PERSONALIZING INFORMATION RETRIEVAL USING TASK FