation base from the narrower search task to the broader context of people's everyday lives, which may be more suitable in the situation of interactive IR (Belkin, Cole, & Liu, 2009).

Researchers have spent a fairly extensive amount of effort examining the effects of different tasks on information searchers' behaviors and performance. A common approach is to classify user tasks into different types along some task feature(s). These include, for example: closed versus open-ended tasks (Marchionini, 1989); specific versus general tasks (Qiu, 1993); factual, descriptive, instrumental, and exploratory tasks (Kim, 2006); fact-finding versus information gathering (Toms et al., 2007; Kellar, Watters, & Shepherd, 2007); and learning about a topic, making a decision, finding out how to, finding facts, and finding a solution (Freund, 2008). The various standards and definitions of task classification make it difficult to compare findings across studies. This makes it necessary to have some standard classification schemes. A rather comprehensive classification scheme is provided by Li and Belkin (2008), which includes a number of dimensions: task product, objective complexity,

DOI: 10.1002/asi

subjective complexity, and difficulty, to name a few. This scheme has been used in a number of studies, which showed the effects of these task features on users' behaviors (e.g., Li & Belkin, 2010; Liu et al., 2010).

Stage of task. Stage of task has received careful investigation regarding the information seeker's affective, cognitive, and physical action changes during the information-seeking process. Li and Belkin (2008) did not include this as a dimension, and this is a reasonable expansion to their classification scheme.

Kelly's (1963) construct theory depicts the process of construction as occurring in six different phases when individuals build their view of the world by assimilating new information: confusion, doubt, threat, hypothesis testing, assessing, and reconstructing. Taylor (1968, 1986) describes four levels of information need along the different stages of search: visceral, an actual but unexpressed need for information; conscious, a within-brain description of the need; formalized, a formal statement of need; and compromised, the question as presented to the information system. In her information-seeking process (ISP) model, Kuhlthau (1991) proposes six stages, including initiation, selection, exploration, formulation, collection, and presentation. The user's feelings, thoughts, and actions vary along the different stages. This body of research indicates that stage of task may be an important factor that relates to the user's judgment of document usefulness.

Vakkari and colleagues, in a series of papers (e.g., Vakkari, 2001; Vakkari & Hakala, 2000), describe their research on the relationship between a user's stage in accomplishing a task and his/her search tactics and the relationship between task stage and relevance assessments. Through a study which involved 11 masters students preparing a research proposal and engaging in IR search three times, in the beginning, middle, and end points during the course, it was found that the user's vocabulary changed from broader to narrower terms. As the task stage progressed, the users were less likely to start their initial queries by introducing all the search terms, were more likely to enter only a fraction of the terms, and tended to use more synonyms and parallel terms. In terms of the relevance criteria, the results supported the overall hypothesis that the user's relevance criteria depend on the stage of his/her task performance process. These findings shows the differences along stages and sheds lights on designing personalization systems that could provide tools helping users build their conceptual structure in the initial stage of tasks. However, it should be noted that the authors did not show statistical significance of the changes in relevance criteria. In another study, Taylor, Cool, Belkin, and Amadio (2007) did find statistically significant relationships between the users' stages in the search process and relevance categories chosen. This demonstrates the differences in user's relevance judgments in different stages of the task; however, it leaves open how such differences could be modeled through the user's behaviors.

One issue to note about task stage is that it is not always easy to split the stages exactly and accurately in empirical research because stages do not often or necessarily have obvious borders or lines, especially when the user is not involved in explicitly expressing such stages. Besides the method used by Vakkari and colleagues, that is, taking different points during a course as different stages, two other strategies have been used in related literature. One way is to split stages equally by time period. For example, in their study analyzing the effect of implicit relevance feedback (IRF), White, Ruthven, and Jose (2005) divided tasks, based on the logged user-system interaction data, into three stages with equal time length: "start," "middle," and "end." They found that IRF is used more in the middle of the search than at the beginning or end, whereas explicit relevance feedback (ERF) is used more toward the end. In their study exploring how searchers' criteria on web pages' relevance evolved in the different stages of tasks, Tombros, Ruthven, and Jose (2004) identified the stages in the users' task progress by identifying the first and last sets of web documents that the users visited. The study found that the users' relevance criteria during a task displayed a certain degree of variation, especially for tasks for which the users had a higher perception of task completion. Another way to operationalize task stage was used by Lin (2001), which manipulates the user's task with different subtasks to be completed in different search sessions. In his study, a task scenario is designed which requires the participants to finish a task that involves making a vacation plan. This plan is accomplished through three steps/sessions: identify candidate places for the trip, compare the different candidate places and decide on one place to go, and to make a plan for the trip. All these means to operationalize task stage are arbitrary to some extent, but the Lin (2001) approach seems closest to the situation in people's daily lives when solving complex tasks.

Task structure. If the work task consists of multiple subtasks, the relationship between the subtasks has to be taken into account because the task-doer could take different orders of these subtasks during the process of accomplishing the work task. This dimension was not included in Li and Belkin's (2008) classification scheme, but a similar idea can be found in other works. For example, Toms et al. (2007) classify tasks based on their conceptual structures. The two types of tasks in their approach are: the parallel, where the search uses multiple concepts that exist on the same level in a conceptual hierarchy, and the hierarchical, where the search uses a single concept for which multiple attributes or characteristics are sought. This opens a way to extend Li and Belkin (2008) by adding to their classification scheme a new dimension of task structure.

Summary. As can be seen from the review, there is a rich body of literature on task in terms of how it affects users' information-search behaviors and performance. However, further research efforts are needed. For instance, previous studies typically concluded by finding behavioral or performance differences among users performing different types of tasks, but lacked further approaches that made use of these findings in developing systems to personalize retrieval. Another reason for further research is that a majority of previous studies looked at behavioral or performance variables on a whole task-session level, such as time spent to complete the whole task, total number of queries, total number of pages viewed and saved, and effectiveness (recall, precision) or efficacy (number of saved documents out of all viewed) of the search. These variables cannot be obtained until the end of a session. While these results can in general be used to predict task type a posteriori, it is not easy to make use of these findings in designing adaptive search. Lower-level behavioral variables that can be captured and used in real time are needed, for example, document dwell time, number of pages per query, etc.

Task, Search Behavior, and Document Usefulness

The reviewed studies mostly concern a two-way relationship between tasks and user behaviors, or between tasks and search performance. No consideration was made with respect to the usefulness of the documents with which the users interacted. Document usefulness is an important element in the search system. Not only do systems want to return useful documents for the users at top ranks, but systems can also learn user interest from useful documents and extract significant terms from them for query expansion, helping the users find what they need more efficiently. Being able to predict document usefulness based on user behaviors would benefit IR systems in personalizing search. There have been previous studies looking at another two-way relationship between document usefulness and user behaviors including document reading time, scrolling, etc. (e.g., Morita & Shinoda, 1994; Kelly & Belkin, 2001). These studies, with different experimental settings, have generated seemingly conflicting findings concerning the relationship between document reading time and preference/relevance judgment. While attempting to design a filtering system based on monitoring user behaviors, Morita and Shinoda (1994) found a strong tendency for users to spend a greater amount of time reading those articles rated as interesting than those rated not interesting. In a different setting which was interactive in nature, asking the users to perform search tasks, Kelly and Belkin (2001) found that the length of time that a user spent viewing a document was not significantly related to the user's subsequent relevance judgment. Kelly and Belkin (2004) further suggest that contextual factors should be taken into account in interpreting the evidence of user behavior, which leads to a three-way relationship among document usefulness, user behavior, and contextual factors.

Generally speaking, in examining such a three-way relationship, task has been found to be helpful in predicting document usefulness from the user's behaviors, such as dwell time (or display time, i.e., the time that a user spends on a retrieved information object). Kelly and Belkin (2004) found that using display time averaged over a group of users to predict document usefulness is not likely to work, nor does it work using display time for a single user without taking into account contextual factors. Specifically, display time on relevant and nonrelevant objects differed significantly according to specific tasks and specific users. This demonstrated that inferring the usefulness of a document from dwell time should be tailored toward individual tasks and users. This study, however, did not examine how to incorporate the contextual factors and what the actual effectiveness would be.

The issue of incorporating contextual factors into prediction of usefulness based on dwell time was addressed by White and Kelly (2006). They explored the interaction effects between dwell time and the two factors of user and task on document usefulness. They examined whether additional information from the user and/or the task helps reliably establish a dwell time threshold to predict document usefulness, and how effective this method would be. They found that tailoring the display time threshold based on task type information improved an implicit relevance feedback algorithm performance. In other words, display time was shown to be able to successfully predict document usefulness when task information is considered. This study is a successful attempt examining the interaction effect of contextual factors and display time in predicting document usefulness.

Nevertheless, there are still research problems calling for further efforts. In White and Kelly's (2006) approach to classifying tasks, they collapsed the different everyday life tasks identified by their seven participants, according to the task contents, into several categories such as online shopping, emailing, researching, etc. However, the different tasks cannot be more effectively used for personalization in a more general sense unless they are classified into some common types according to a certain generic features. Such efforts could be conducted following some task classification or ontology; for example, the task classification scheme of Li and Belkin (2008).

Topic Knowledge and Personalization

Another contextual factor that is potentially helpful in providing additional user interest information for personalization is the user's knowledge. The literature on IR has seen two different types of knowledge being studied regarding their effects on users' search behaviors: one is domain knowledge, and the other is topic knowledge. These two types of knowledge have been shown to be different in affecting users' search behavior (Zhang, Liu, & Cole, 2013). What is relevant to the current paper is topic knowledge (i.e., topic familiarity; topic knowledge and topic familiarity are used interchangeably in this article), and this section reviews the literature on topic knowledge and personalization.

In looking at users' topic knowledge and their search behaviors, the examined behaviors often include document features related to reading behaviors, dwell time, the ratio of saved to all viewed documents, etc. Hembrooke, Granka, Gay, and Liddy (2005) found that experts with high topic knowledge issued longer and more complex queries than novices. They also used elaboration as a reformulation strategy more often as compared to simple stemming and backtracking modifications used by novices. Allen (1991) found that compared with their counterparts, people with high topical knowledge used more search expressions, and employed more search expressions that had not been contained in their statement of information need. Kelly and Cool (2002) found that an increase in familiarity with topics, reading time tended to decrease and efficacy, measured by the ratio of the number of saved documents to the total number of viewed documents, increased. These results indicate that it may be possible to infer topic familiarity implicitly from search behaviors. However, further efforts are needed in order to tell which specific documents may be predicted as useful based on topic knowledge and reading time, the user's saving, viewing, and other behaviors. Kelly (2006b) found that user topic familiarity, as a contextual factor, had a significant effect on user behaviors, specifically, document display time.

Using a different approach, Kumaran, Jones, and Madani (2005) attempted to differentiate documents that match different levels of topic familiarity by document features. They defined two types of web pages: *introductory* web pages, which do not presuppose their readers to have any background knowledge of the topic and may introduce or define key terms in the topic; and *advanced* web pages, which assume their readers have sufficient background knowledge and familiarity with the key technical/important terms in the topic. A classifier was built to classify the documents according to different features (e.g., stop-word, line-length) that could predict assumed topic familiarity. An experiment to re-rank search results for people with lower topic familiarity showed that the classifier was effective: the proportion of introductory pages at top 5 and top 10 result lists using this method was significantly higher than in a baseline run using default search engine ranking. Their method could be effective in biasing result ranking for topic familiarity when it is known. Their study indicated that certain features of the document 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 familiarity with the topic. This could be useful in implicitly inferring one's topic familiarity, which would accordingly help in personalization system design that takes account of the user's topic familiarity when it is not explicitly known.

While these studies mostly concerned the two-way relationships between knowledge and user behaviors, there is a need to examine the three-way relationship among knowledge (as a contextual factor), user behaviors, and document usefulness. In their proposed user modeling system that accounts for contextual factors, Kelly and Belkin (2002) addressed the user's topic familiarity. They pointed out that topic familiarity might affect the types of information search and behaviors exhibited by the user. They illustrated the likely ways in which topic familiarity might affect a user's reading time on a document: the relationship between reading time of relevant and nonrelevant documents is not simply linear; rather, it could vary in two very different ways according to topic familiarity. For those with a low degree of familiarity, reading time for both relevant and nonrelevant documents may be similar, but for those with high degree of familiarity, reading time for relevant and nonrelevant documents may be very different. Their concept makes intuitive sense, but there has been no further research hypothesis developed or effort spent on verifying this type of relationship in a systematic way.

Theoretical Stance

Research Model

Personalizing IR requires understanding of the user's goal (or task), context, and search behaviors within the current search episode. This sets up a three-way relationship among document usefulness, user behaviors, and contextual factors, toward achieving a certain goal. We propose a model (Figure 1) involving key factors in an IR episode that address basic relationships and interactions among these factors, which could provide significant evidence for personalization.

Four sets of elements are included in this model: goal, contextual variables, user behaviors, and document

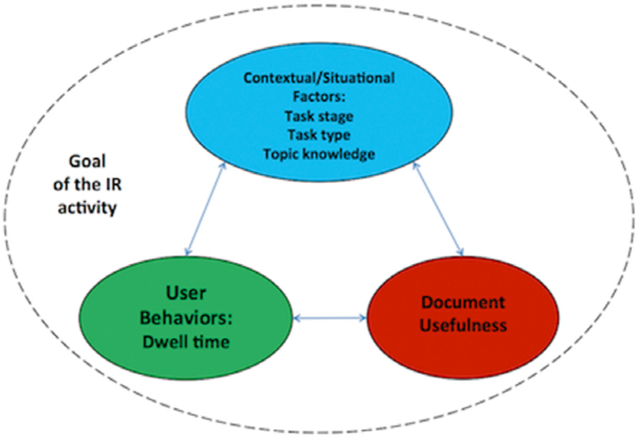

FIG. 1. A research model: Factors and relations in an IR activity. (Factors shown are those addressed in the paper, among others.) [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

usefulness. Given a certain goal that drives the user to engage in information search, there is a three-way relationship: one element is document usefulness, which is the core value that a personalization system tries to learn, infer, or predict; another element is behavioral information, which is what users do and which can be observed by systems; and the third element refers to contextual factors, which set and convey the background and contextual information about the users who are conducting information search.

The behaviors that have been studied in IR research, specifically IR personalization research, include querying, dwelling behavior (one measure is the duration of dwelling on a document, called dwell time), saving behavior, clicking, and using behavior. Behaviors in IR personalization can be understood in a two-fold sense. On the one hand, behaviors can be viewed as the concrete expressions that a user shows under his/her specific situation in the IR episode. For example, a user's dwell time on a retrieved information object may be a function of his/her level of domain knowledge, and his/her task features. On the other hand, behaviors are also sources for the IR system to learn about the user. Predictions of a user's preference, that is, a document's usefulness to the user, can be made according to the user's behaviors. For example, the user's dwell time on a document, or the saving or using behaviors, may tell, at least to some extent, how useful the document is to the user. In the latter sense, behaviors can have interactions with situational factors, and such interactions can also possibly help predict a document's usefulness. For instance, a user's dwell time on a document, together with the consideration of the user's knowledge, and/or task, may tell how useful this document is to him/her.

The contextual factors that were considered in the research reported here include the stage of task, task type (specifically, task structure), and topic knowledge. The review of the related literature on task stage showed that users' behaviors and cognitive status (e.g., relevance judgment criteria) did vary along different stages in the search task (e.g., Vakkari & Hakala, 2000; Taylor et al., 2007). Thus, it is of interest to see how information about task stage can be used for personalizing search results. One question of interest is whether or not task stage, as a contextual factor, can provide useful information for implicitly inferring a document's usefulness from some dwell time.

This study also considers different types of tasks classified by subtask relationships, that is, task structure. Using a method similar to Toms et al. (2007), two basic types of tasks are conceptualized: the parallel and the dependent. In some tasks, subtasks are parallel to one another, and the accomplishment of one is not necessarily dependent on the accomplishment of other subtasks. The knowledge needed for, and acquired after, one subtask does not necessarily contribute to the knowledge needed for subsequent subtasks, nor is it necessarily based on knowledge acquired in the conduct of previous subtasks. Accordingly, the order of the subtasks is not fixed, but rather can vary. These tasks are called parallel tasks in this study. On the other hand, some subtasks could be dependent on others, and the accomplishment of one is based on the accomplishment of others. In this case, the knowledge needed for, and acquired after, one search subtask is usually built on that of the previous ones. The order of the subtasks in such a task is usually fixed. These tasks are called dependent tasks in the study. This study considers only these two simple and basic types of subtask relationships for the purpose of easily detecting the effects (or differences of the effects) of these two types of tasks on search behaviors, and leaves other types of tasks to future studies.

A user's topic knowledge usually changes during the search process (cf. Kuhlthau's ISP model), which makes it appealing to see if, and how, one's topic knowledge can be used as a significant factor for personalization. In addition to the effect of topic knowledge on user behaviors, as has been studied in the literature, this study also looks at the interaction effect of the user's topic familiarity and dwell time on predicting document usefulness. This approach can add knowledge to the related literature concerning how to infer document usefulness from user's behaviors and contexts.

Specifically, this study attempts to answer the following research questions:

RQ1. Can dwell time be used as a reliable indicator of document usefulness?
RQ2. Does the stage of the user's task help in interpreting dwell time as an indicator of document usefulness?
RQ3. If the stage information helps in predicting document usefulness from dwell time, does its role vary in different task types?
RQ4. Does taking account of the user's topic knowledge help in interpreting dwell time as an indicator of document usefulness?
RQ5. If the topic knowledge information helps in predicting document usefulness from dwell time, does its role vary in different task types?

Method

Operationalization of Task Stage

Several ways have been used in the literature to operationalize task stage, each having different levels of arbitrariness and degrees of difficulty in dividing task activities. One is to divide a task episode into different stages according to the time elapsed, as is done by White et al. (2005). Another way is based on Kuhlthau's (1991) ISP theory that the information-search process is a six-stage model including initiation, selection, exploration, formulation, collection, and presentation. However, the search activity may not be easily, accurately, and necessarily divided into six stages. A third way can be seen in some tasks with clear subtask boundaries that can be easily divided into stages. A "complex" work task that includes subtasks usually requires task-doers to engage in many activities to accomplish it. Task-doers may not be able to finish the task at once due to the complexity of the task, the limitation of the task-doer's time, efforts, and knowledge. Therefore, such tasks may often be conducted in sessions.

Due to the ease of identifying task boundaries, the current study employed the third approach using different subtasks accomplished in different search sessions at different times. It should be noted that this approach does not necessarily contradict the ISP. In this case, for each subtask or the overall task, the user may still have gone through the six-stage ISP. The two methods just differ in that they are from different perspectives with different criteria for separating task stages.

Experimental Design

This experiment was conducted on a between-subjects model. Twenty-four participants were recruited, alternately assigned to one of the two three-session tasks, described below. In this experiment, the three sessions were treated as three stages in task completion.

Tasks

The study used journalists' assignments as tasks. The major reason for this was that they could be relatively easily set as realistic tasks in different subject domains; another was that we had access to a population at least somewhat experienced in journalism. As mentioned earlier, among the many facets in task type, this study focused on task structure. Task design varied the values of this facet only and kept those of others as constant as possible.

Two tasks were used in the study: one was a parallel task, in which the accomplishment of one subtask is not necessarily based on that of others; the other was a dependent task, in which the accomplishment of some subtask depends on that of others. They both had three subtasks, each of which was worked on by the participant during a single session, with the three sessions representing the three stages of the task. To maintain the consistency of other facets as much as possible, the design took into account the following considerations, which focused on two significant facets that have been demonstrated (e.g., Li & Belkin, 2010) to influence user's search behaviors: product and objective task complexity. First, the task product was set as intellectual for all three subtasks, specifically, each subtask asked the participants to submit a report, which by its nature embedded new ideas or findings (Li & Belkin, 2010). Second, the objective complexity of the two tasks was roughly the same, both being low complexity, using Li and Belkin's (2008) definition of task complexity. This meant that each subtask of the two tasks could be actually finished by searching only one type of information source. In addition, the two tasks were in the same domain. The two tasks are described as follows.

The parallel task. As a beat reporter for automobiles, you want to write a feature story about cost-effective cars, specifically, hybrid cars for low to mid-income level families. You want to focus on three models of cars from auto manufacturers that are famous for good warranties and fair maintenance costs, and the three models are: Honda Civic sedan, Toyota Camry sedan, and Nissan Altima sedan.

You want to write about the features of each of the three models, including aspects such as: standard features and specifications, safety, pricing, reviews, possible pictures, and so on. You have three sessions to finish this assignment, and you will need to finish the writing on one car in each session. At the final session, you will need to integrate the three reports.

The dependent task. As a beat reporter for automobiles, you want to write a feature story about cost-effective cars, specifically, hybrid cars for low to mid-income level families. To do it, you need to learn what makes and models have hybrid cars, what are their features, prices, and safety levels, etc. Specifically, you will need to accomplish the following activities:

Collect information on what manufacturers have hybrid cars. You want to list the different models that are good for mid-level income families. Select three models that you will mainly focus on in this feature story. You want to introduce their specific features that make you choose them out of other models.

Compare the pros and cons of three models of hybrid cars. You will have three sessions to finish this assignment. You will need to finish one activity in each session, but the order of the three sessions is up to you.

Subtask Orders

An assumption underlying the task description is that the subtask order in the parallel task is not fixed, while that in the dependent task is at least to some extent fixed. To maintain consistency, however, the experiment did not control the order of subtasks, but rather chose to let the participants determine the subtask orders that they wanted to follow. Subtask orders that appeared in the task descriptions were rotated following a Latin square design.

Participants

This study used journalists' assignments as work tasks, and accordingly the participants were recruited from those who had certain knowledge and skills to deal with such kinds of assignments. Participants in this study were recruited from the Journalism/Media Studies and Communication undergraduate student community in the School of Communication and Information (SC&I) at Rutgers University as a convenience sample. Recruitment was conducted through sending recruitment emails to student listservs and posting recruitment ads on post-boards in the SC&I building.

Each participant was invited to come to our lab to work on the assigned task three times within a 2-week interval, at their convenience. Each participant was given $30 upon completing the whole task. In order to encourage participants to work on the assignment in a serious manner, participants were told in advance that the six who submitted the most detailed reports on the assignment, as judged by an assessor, would each receive an additional $20.

Data Collection

User-system interactions, including mouse movement, keyboard activities, application displayed, web page opened, were logged by Morae software (http://www.techsmith.com/morae.asp). Various types of questionnaires were used, including: background questionnaire to elicit users' background information, pre- and postsession general task questionnaires to elicit their topic knowledge and perceived difficulty of the general tasks, pre- and postsession subtask questionnaires to elicit their topic knowledge and perceived difficulty of the subtasks, as well as a usefulness judgment questionnaire to elicit their usefulness judgment ratings on all viewed documents.

Experiment Procedure

Participants came individually to our interaction lab to take part in the experiment. Upon arrival in the first session, they completed a consent form and the background questionnaire. They were then given a form describing the general work task that they were assigned (either *Parallel* or *Dependent*) to be accomplished in the whole experiment. The presession task questionnaire specifying the topic was then administered, and the participants were then asked to pick one subtask to work with in the current session. The presession subtask questionnaire followed, after which participants were directed to IE 6.0 to work on that subtask: searching for useful sources and writing reports. They were given up to 40 minutes and were allowed to search freely on the web for resources in their report writing. For logging purposes, users were asked to keep only one IE window open and use back and forward buttons to navigate between web pages.

After report submission for the first session, participants went through the process of evaluating, on a 7-point scale, each document that they had viewed, in the order of viewing them in the actual search process, with respect to its usefulness to the overall task (the usefulness evaluation questionnaire). The postsession subtask questionnaire and a postsession general task questionnaire were then administered. This ended the first session.

In the 2nd and the 3rd sessions, participants went through the same processes except for the consent form and background questionnaire. In the 3rd session, after the postsession general task questionnaire, the exit interview asked them to reflect on their overall knowledge gain and to comment on the whole experiment.

Results

Participants' Characteristics

Of the 24 participants, 21 were female and 3 were male. There were 10 seniors, six juniors, and eight sophomores. Their ages varied between 18 and 23, with an average of 20.4 (standard deviation [SD] 1.3) years. On average, participants had 8.4 years (SD = 2.9) of online searching experience. They self-rated their levels of computer expertise as 4.6 (SD = 1.0), and levels of searching expertise as 5.4 (SD = 0.9), on a 7-point scale, 1 being for *novice* and 7 for *expert*.

Various Types of Time at the Document Level

At the document level, dwell time measures the time that a user spends on a retrieved document. In the current study, since users were asked to generate reports based on their information searching, they often wrote in a word processing program in parallel with searching for information on the web. Retrieved documents could be open for a long time but not always be active, especially when the users were writing. There is therefore a need to differentiate several types of time on the document level, which we term dwell time, display time, and decision time (Figure 2).

Dwell time. The time duration from each point when the user starts viewing a document (usually when a document is opened) to the point when the user leaves the document (the user may close the document, or he/she may leave the document while it is open and go to another application). Each dwell time is the time that a user dwells on the document, or in other words, that a document is active for the user to read. This is denoted "a" in Figure 2.

Total dwell time. The sum of all dwell times that a user interacts with a document.

Display time. The total duration of a document between when it is opened to when it is closed. This is the total time that the document remains open, no matter if it is active, that is, if the user views it or not, after it is opened. It is possible that a document was opened multiple times in an experiment session, so one document could have multiple display times at different points. This is denoted "d" in Figure 2.

Total display time. The sum of all display times for a document that is revisited during a session.

Decision time. The first dwell time a user spends in a document. This time is called decision time in the current study in the sense that by the end of this duration, the users would typically have made some internal decision on the usefulness (being useful or not, or to use the document or not) of the document. For example, opening a new document or going to an existing document and starting writing most likely means that the web page that the user has just viewed

DOI: 10.1002/asi

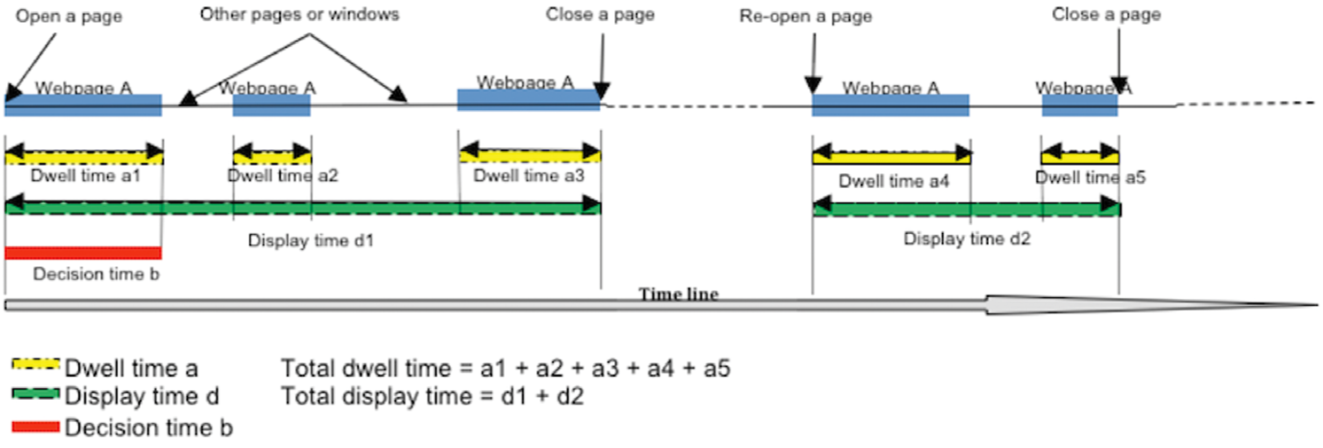

FIG. 2. Different types of time. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

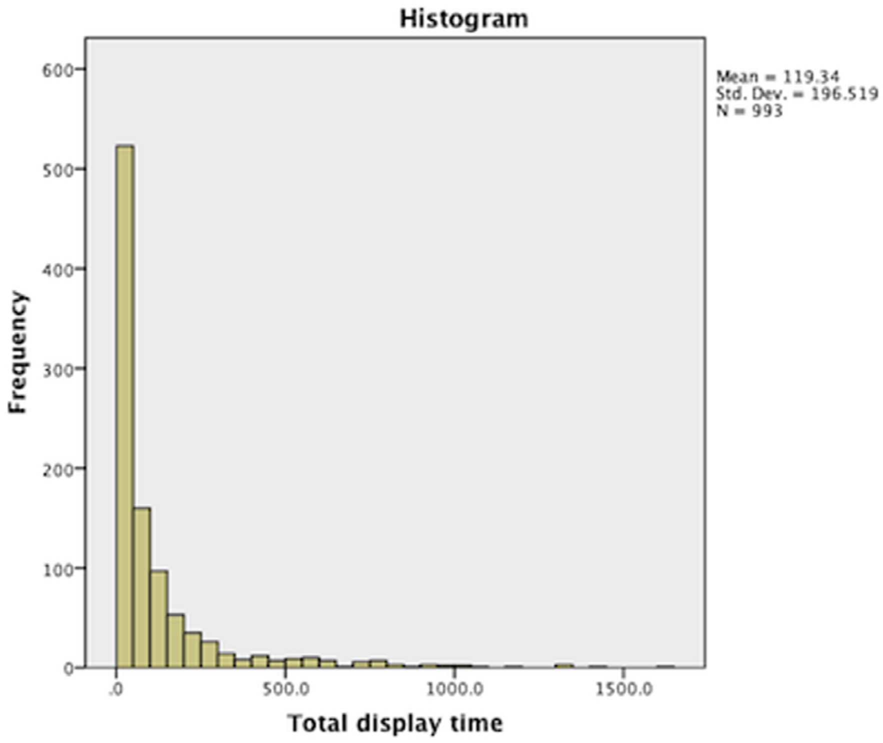

FIG. 3. The distribution of total display time in both tasks combined. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

was useful; leaving a web page and going back to the search result page (to refine queries or open another search result) perhaps means that the page just viewed was not useful. This is denoted "b" in Figure 2.

Among all the aforementioned types of time, total display time, total dwell time, and decision time best represent the features (e.g., usefulness) of a certain document across the whole session and so they were used for analysis.

Transformation of time. An exploration of the time distributions found that they are not normal. For example, Figure 3 shows the distribution of total display time in both tasks combined. In order to adjust these distributions and to improve the interpretability of results on relationships between factors, a logarithmic transformation was performed using the log base of 10, as has been previously done in the literature (e.g., Kelly, 2006b). Figure 4 shows the distributions of total display time under this transformation in both tasks combined, which were much more bell-shaped, even though some were not perfectly normal.

Usefulness Rating Scores and Grouping

Document usefulness in this study was obtained at the end of each session by asking the users to rate, based on a

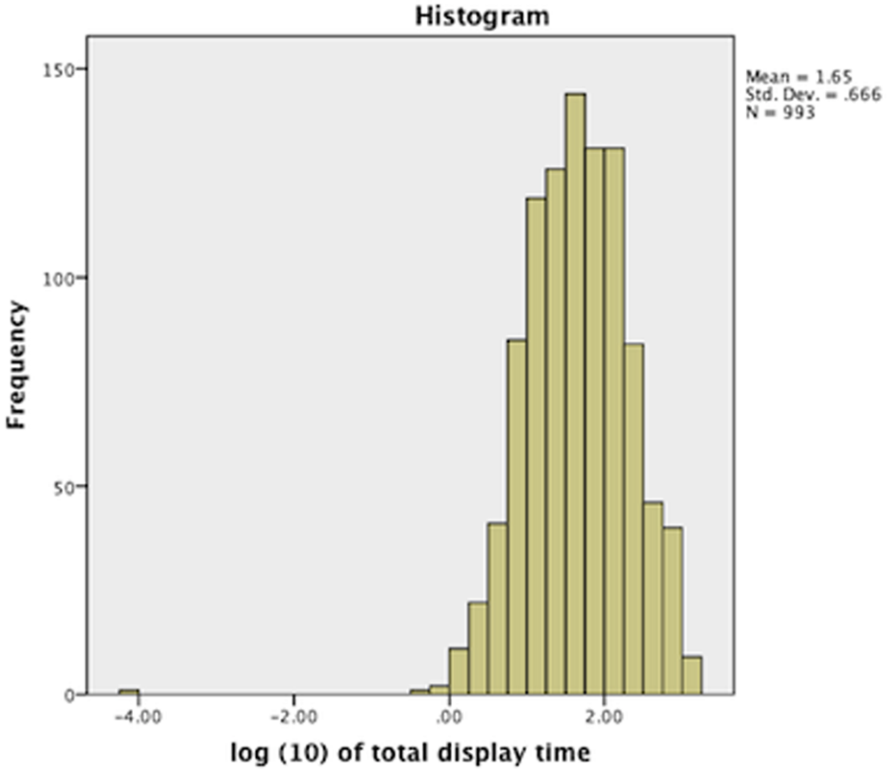

FIG. 4. The distribution of log(10) total display time in both tasks combined. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

7-point Likert scale, how useful the document was for accomplishing the task. The 7-point scale was appropriate for collecting user assessments (Tang, Shaw, & Vevea, 1999), but could be too fine-grained for a system in algorithms and calculation. Therefore, document usefulness was collapsed into fewer groups in a manner similar to what was used in White and Kelly (2006), who collapsed the original 7-point scores elicited from user ratings into three groups: not-useful, mid-useful, and highly useful, based on the examination of the distribution of responses to the question. For example, Figure 5 shows the distribution of the original usefulness data in both tasks. From this distribution, it is reasonable to combine scores 1–2 into a not-useful group, 3–5 into a mid-useful group, and 6–7 into a highly useful group. Figure 6 shows the distribution after grouping, where the three groups were quite balanced. In the following part of this subsection, unless specified, usefulness is denoted as the grouped usefulness.

Results Related to Time, Stage, and Usefulness

The RQs ask about the relationships among time, stage, and usefulness; and those among time, knowledge, and usefulness. The general linear model (GLM) test (Madsen & Thyregod, 2011) was used for statistical examination for the RQs because it can detect interaction effects between/among variables, in addition to the main effects of independent variables on dependent variables. In the analysis, time was used as dependent variable, and stage, knowledge, and usefulness were used as independent variables.

This subsection reports the results for RQs 1, 2, and 3. In the analysis, we looked at both tasks combined, and each task individually, to detect if there were differences in the relationship patterns when task type is known or not, and for different tasks.

Both tasks combined. For total display time, the results (Table 1) show that there was a significant main effect of usefulness, meaning that the relation between usefulness and total display time was significant, and therefore total display time could be a reliable indicator of document usefulness. There was also a significant main effect of stage. In addition, there was a significant interaction effect between stage and usefulness, meaning that the patterns of the relation between usefulness and total display time varied across stages.

For total dwell time, usefulness was found to have a significant main effect, meaning that usefulness and total dwell time had a significant relationship, and total dwell time could be a reliable indicator of document usefulness. The relation between time and stage was not significant, nor was the relation between time and the interaction of usefulness and stage. In fact, in stages 2 and 3 the relationship patterns were almost identical.

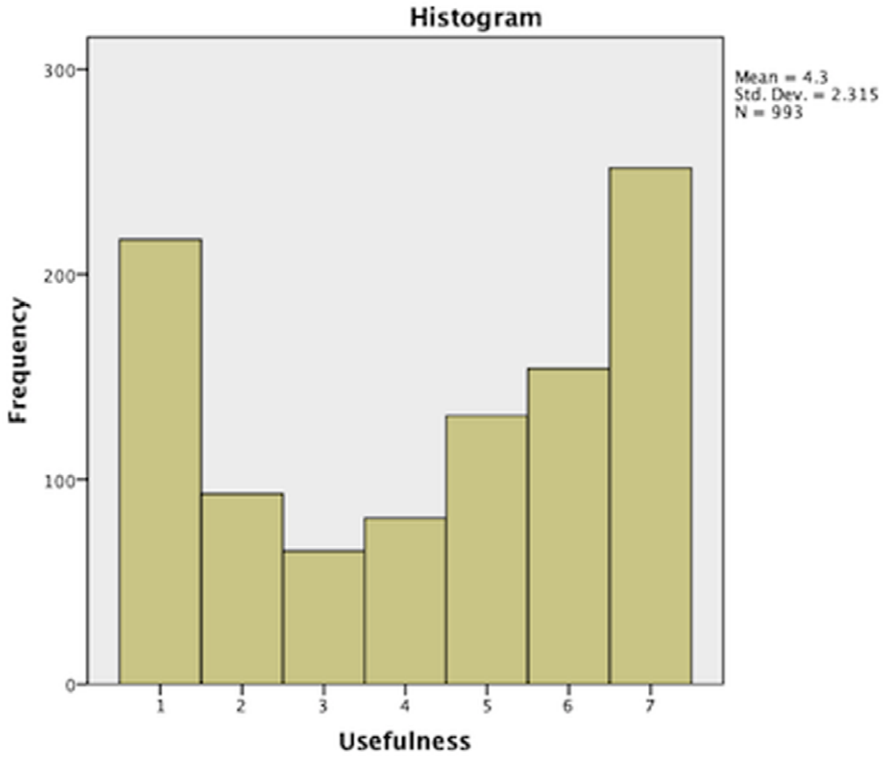

FIG. 5. The distribution of original usefulness data in both tasks combined. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

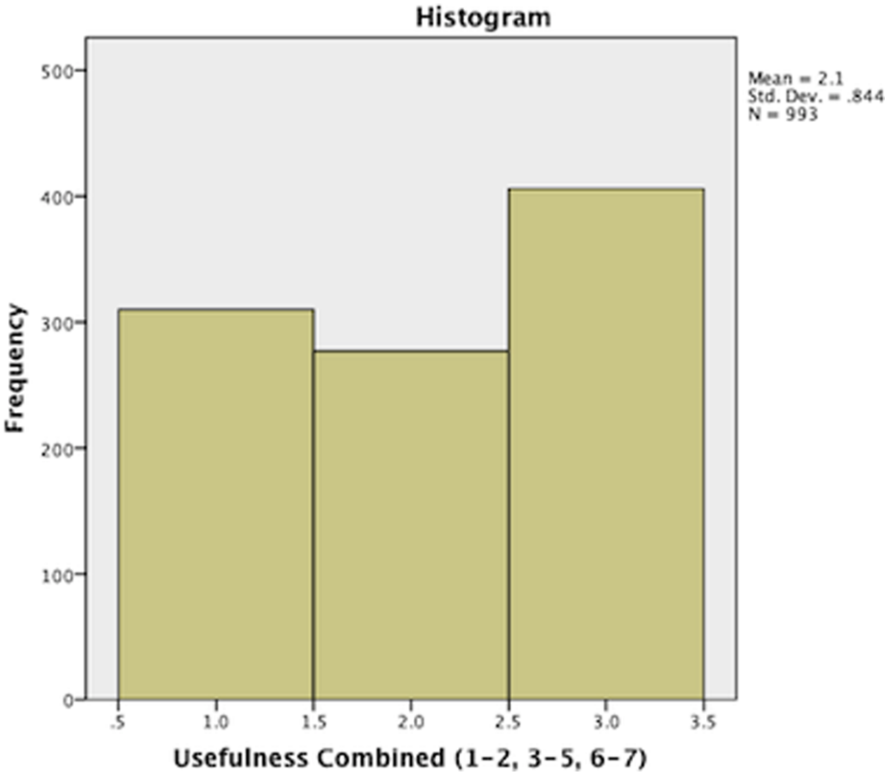

FIG. 6. The distribution of combined usefulness data in both tasks combined. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

TABLE 1. Summary of the *F*(*p*) values of stage and usefulness (results of GLM analyses).

Task	Type of time	Stage	Usefulness	Stage*usefulness
In both tasks combined	Log(10) total display time	**4.150 (.016)**	**123.779 (.000)**	**2.658 (.032)**
	Log(10) total dwell time	1.682 (.187)	**75.402 (.000)**	.817 (.514)
	Log(10) decision time	.326 (.722)	2.158 (.116)	**3.619 (.006)**
Dependent task	Log(10) total display time	1.959 (.142)	**63.404 (.000)**	1.905 (.108)
	Log(10) total dwell time	1.290 (.276)	**35.225 (.000)**	.829 (.507)
	Log(10) decision time	.790 (.454)	**3.336 (.036)**	1.572 (.180)
Parallel task	Log(10) total display time	2.402 (.092)	**61.110 (.000)**	1.393 (.236)
	Log(10) total dwell time	.477 (.621)	**40.781 (.000)**	.425 (.791)
	Log(10) decision time	.449 (.639)	.140 (.869)	**2.478 (.043)**

Note. Those in bold were statistically significant. The *p* values are in parentheses.

For decision time, the results show that neither stage nor usefulness had a main effect on time, meaning that the relationship between time and usefulness or that between time and stage was not significant. Nevertheless, there was a significant interaction effect between usefulness and stage on the log(10) of decision time. The patterns of decision time in the three stages were very different. In stage 1, decision time for mid-useful documents was the lowest, but users spent more decision time on not-useful documents, and even more time on highly useful documents. However, in stage 2, the pattern was exactly reversed. The decision time for mid-useful documents was the longest, followed by that for highly useful documents, and then the not-useful documents. In stage 3, decision time for both not-useful and highly useful documents was shorter than that for mid-useful documents.

In the dependent task. For total display time, the results show that usefulness had a main effect, stage did not show a significant main effect; and there was no significant interaction effect between stage and usefulness on log(10) of total display time.

For total dwell time, the results show that usefulness had a significant main effect on log(10) of total display time, stage did not have a significant main effect, and the interaction between stage and usefulness did not show a significant effect. Figure 7 shows that the total display time patterns in stage 2 and stage 3 were almost identical.

For decision time, usefulness showed a significant main effect on log(10) of decision time, meaning that the relation between usefulness and log(10) of decision time was significant. This shows that the longer the decision time the more likely that the documents were useful. However, stage did not have a significant main effect, nor was there a significant interaction effect between stage and usefulness on decision time.

In the parallel task. For total display time, the results show that usefulness had a significant main effect on log(10) of total display time, stage did not have a significant main effect, and the interaction between stage and usefulness did not show a significant effect.

For total dwell time, the results show that usefulness had a significant main effect, stage did not have a main effect; nor was there a significant interaction effect between stage and usefulness on total dwell time.

For decision time, neither usefulness nor stage was found to have a significant main effect. However, the interaction between stage and the combined usefulness was significant.

Summary. From Figure 7, one can see that decision time had very different patterns than the other two types of time, with which usefulness had a positive correlation. From Table 1, one can further see that the effects of stage and usefulness also varied for different types of times, as well as in different task types. As for usefulness, it always showed a significant main effect on both the total display time and the total dwell time. For decision time, the effect of usefulness is as follows. In the dependent task, usefulness showed a significant main effect on decision time. In the parallel task and in both tasks combined, usefulness did not show a main effect on decision time; instead, there was an interaction effect of stage and usefulness on decision time.

In terms of the effect of stage, in the dependent task, stage did not show any effects on any types of time. However, in the parallel task, stage showed an interaction effect (with usefulness) on the decision time. In both tasks combined, for the total display time, stage showed both main effect and interaction effect (with usefulness); for the decision time, stage showed interaction effect (with usefulness) on the decision time.

Topic Knowledge

Knowledge variables elicited and used in analysis. Topic knowledge in this study was measured by the user's self-assessed familiarity on the topic based on a 7-point scale. In the experiment, users' familiarity degrees with the general task were evaluated both before and after each session, and both for general tasks and subtasks. This generated four scores about topic knowledge: pre- and postsession general task topic knowledge, and pre- and postsession subtask topic knowledge.

DOI: 10.1002/asi

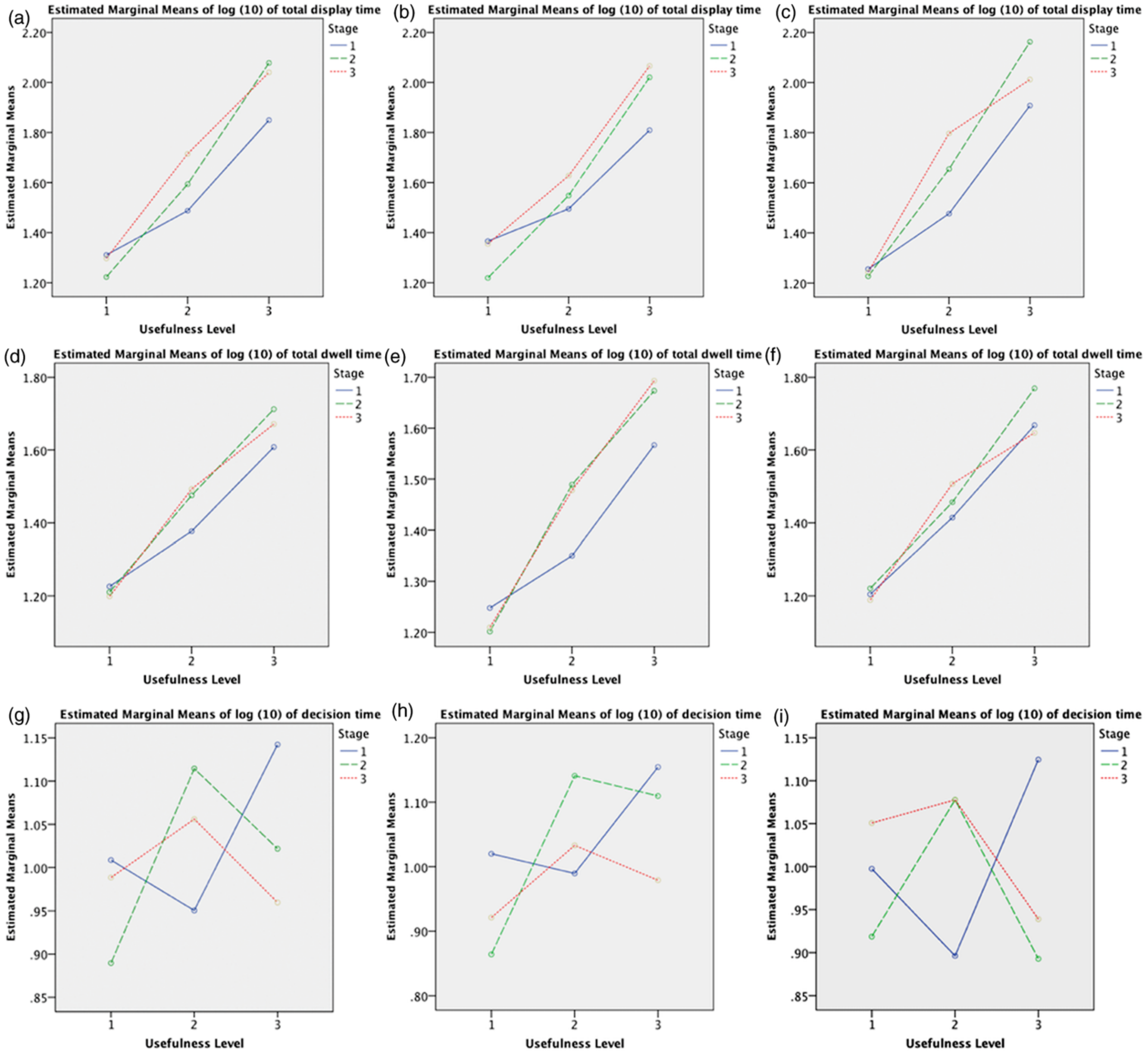

FIG. 7. Relations of time, usefulness, and stage. (a) In both tasks combined: Log(10) total display time. (b) In both tasks combined: Log(10) total dwell time. (c) In both tasks combined: Log(10) decision time. (d) In the dependent task: Log(10) total display time. (e) In the dependent task: Log(10) total dwell time. (f) In the dependent task: Log(10) decision time. (g) In the parallel task: Log(10) total display time. (h) In the parallel task: Log(10) total dwell time. (i) In the parallel task: Log(10) decision time. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

For the purpose of helping interpret time as an indicator of document usefulness, so that the system might help users in their search, it seems to make the most sense to use the presession topic knowledge. In addition, general task topic knowledge measured users' knowledge increase of the same overall task, but subtask topic knowledge in the three stages measured users' knowledge of the different subtasks. Since users in the study worked with a multisession task, it is natural to consider their knowledge of the whole task. Therefore, only the presession general task topic knowledge was used in investigating the relationship of general task topic knowledge with document usefulness and time.

Topic knowledge in three stages and two tasks. Table 2 shows the comparison data of pre- and postsession general task topic knowledge across three stages. Figure 8 depicts the change of presession general task topic knowledge across three stages. As can been seen, knowledge levels increased across stages when both tasks were combined and in each task individually. At stage 3, knowledge levels had significantly improved over those at stage 1.

Grouping knowledge scores. Just as was done for the usefulness scores, knowledge scores were also combined into fewer groups since it is appropriate for the system to differentiate user knowledge based on three levels: not familiar

TABLE 2. Mean and standard deviation of presession general task topic knowledge in two types of tasks at 3 stages.

Task	Stage 1	Stage 2	Stage 3	*F(p)*
Dependent	2.33 (1.073)	3.58 (1.443)	4.08 (1.676)	**4.84 (.014)**
Parallel	3.17 (1.801)	4.00 (1.651)	5.42 (1.730)	**5.20 (.011)**
Total	2.75 (1.511)	3.79 (1.532)	4.75 (1.800)	**9.16 (.000)**

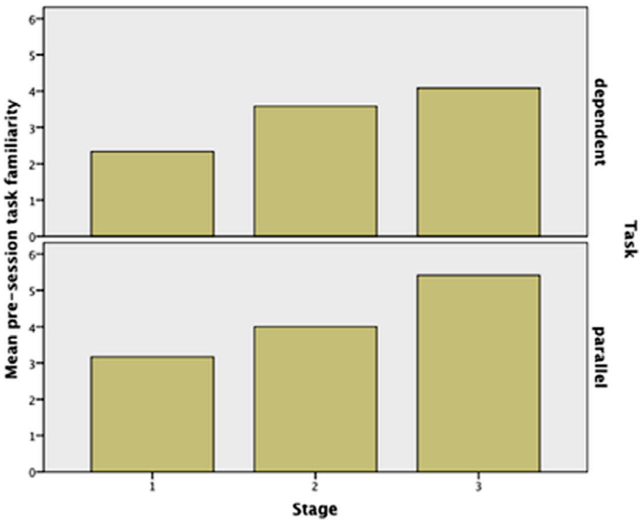

FIG. 8. Presession general task knowledge at three stages in two tasks. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

(little knowledge; the original familiarity ratings 1–2), somewhat familiar (some knowledge; the original familiarity ratings 3–5), and very familiar (much knowledge; the original familiarity ratings 6–7). In the rest of this section, unless specified, usefulness and topic knowledge refer to grouped usefulness and grouped topic knowledge.

Results Related to Time, Topic Knowledge, and Usefulness

This subsection reports results answering RQs 1, 4, and 5. Again, we looked at all tasks combined, as well as each task individually to detect if there were differences in the relationship patterns in different tasks.

Both tasks combined. For total display time, as Figure 9 indicates, usefulness was found to have a significant main effect, meaning that the relation between usefulness and log(10) of total display time was significant, that is, the more useful the documents, the longer the total display time. Topic knowledge did not have a significant main effect on log(10) of total display time. However, the interaction effect between topic knowledge and usefulness was significant, meaning that the relationship patterns of usefulness and total display time varied according to different levels of topic knowledge.

For total dwell time, usefulness showed a significant main effect, but topic knowledge did not. There was also a significant interaction effect between usefulness and topic knowledge.

For decision time, although usefulness had a significant *p* value (.042) (Table 3), a closer examination in the post-hoc analysis detected that the three levels of usefulness scores did not actually have any differences. Topic knowledge showed a significant main effect on log(10) of decision time. In addition, there was a significant interaction effect between usefulness and topic knowledge on log(10) of decision time.

These findings indicate that, if topic knowledge is not considered, all users seemed to have equally quickly determined the usefulness of retrieved documents that had different levels of usefulness. However, when considering different levels of topic knowledge, users with different levels of knowledge spent variable time judging the usefulness of the documents.

In the dependent task. For total display time, usefulness showed a significant main effect; topic knowledge did not show a significant main effect; and there was no significant interaction effect between knowledge and usefulness.

For total dwell time, usefulness showed a significant effect, but topic knowledge did not. The interaction between usefulness and knowledge did not show a significant effect either.

For decision time, as what was found above with regard to total display time or total dwell time, usefulness had a significant main effect. Topic knowledge did not have a significant main effect on decision time. The interaction between topic knowledge and usefulness did not show a significant effect on decision time either.

In the parallel task. For total display time, usefulness showed a significant main effect; topic knowledge did not have a significant main effect; and the interaction between usefulness and knowledge had a significant effect on total display time, meaning that the relation between usefulness and total display time varied across different levels of topic knowledge.

For total dwell time, usefulness showed a significant main effect; topic knowledge did not show a significant main effect; and the interaction between usefulness and knowledge showed a significant effect.

For decision time, unlike the previous results for total display time and total dwell time, usefulness did not appear to have a significant main effect on decision time, but topic knowledge did. The interaction effect between usefulness and knowledge on decision time was marginally significant.

Summary. As can be seen from the results, significant relationships were found between usefulness and total display time, as well as usefulness and total dwell time in all types of tasks. Topic knowledge had a significant relationship with decision time in both tasks combined and in the parallel task,

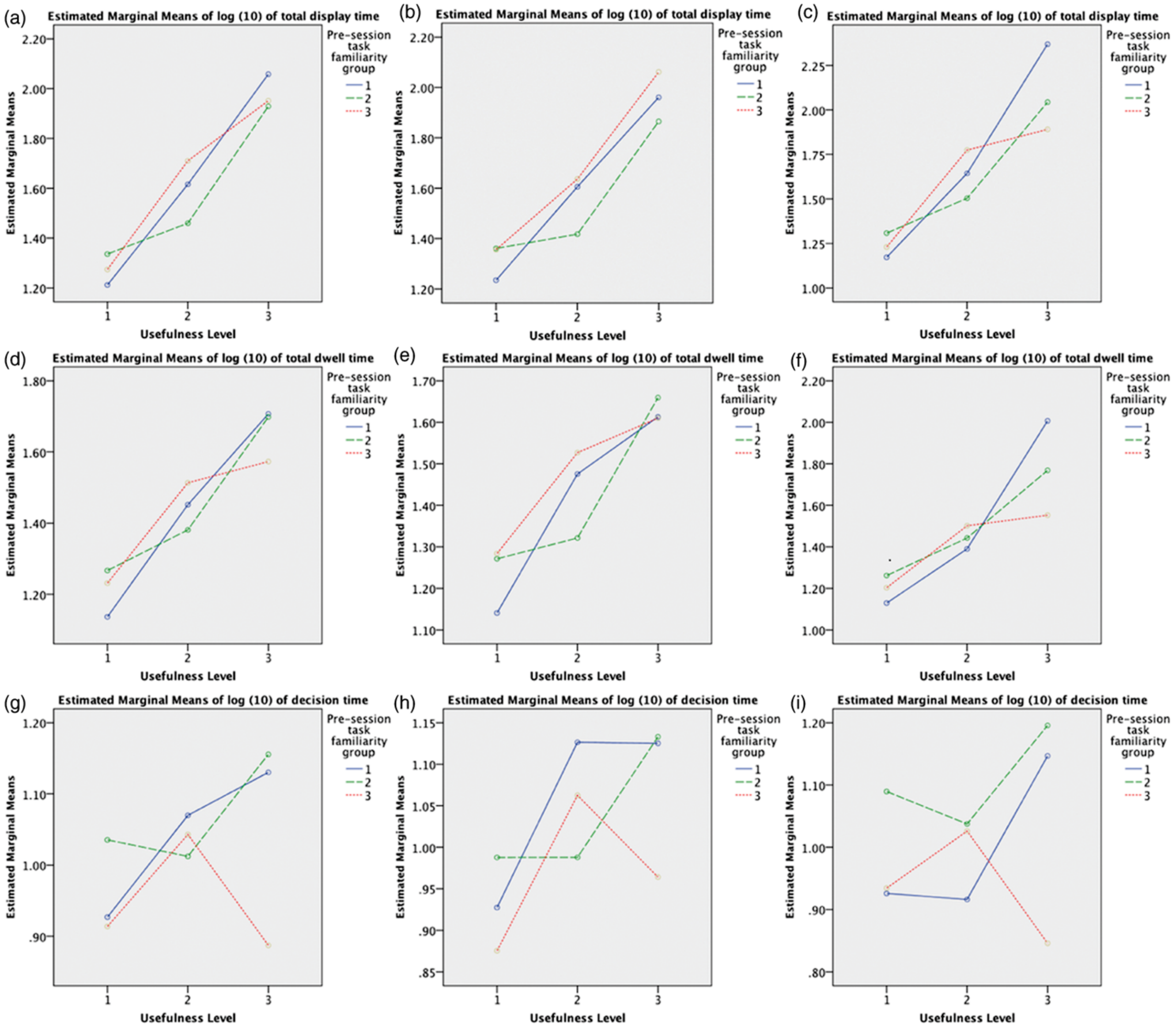

FIG. 9. Relations between time, topic knowledge, and usefulness. (a) In both tasks combined: Log(10) total display time. (b) In both tasks combined: Log(10) total dwell time. (c) In both tasks combined: Log(10) decision time. (d) In the dependent task: Log(10) total display time. (e) In the dependent task: Log(10) total dwell time. (f) In the dependent task: Log(10) decision time. (g) In the parallel task: Log(10) total display time. (h) In the parallel task: Log(10) total dwell time. (i) In the parallel task: Log(10) decision time. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

TABLE 3. Summary of the *F*(*p*) values of topic knowledge and usefulness (results of GLM analyses).

Task	Type of time	Topic knowledge	Usefulness	Topic knowledge*usefulness
In both tasks combined	Log(10) total display time	1.314 (.269)	**123.112 (.000)**	**3.050 (.016)**
	Log(10) total dwell time	.102 (.903)	**76.114 (.000)**	**3.501 (.008)**
	Log(10) decision time	**4.498 (.011)**	3.170 (.042)	**3.039 (.017)**
Dependent task	Log(10) total display time	2.252 (.106)	**60.574 (.000)**	1.345 (.252)
	Log(10) total dwell time	.751 (.472)	**29.806 (.000)**	1.865 (.115)
	Log(10) decision time	1.097 (.335)	**3.312 (.037)**	.940 (.440)
Parallel task	Log(10) total display time	.895 (.410)	**65.704 (.000)**	**3.938 (.004)**
	Log(10) total dwell time	1.478 (.229)	**51.787 (.000)**	**4.688 (.001)**
	Log(10) decision time	**4.666 (.010)**	.880 (.415)	**2.302 (.058)**

Note. Those in bold were statistically significant. The *p* values are in parentheses.

but not in the dependent task. Significant interaction effects between usefulness and topic knowledge were found on all types of time in the parallel task, as well as in both tasks combined, but not in the dependent task. These results indicate that in the dependent task, any of the three types of time can be a reliable indicator of document usefulness without consideration of topic knowledge. However, in the parallel task and in both tasks taken together, any of the three types of time only cannot be a reliable indicator of document usefulness. Taking topic knowledge into consideration does help in interpreting these times as indicators of usefulness.

Comparison Between the Roles of Task Stage and Topic Knowledge

Descriptive comparison. Since the roles of task stage and topic knowledge appeared in both tasks combined and the parallel task, this section compares their roles in these two settings. When considering both tasks together, for not-useful documents, users in stage 2 had shorter decision times but those in stages 1 and 3 had longer decision times. Users with little or much topic knowledge had shorter decision times than those with some knowledge. This seems to indicate that in determining a document's usefulness when it was not-useful there was a correspondence between stage 2 and knowledge levels 1 and 3 (shorter decision time), and a correspondence between stages 1 and 3 and knowledge level 2 (longer decision time). Those with either little or much knowledge make the decision rather quickly, just the same as people in stage 2. On the other hand, those with some knowledge took a long time to make the decision, just as people in stage 1 or stage 3.

For mid-useful documents, in stage 1 users had the shortest decision time; in stage 2 they had the longest decision time; in stage 3 the decision time was in between that in the other two stages. Meanwhile, people with different levels of topic knowledge did not seem to differ in decision time.

For highly useful documents, in stage 1 users spent a long time to make a usefulness decision, in stage 2 they spent less time, and in stage 3 they spent very little time. However, those with some or little knowledge took a long time to make a usefulness decision, but those with much knowledge took a very short time to make a usefulness decision. This seems to indicate that in making a usefulness judgment when it was actually highly useful, there was a correspondence between stage 1 and knowledge levels 1 and 2 (long decision time), and stage 3 and knowledge level 3 (short decision time). Those with much knowledge made the decision rather quickly, just the same as people in stage 3. On the other hand, those with only some or little knowledge took a long time to make the decision, as people in stage 1.

In the parallel task, the patterns were a bit different than those in both tasks together. For not-useful documents, in stage 2 users had the shortest decision time; in stage 3 they had the longest decision time; while in stage 1 their decision time was in between that in stages 1 and 3. Those with little or much knowledge had shorter decision time, and those with some knowledge had longer decision time. This seems to indicate that in making a usefulness judgment when it was actually not useful, there was a correspondence between stage 2 and knowledge levels 1 and 3 (short decision time), and stage 3 and knowledge level 2 (long decision time). Those with either little or much knowledge make the decision rather quickly, just the same as people in stage 2. On the other hand, those with some knowledge took a long time to make the decision, as people in stage 3.

For mid-useful documents, in stage 1 users had a shorter decision time, but in both stages 2 and 3 they had longer decision times. Users with little knowledge had a shorter decision time, and those with at least some knowledge had a longer decision time. This seems to indicate that in making a usefulness judgment when it was actually mid-useful, there was a correspondence between stage 1 and knowledge level 1 (shorter decision time), and stages 2 and 3 and knowledge levels 2 and 3 (longer decision time). Those with little knowledge made the decision rather quickly, just the same as people in stage 1. On the other hand, those with at least some knowledge took a long time to make the decision, as people in stages 2 and 3.

For highly useful documents, in stage 1, users had a very long decision time, but in stages 2 and 3 they had a short decision time. Users with only some or even little knowledge had long decision time, but those with much knowledge had short decision times. This seems to indicate a correspondence between stage 1 and knowledge levels 1 and 2 (long decision time), and stages 2 and 3 and knowledge levels 3 (short decision time). This seems to indicate that in making a usefulness judgment when it was actually highly useful, there was a correspondence between stage 1 and knowledge levels 1 and 2 (long decision time), and stage 3 and knowledge level 3 (short decision time). Those with much knowledge make the decision rather quickly, just the same as people in stage 3. On the other hand, those with only some or little knowledge took a long time to make the decision, as people in stage 1.

Further comparison in a GLM model. Further analysis using the GLM model was conducted to confirm these findings and to compare the factors considered in RQ2, that is, knowledge, with that in RQ1, that is, stage. Since both topic knowledge and stage showed effects only when decision time was considered but not the other two types of time, this analysis only looks at decision time. The results show that when both task stage and task familiarity were considered, stage did not appear to be a significant factor, nor did task familiarity. Usefulness did not either. However, the interaction of usefulness and task familiarity (i.e., topic knowledge) had a significant effect, $F(4, 988) = 2.425$, $p < .05$, but not the interaction of stage and usefulness. This seems to suggest that presession task familiarity plays a more significant role than task stage in interpreting decision time as an indicator of usefulness.

Discussion and Implications

Three Types of Time and Their Use in Modeling Users and Personalizing Search

The current study identified and used three types of time: total display time, total dwell time, and decision time. Previous research has only used dwell time or display time to represent the duration that a web page is displayed for a user to view, with no consideration of how long a specific web page is viewed at different times in a given period. Also, previous studies have not differentiated between the time that a document was viewed by the users (i.e., total dwell time) and the time that a document was opened, even though it was not viewed (i.e., total display time), possibly because they were rarely conducted in the context of doing a task with an output other than finding documents.

Among the three types of time identified in the current study, total display time and total dwell time were both measured at a whole-session level and thus cannot be captured until a session is finished. Although in a multisession task, at the end of a session, these two types of time can be captured and may then be used to personalize search in the following sessions, they cannot be used for personalizing search for the ongoing session. On the other hand, decision time can be captured at a much earlier phase in a session and therefore can be used for adapting search for the current session, in addition to adapting search in the following sessions. Detailed discussion about how each type of time can be used is presented in the following sections.

Time, Task Stage, and Document Usefulness: The Stage Model

This section discusses the use of time as an indicator of usefulness when stage is considered in the GLM model, simplified as the stage model.

Time as an indicator of usefulness. As was found when both tasks were combined, those documents that had longer total display time or total dwell time were more likely to be useful. This is reasonable, considering that when working with their tasks, users often moved back and forth between reading useful documents and writing reports, and the length of total dwell time and total display time of those documents which were more useful was therefore increased. These findings indicate that when task type was not specified, both total display time and total dwell time were rather reliable indicators of document usefulness. Meanwhile, when task type was not specified, decision time alone could not be used as a reliable indicator of document usefulness.

In the dependent task, each of the three types of time appeared to be reliable indicators of document usefulness. Simply put, the findings were that the longer the time (any type), the more useful the document was. For total dwell time, this could be explained by the same point as that for both tasks considered together (see the preceding paragraph), that for the useful documents, users kept referring back to them when they wrote the reports so that the total dwell time of such documents was prolonged. For total display time, this could be explained by the observation that in the dependent task, even when the users did not read the documents, for example, when they were writing the reports, if the documents were more useful, users still left them open, which extended their total display time. For decision time, this finding was that the longer the users spent on the document before leaving it for the first time after the documents was opened, the more useful the document was. This could be explained by the fact that in the dependent task, for more useful documents, the users probably needed to read longer to get the useful pieces in the document before starting to use them in writing the reports. Possible reasons for this may be that they had little knowledge of the documents (see more in the next paragraph about the parallel task).

By contrast, in the parallel task, total display time and total dwell time were shown to be reliable indicators of usefulness, but decision time was not. Possible explanations for the findings on total display time and total dwell time in the parallel task could be the same as those in the dependent task, namely, that the users kept referring back during their writing to the more useful documents (prolonged total dwell time), and that they left the more useful documents open when they wrote the reports (prolonged total display time). Concerning decision time, the finding was that users did not necessarily spend a longer time on more useful documents before leaving them the first time after the documents were opened. This could be explained by the fact that, in the parallel task, users may have already obtained some knowledge of the documents in previous sessions, so that they did not need to spend time getting familiar with the documents.

To sum up, total dwell time and total display time were shown to be reliable indicators of document usefulness in both tasks combined, and in either the parallel or the dependent task. However, decision time as a single indicator of usefulness only worked in the dependent task. Given the aforementioned limitation of total dwell time and total display time, using time as a reliable indicator of usefulness to personalize for the current session can only be applied in the dependent task, when decision time alone is used. Table 4 is a summary of the indicators of usefulness in the stage model.

Stage as a helpful contextual factor in inferring usefulness from time. As can be seen from Table 4, in both tasks combined, stage appears to have a significant interaction effect with usefulness on time. This means that stage played a role in interpreting time as an indicator of document usefulness without regard to task type. The role of task stage in both tasks combined could be due to the strong influence of the parallel task.

In the parallel task, it was found that decision time alone cannot be a reliable indicator of document usefulness; however, when task stage information is also considered, it

TABLE 4. Summary of indicators of usefulness in the stage model.

Time type	Role as indicator of usefulness	Applicable task type			Applicable sessions
		Both	Dependent	Parallel	
Total display time	Single	√	√	√	Following
	With stage	√			Following
Total dwell time	Single	√	√	√	Following
	With stage				Following
Decision time	Single		√		Current
	With stage	√		√	Current

can. One possible explanation of this role of task stage in helping interpret decision time as an indicator of usefulness could be that, in the parallel task, subtask topics changed across stages, but subtask patterns did not, and users were dealing with roughly the same things (e.g., exterior and interior features, performance, safety, prices, and colors) on different car models. Users were very likely to have gained some knowledge of subtasks, or come across the same documents (or similar documents in the same websites) in later stages, which may greatly reduce their time spent on deciding the usefulness of these documents (i.e., decision time).

In the dependent task, however, knowledge of stage did not seem to contribute to the informative value of any of the three types of time. An explanation of why stage did not play a role in the dependent task could possibly be that in the dependent task, not only were subtask topics different, but subtask patterns were also different. This is different from the case in the parallel task, as explained in the preceding paragraph. In the dependent task, the users were dealing with quite different types of subtasks in the three sessions, and they most likely looked at different web pages on different websites. Users would not have gathered knowledge over stages that may have changed the time that they needed to spend on determining the usefulness of the documents (decision time), on reading the documents (total dwell time), or on keeping the document display (total display time).

In sum, task stage was found in this study to be a significant factor that may help interpret time as an indicator of usefulness. When no task information was specified, task stage was found to help interpret total display time as an indicator of usefulness. This finding can be used for subsequent search/work sessions on the same task, although it cannot be applied to the ongoing session due to the limitation that total display time cannot be captured until the end of a session. In addition, task stage was found to help in interpreting decision time (which can be captured within the session) as an indicator of usefulness, which can be used for personalization in the current session. This role of task stage seems to be due to its strong influence in the parallel task, where task stage helped interpret decision time as an indicator of usefulness. These findings can help personalize search for specific users in that decision time can be a reliable indicator of document usefulness given the task stage (and task type) information.

Task type as a helpful contextual factor in inferring usefulness from time. Generally speaking, our most important finding is that when interpreting decision time as an indicator of document usefulness, in the dependent task, task stage did not actually play a role, but in the parallel task, it did. The possible explanation is that in the dependent task, subtasks that the users worked with in the three sessions were different, not only in their topics, but also in the subtask patterns. However, in the parallel task, subtasks that the users worked with in the three sessions were different only in their topics; they had the same subtask patterns, and users only changed car models across sessions, but they worked on the same or similar aspects including cars' exterior or interior features, performance, safety, etc. It is possible that users gained knowledge across stages in the parallel task on the usefulness of some documents or types of documents, and hence, their decision time on useful documents in later stages was reduced; while in the dependent task, users would not have gained such knowledge, hence their decision time remained the same across stages.

When there is no task information specified, that is, in both tasks combined, stage also played a role when decision time is used for personalization. This is due to, we think, the strong role of task stage in the parallel task. Although inferring document usefulness based on decision time and stage still works in the absence of task type information, taking it into account should be able to increase the interpretation accuracy, that is, the overall correctness of usefulness prediction.

What also needs to be mentioned is that if total display time were to be used for personalizing search for subsequent sessions, stage was also found to play a role when no task type information was specified. Interestingly, this role did not hold true in each individual task. This again indicates that task type information is important in order to accurately interpret total display time as an indicator of document usefulness from total display time.

Time, Topic Knowledge, and Usefulness: The Knowledge Model

This section discusses the use of time as an indicator of usefulness when knowledge is considered in the GLM model, simplified as the knowledge model.

TABLE 5. Summary of indicators of document usefulness in the topic knowledge model.

Time type	Role as indicator of usefulness	Applicable task type			Applicable sessions
		Both	Dependent	Parallel	
Total display time	Single	√	√	√	Following
	With topic knowledge	√		√	Following
Total dwell time	Single	√	√	√	Following
	With topic knowledge	√		√	Following
Decision time	Single		√		Current
	With topic knowledge	√		√	Current

Time as an indicator of usefulness. In general, it was found that all three types of time as indicators of usefulness in the topic knowledge model were consistent with those in the stage model. Table 5 is a summary of the indicators of document usefulness when topic knowledge was considered in the model.

Topic knowledge as a helpful contextual factor in inferring usefulness from time. Although in the dependent task, topic knowledge did not seem to play any role in interpreting any of the three types of time as indicators of usefulness, in the parallel task and in both tasks combined, topic knowledge was found to have played a significant role.

In the parallel task, for highly useful documents, those users with high levels of topic knowledge viewed the documents (i.e., total dwell time) or had them displayed (i.e., total display time) for less time than those with medium level of knowledge, and than those with low level of knowledge. A possible explanation could be that users with higher levels of knowledge knew where to look and how to use the useful pieces in the useful documents in their writings (the writing process was going on in parallel with document reading) so that the total dwell time and total display time of useful documents of this group of people was shorter. In addition, for highly useful documents, users with high levels of topic knowledge made decisions about document usefulness (i.e., decision time) very quickly, while those with medium and low levels of knowledge did this relatively slowly. This can be explained, at least in part, by the fact that users with higher levels of knowledge knew whether or not the document was useful, which part(s) of the document was useful, and how to use the useful information before they started to write using these pieces of information.

When task type was not specified, findings were similar to those in the parallel task. Again, this means that the parallel task had a great impact on the combined group, so that the findings still hold true even when the dependent task was also included in analysis.

These findings indicate that in the parallel task, or when task type was not specified, topic knowledge played a significant role in interpreting all three types of time as indicators of document usefulness, especially from decision time (when time only was not able to reliably infer usefulness). The role of topic knowledge in inferring usefulness when total display time and total dwell time are used can be applied to the subsequent search/work sessions but not to the ongoing session. The role that topic knowledge plays in inferring usefulness from decision time can be applied to both the subsequent sessions and the ongoing session since decision time of a page can usually be captured in an early phase of a session.

Task type as a helpful contextual factor in inferring usefulness from time. Regarding task type, our results had a similar pattern in the knowledge model as in the stage model, that in the dependent task, topic knowledge did not help interpret usefulness using time, but it did in the parallel task. The explanation to this would be similar as in the stage model too, that users become more knowledgeable about web page patterns in the parallel task, but not in the dependent task. Again, it is important to know task type information in order to better infer document usefulness from time.

Comparison of the Roles of Task Stage and Topic Knowledge in Interpreting Time as an Indicator of Usefulness

Both could help, in general. The results showed that task stage and topic knowledge were both found to have the potential to help infer document usefulness from time in general. Both were found to help in the parallel task or when no task type information was specified but not to the dependent task. Both were found to have especially significant relationships when decision time was considered with respect to usefulness, under which situations time only cannot be used to infer usefulness at all. Recall that users' topic knowledge was found to increase with stage, which meant that these two factors were positively correlated to some degree, so it is reasonable to see that they both could help in general.

Examining their potential in detail. Although task stage and topic knowledge were both found to be potentially important in interpreting time as an indicator of document usefulness, the roles they played were not always the same when considering the specific values of these two variables. In other words, it is not the case that stage 1 corresponded to

TABLE 6. Frequency of knowledge levels by task stages in the parallel task.

	Presession topic knowledge level		
Stage	1	2	3
1	64	64	48
2	11	78	52
3	8	21	104

TABLE 7. Frequency of knowledge levels by task stages in both tasks combined.

	Presession topic knowledge level		
Stage	1	2	3
1	191	161	48
2	73	142	111
3	19	77	171

topic knowledge level 1, stage 2 to topic knowledge level 2, and stage 3 to knowledge level 3.

As the results in Tables 6 and 7 show, in either the parallel task or when task type was not specified, when decision time was used, task stage 3 and topic knowledge level 3 appeared to have very similar roles in helping infer usefulness from time. Specifically, in stage 3 or when the user had much topic knowledge, decision time was short for not-useful documents, it increased for mid-useful documents, and it dropped down for highly useful documents to a similar or lower decision time for not-useful documents. This could be explained by the fact that stage 3 corresponded to topic knowledge level 3. In other words, in stage 3 users should have a high level of topic knowledge. This was supported by the observation of frequencies of knowledge levels in three stages, as shown in Tables 6 and 7, and Figure 10.

Nevertheless, both in the parallel task and when task type was not specified, stages 1 and 2 and knowledge levels 1 and 2 did not correspond as well as stage 3 and knowledge level 3 did. Observation of knowledge levels' distributions at stages showed that in the parallel task and in both tasks combined, in stages 1 and 2, there was not a single dominant knowledge level. In the parallel task, in stage 1, knowledge levels 1 and 2 had the same frequencies ($n = 64$), which were descriptively but not statistically significant higher than level 3 ($n = 48$) (Table 6: frequency of knowledge levels by task stages in the parallel task). In stage 2, both knowledge levels 2 ($n = 78$) and 3 ($n = 52$) had higher frequencies than level 1 ($n = 11$) ($p < .001$). The observations indicate that in stage 1, users' knowledge levels basically evened out, with roughly equal numbers of users with little, medium, and much topic knowledge (levels 1, 2, and 3); in stage 2, most of them already had medium or more knowledge (levels 2 and 3), by stage 3, most of them had much knowledge (level 3). So the role of task stage and topic knowledge in helping infer usefulness does not match by the stage and knowledge level values.

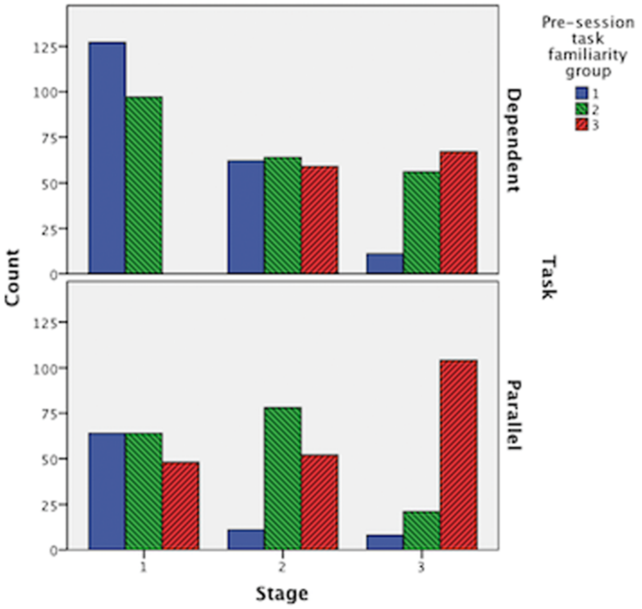

FIG. 10. Frequency of knowledge levels by task stages in two tasks. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

TABLE 8. The effects of stage, knowledge, and usefulness on decision time (results of GLM analysis).

Variables	F	*p*
Usefulness	2.760	.064
Stage	.195	.823
Knowledge	1.956	.142
Usefulness*stage	1.790	.129
Usefulness*knowledge	2.425	**.047**
Stage*knowledge	.941	.440
Usefulness*knowledge*stage	1.835	.087

When task type was not specified, in stage 1, knowledge levels 1 ($n = 191$) and 2 ($n = 161$) had similar frequencies, which were statistically higher than level 3 ($n = 48$) ($p < .001$). In stage 2, knowledge levels 2 ($n = 142$) and 3 ($n = 111$) had higher frequencies than level 1 ($n = 73$) ($p < .001$). The observations basically indicated that in stage 1, most users had little or medium topic knowledge (levels 1 and 2), until stage 2, most of them already had medium or more knowledge (levels 2 and 3), by stage 3, most of them had much knowledge (level 3). So again, the role of task stage and topic knowledge in helping infer usefulness was not exactly the same match by values.

Further comparison of their roles. The results when both task stage and topic knowledge were considered in GLM analyses (Table 8) indicate that task stage did not play a significant role any more, but topic knowledge did. This could be interpreted that topic knowledge could play a more significant and maybe more accurate role than task stage in

DOI: 10.1002/asi

helping infer usefulness. If information about both users' topic knowledge and task stage is available, then it would be good to use topic knowledge to personalize search. However, if only one of them is available, it should also be good to use that information, either the stage or topic knowledge.

Implications for System Design

Our findings have implications for personalization system design. Time as a user behavior can be easily detected by the system, and stage or knowledge may well be detectable, too, as explained later in this section. It is hoped that task type can also be learned through means that will be discussed. Using these types of information, systems could make relatively accurate predictions on the usefulness of documents that searchers have interacted with, based on the findings of this study.

If a personalization system based on dwell time of some sort does not consider stage information, it would have a single threshold criterion across all stages for not useful, somewhat useful, and very useful documents. For instance, based on our results, a system might simply classify those with a decision time of less than 10 seconds (the average time length of not-useful documents in both tasks combined) as not-useful documents, those with decision time of longer than 11.2 seconds (the average time length of high-useful documents in both tasks combined) as highly useful, and those in between as mid-useful. However, if the system knows that the task is parallel, when taking stage information into consideration, the system would set different thresholds at different stages. For example, at stage 1 the system would classify documents with a decision time of less than 1.4 seconds (the time length corresponding to the not-useful documents in the parallel task at stage 1) as not-useful documents, and those with a decision time of longer than 1.5 seconds (the time length corresponding to the highly useful documents in the parallel task at stage 1) as highly useful documents. At stage 2, the thresholds would be different. Those with decision time of less than 12.6 seconds (the time length corresponding to the useful documents in the parallel task at stage 1) would be not-useful, longer than 12.6 seconds would be mid-useful, and those with decision time of 12.6 seconds would be highly useful. At stage 3, the thresholds would again be different than in previous 2 stages. In general, based on the findings of the roles of stage, this approach to setting different thresholds at different stages may lead to better performance in predicting document usefulness. Further studies will attempt to discover the thresholds for different stages, making predictions, and generating the receiver operating characteristic (ROC) curve that describes the prediction performance (both correctness and error rate). It should be noted, however, that the threshold values shown in the current study was just based on the mean values. These numbers may not be the same as in other studies, but it is in practice feasible that the system generates the threshold for the search session in the process that it is monitoring. The threshold is not necessarily the same for all searches, but the method developed in the current study can be used in other studies and system design.

When taking topic knowledge into consideration, the system would also set different thresholds for people with different levels of topic knowledge instead of setting the same thresholds for all people. For example, if the system learns that the user is working on a parallel task, for those with low knowledge, the system would classify documents with a decision time of less than, say, 10.5 seconds as not-useful documents, and those with a decision time of longer than 10.5 seconds as highly useful documents. For those with medium level of knowledge, the thresholds will be different from those with little knowledge. Documents with a decision time of less than, say, 1.6 seconds would be not-useful, and longer than 1.6 seconds would be highly useful. For those with a high level of knowledge, the threshold will be different again. In general, this way of setting different thresholds based on different levels of knowledge should enhance usefulness prediction. Future studies will attempt to discover the thresholds for people with different levels of knowledge, making predictions, and generating the ROC curve that describe the prediction performance (both correctness and error rate).

It may seem that sometimes using decision time and stage (or topic knowledge) information was not a perfect way to infer usefulness. For example, at stage 3 it is difficult to differentiate the very useful and not useful documents since the means of decision times for these two groups were roughly the same. However, the purpose of this study was to explore the role of stage, task type, and topic knowledge in helping to interpret time as an indicator of usefulness, and the results have provided strong evidence for it. The seemingly difficult classification at stage 3 could possibly be improved by some other behavioral signals; for example, if viewing the document was followed by writing in MS WORD documents (the heuristic is to differentiate very useful and not useful documents, which had similar decision time). Future studies will look more into other behavioral signals, how they can be combined with the findings of this study, as well as how to design and evaluate a prototype using these promising findings. It is our hope that by designing systems that can bring the most useful documents to the top ranks in the results list, the information searchers will find it easier to obtain useful documents to solve their tasks, which will lead to more satisfaction with the search systems.

Ways to detect stage, knowledge, and task type. In order for the findings to be applicable in personalization system design, systems will need to be able to know about task stage and topic knowledge, and task type. Other than explicit ways of elicitation of this information, it is possible to infer such information implicitly from users' past and current behaviors.

For task stage, the system could possibly learn it through monitoring users' behaviors. For example, Wang et al. (2013) developed a method to identify cross-session tasks from search logs by investigating interquery dependencies learned from users' searching behaviors.

For topic knowledge, similar approaches can also be used to estimate users' knowledge as low, medium, or high (as our results showed, stage and topic knowledge were correlated). It is also possible to infer topic knowledge from users' domain knowledge. For instance, domain knowledge can be inferred using the way described in White et al. (2009), namely, that users frequently going to the medical database PubMed, etc., are likely to be domain experts, while those who rarely use such databases are likely to be domain novices. In addition, topic knowledge can possibly be learned according to the readability and specificity of the documents that the users like to read, based on the ideas of Belkin et al. (2004).

With respect to task type being parallel or dependent, the system may possibly learn this from the users' query formulation and reformulation behaviors. Liu, Gwizdka, Liu, Xu, and Belkin (2010) found that in parallel tasks users tended to employ more frequently a query reformulation strategy called Word Substitution, that is, to substitute some of the terms in a query while keeping the total number of query terms unchanged. This makes sense. For example, if a user just changes the query from "Honda price" to "Toyota price," it is very likely that the user is working in a parallel task. In short, information relating to all these contextual factors is detectable by the system.

Predicting stage, topic knowledge, or task type. Another way to apply the findings in this study is for the system to learn information about task stage, topic knowledge, and task type based on the users' behaviors, given the users' behavior features and/or document usefulness inferred by other heuristics. For instance, a user using a document in writing indicates that the document is very likely to be very useful. If the decision time of this document, that is, the time before the user starts using this document, is very short, then based on the findings of this study, this user is very likely to be working on a parallel task, and that he/she is in a later stage of his/her task, or his/her knowledge level on this task is pretty high. On the contrary, if the decision time of this document is pretty long, then it is not very likely that the user is engaged on a parallel task, or he/she is in the later stage of the task, or his/her knowledge level on this task is pretty high.

Limitations, Generalization, and Future Studies

The study was a controlled lab experiment with college students working on certain assigned tasks, so care should be taken when generalizing the findings. The three-subtask design to operationalize task stages is different from the real tasks in people's lives, which do not usually have clear stage boundaries. The findings on the roles that task stage plays are valid, but it is an issue to accurately obtain task stage information. Also, despite the fact that the tasks were designed to mimic real-life work tasks, they are still tasks assigned to the users. A longitudinal and naturalistic study is needed to see if there are differences in findings on task stage and topic knowledge when users work with their own tasks.

Nevertheless, the study was carefully designed: the tasks' topics were frequently seen in everyday life, the tasks were designed as simulated task scenarios (Borlund, 2000), and the journalism students were mimicking journalists who are usually not restricted to a certain domain. In addition, the tasks were designed to follow the classification scheme of Li and Belkin (2008) and vary only in one feature while keeping others constant, which makes it possible to generalize the findings relatively safely to other tasks of the same type without concerns about topicality.

With regard to the contextual factors, the study used only two tasks varied along one task feature. Other task features can be considered; for example, task difficulty, task product (being factual or intellectual), and so on. In addition, other factors that may possibly play roles in inferring usefulness from time, for example, some cognitive characteristics such as the need for cognition, can also be considered in future studies.

The study found a significant three-way relationship between contextual factors, usefulness, and time. However, it should be noted that the effect size was not big. Partial eta squared varied from 0.01–0.03, indicating that time only is not enough to predict usefulness. Other types of behaviors, such as saving and using behaviors, will also need to be considered in future studies in order to achieve better prediction of document usefulness based on user behavior.

In addition, this study did not consider document length as an influential factor on the time users spent on documents. The reason is that, as the given task and topic are in general familiar to people, and the retrieved documents for these tasks are in general easy to read, it is not very likely that users will have to spend a significantly longer time on longer documents. Future studies can examine the relation between usefulness, time, and document length, to confirm the findings.

Conclusions

Through a controlled lab experiment, we collected data and analyzed the relationships among information searchers' time spent on documents, the usefulness of the documents, as well as searchers' stage in task accomplishment and their familiarity with search tasks. Our research is important in several ways. We studied information-seeking behavior during the performance of a real-world task, which allowed distinguishing between the different types of time in useful ways. It is also the case that there have been very few studies of information seeking over multiple search sessions on the same motivating task, and our study contributes to knowledge of behaviors in this context. Contextual factors including task stage, user's topic knowledge, and task type were found to be helpful in inferring document usefulness from the time the user spent on a document. The research extends the literature by discovering the conditions of some relationships between certain factors that have been found in previous studies, and providing a systematic way to examine the roles of contextual factors in helping to infer document usefulness. The study also indicates that many aspects of user behaviors are not isolated, but instead, are related to each other. This study has implications for search system design, both theoretically and practically. The results clearly demonstrate that user behaviors are affected by the context in which the user seeks information, instead of being uniform in all circumstances. It has also demonstrated that contextual factors should be taken into consideration when inferring document usefulness from behaviors. Accurately inferring document usefulness for personalizing IR is possible, based on the behavior of decision time, together with consideration of some contextual factors (task stage, topic knowledge, task type), or maybe also other user behaviors (querying, using documents, etc.). It is also possible to predict stage, knowledge, and task type based on usefulness.

Limitations of the study include its being a controlled laboratory experiment using preassigned tasks instead of being a naturalistic setting, relatively small effect size, limited types of information problems, and not having considered document length. Future directions include examining the relationships between user behaviors, document usefulness, and contextual factors in a naturalistic setting, considering more behavioral factors than dwell time in inferring document usefulness, examining more contextual factors such as users' cognitive features, different task types with different task products, difficulty, task goals, etc., and building document usefulness prediction models and prototypes, which may give users better search experience and more search satisfaction.

In conclusion, this research has contributed to a better understanding of how information-seeking behaviors, specifically the time that users spent on documents, can be used as implicit evidence of document usefulness, as well as how contextual factors of task stage, topic knowledge, and task type can help in interpreting time as an indicator of document usefulness. The findings have theoretical and practical implications for using behaviors and contextual factors in the development of personalization systems.

Acknowledgments

This article was based on the dissertation of the first author. We would like to thank the committee members, Jacek Gwizdka, Xiangmin Zhang, and Diane Kelly, for their contributions to this work. This work was supported by IMLS Grant LG#06-07-0105-05. The opinions, findings, and conclusions do not necessarily reflect those of the sponsor.

References

Agosto, D.E., & Hughes-Hassell, S. (2005). People, places, and questions: An investigation of the everyday life information-seeking behavior of urban young adults. Library & Information Science Research, 27, 141–163.

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

Allen, B. (1996). Information needs: A person-in-situation approach. In P. Vakkari, R. Savolainen, & B. Dervin (Eds.), Information seeking in context: Proceedings of an International Conference on Research in Information Needs, Seeking and Use in Different Contexts (pp. 111–122). London: Taylor Graham.

Belkin, N.J. (1993). Interaction with texts: Information retrieval as information-seeking behavior. In: Information retrieval '93. Von der Modellierung zur Anwendung. Konstanz: Universitaetsverlag Konstanz, 55–66.

Ingwersen, P., & Javerlin, K. (2005). Information retrieval in context: IRix, SIGIR Forum, 39, 31–39.

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

Belkin, N.J., Chaleva, I., Cole, M., Li, Y.-L., Liu, Y.-H., Muresan, G., . . . (2004). Rutgers' HARD track experiences at TREC 2004. In Proceedings of TREC 2004.

Belkin, N., Cole, M., & Liu, J. (2009). A model for evaluation of interactive information retrieval. In S. Geva, J. Kamps, C. Peters, T. Sakai, A. Trotman, & E. Voorhees (Eds.), Proceedings of the 31th International Conference on Research and Development in Information Retrieval (ACM SIGIR '09) Workshop on the Future of IR Evaluation, (pp. 7–8), Boston.

Borlund, P. (2000). Experimental components for the evaluation of interactive information retrieval systems. Journal of Documentation, 56(1), 71–79.

Byström, K. (2002). Information and information sources in tasks of varying complexity. Journal of the American Society for Information Science and Technology, 53(7), 581–591.

Byström, K., & Järvelin, K. (1995). Task complexity affects information seeking and use. Information Processing and Management, 31(2), 191–213.

Cool, C. (2001). The concept of situation in information science. In Annual Review of Information Science & Technology. M.E. Williams. Medford, NJ, Information Today, 35, 5–42.

Dervin, B. (2003). Given a context by any other name: Methodological tools for taming the unruly beast. In B. Dervin, & L. Foreman-Wernet (with E. Lauterbach) (Eds.), Sense-making methodology reader: Selected writings of Brenda Dervin (pp. 111–132). Gresskill, NJ: Hampton Press.

Dumais, S. (2007). Information retrieval in context. In T. Lau, & A. Puerta (Eds.), Proceedings of the 12th International Conference on Intelligent User Interface (ACM IUI '07) (pp. 2). New York: ACM.

Freund (2008). Exploring task-document relations in support of information retrieval in the workplace. Unpublished Dissertation. University of Toronto.

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 & Technology, 56(8), 861–871.

Ingwersen, P., & Järvelin, K. (2005). The turn: Integration of information seeking and retrieval in context. Heidelberg: Springer.

Kellar, M., Watters, C., & Shepherd, M. (2007). A field study characterizing web-based information-seeking tasks. Journal of the American Society for Information Science & Technology, 58(7), 999—1018.

Kelly, G.A. (1963). A theory of personality: The psychology of personal constructs. New York: Norton.

Kelly, D. (2006a). Measuring online information-seeking context, part 1. Background and method. Journal of the American Society for Information Science & Technology, 57(13), 1729–1739.

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

Kelly, D., & Belkin, N.J. (2001). Reading time, scrolling and interaction: Exploring implicit sources of user preference for relevance feedback. In D.H. Kraft, W.B. Croft, D.J. Harper, & J. Zobel (Eds.), Proceedings of the 24th International Conference on Research and Development in Information Retrieval (ACM SIGIR '01) (pp. 408–409). New York: ACM.

Kelly, D., & Belkin, N.J. (2002). A user modeling system for personalized interaction and tailored retrieval in interactive IR. In E.M. Rasmussen, E. Toms (Eds.), Proceedings of Annual Conference of the American Society for Information Science and Technology (ASIST '02) (pp. 316–325). Medford, NJ: Information Today.

Kelly, D., & Belkin, N.J. (2004). Display time as implicit feedback: Understanding task effects. In K. Järvelin, J. Allan, & P. Bruza (Eds.), Proceedings of the 27th International Conference on Research and Development in Information Retrieval (ACM SIGIR '04) (pp. 377–384). New York: ACM.

Kelly, D., & Cool, C. (2002). The effects of topic familiarity on information search behavior. In G. Marchionini (Ed.), Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries (ACM JCDL '02) (pp. 74–75). New York: ACM.

Kim, J.H. (2006). Task as a predictable indicator for information seeking behavior on the web. Unpublished Dissertation. Rutgers University.

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

Kumaran, G., Jones, R., & Madani, O. (2005). Biasing web search results for topic familiarity. In H-J. Schek, N. Fuhr, A. Chowdhury, & W. Teiken (Eds.), Proceedings of the 14th International Conference on Information and Knowledge Management (ACM CIKM '05) (pp. 271–271). New York: ACM.

Li, Y., & Belkin, N.J. (2008). A faceted approach to conceptualizing tasks in information seeking. Information Processing & Management, 44, 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, 61(9), 1771–1789.

Liu, J., Cole, M., Liu, C., Belkin, N.J., Zhang, J., Bierig, R., Gwizdka, J., & Zhang, X. (2010). Search behaviors in different task types. In C. Lagoze, L. Giles, & Y-F. Li (Eds.), Proceedings of the 10th ACM/IEEE-CS Joint Conference on Digital Libraries (ACM JCDL '10) (pp. 69–78). New York: ACM.

Liu, C., Gwizdka, J., Liu, J., Xu, T., Belkin, N.J. (2010). Analysis and evaluation of query reformulations in different task types. In C. Marshall, E. Toms, & A. Grove (Eds.), Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem (ASIS&T 2010) (pp. 17:1–17:10). Silver Springs, MD: ASIS&T.

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 the American Society for Information Science & Technology 2013 (ASIS&T 2013). Silver Springs, MD: ASIS&T.

Lin, S.-J. (2001). Modeling and supporting multiple information seeking episodes over the web. Unpublished dissertation. Rutgers University.

Madsen, H., & Thyregod, P. (2011). Introduction to general and generalized linear models. New York: Chapman & Hall/CRC.

Marchionini, G. (1989). Information-seeking strategies of novices using a full-text electronic encyclopedia. Journal of the American Society for Information Science, 40(1), 54–66.

Morita, M., & Shinoda, Y. (1994). Information filtering based on user behavior analysis and best match text retrieval. In W.B. Croft, & C.J. van Rijsbergen (Eds.), Proceedings of the 17th International Conference on Research and Development in Information Retrieval (ACM SIGIR '94) (pp. 272–281). New York: ACM.

Qiu, L. (1993). Analytical searching vs. browsing in hypertext information retrieval systems. Canadian Journal of Information and Library Science, 18(4), 1–13.

Savolainen, R. (1995). Everyday life information seeking: Approaching information seeking in the context of "way of life." Library and Information Science Research, 17, 259–294.

Savolainen, R. (2007). Information source horizons and source preferences of environmental activists: A social phenomenological approach. Journal of the American Society for Information Science and Technology, 58(12), 1709–1719.

Talja, S., Keso, H., & Peitilainen, T. (1999). The production of "context" in information seeking research: A metatheoretical view. Information Processing & Management, 35, 751–763.

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

Taylor, R.S. (1968). Question negotiation and information seeking in libraries. College and Research Libraries, 29(3), 178–194.

Taylor, R.S. (1986). Value added processes in information systems. Ablex.

Taylor, A.R., Cool, C., Belkin, N.J., & Amadio, W.J. (2007). Relationships between categories of relevance criteria and stage in task completion. Information Processing & Management, 43, 1071–1084.

Tombros, A., Ruthven, I., & Jose, J.M. (2004). How users assess web pages for information seeking. Journal of the American Society for Information Science and Technology, 56(4), 327–344.

Toms, E., MacKenzie, T., Jordan, C., O'Brien, H., Freund, L., Toze, S., . . . (2007). How task affects information search. In N. Fuhr, N. Lalmas, & A. Trotman (Eds.), Workshop Pre-proceedings in Initiative for the Evaluation of XML Retrieval (INEX) 2007, 337–341.

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

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

Vakkari, P., & Hakala, N. (2000). Changes in relevance criteria and problem stages in task performance. Journal of Documentation, 56(5), 540–562.

Wang, H., Song, Y., Chang, M.-W., He, X., White, R., & Chu, W. (2013). Learning to extract cross-session search tasks. In R. Baeza-Yates, & S. Moon (Eds.), Proceedings of the 22nd International Conference on World Wide Web Conference (WWW '13) (pp. 1353–1364). Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

White, R.W., Dumais, S., & Teevan, J. (2009). Characterizing the influence of domain expertise on web search behavior. In R. Baeza-Yates, P. Boldi, B. Ribeiro-Neto, & B.B. Cambazoglu (Eds.), Proceedings of the 2nd International Conference on Web Search and Data Mining (ACM WSDM '09) (pp. 132–141). New York: ACM.

White, R.W., & Kelly, D. (2006). A study on the effects of personalization and task information on implicit feedback performance. In V. Tsotras, E. Fox, & B. Liu (Eds.), Proceedings of the 15th International Conference on Information and Knowledge Management (ACM CIKM '06) (pp. 297–306). New York: ACM.

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

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

Zhang, X., Liu, J., & Cole, M. (2013). Task topic knowledge vs. background domain knowledge: Impact of two types of knowledge on user search performance. In Á. Rocha, A.M. Correia, T. Wilson, & K.A. Stroetmann (Eds.), Advances in Information Systems and Technologies, AISC 206, (pp. 179–191). Berlin, Heidelberg: Springer.

DOI: 10.1002/asi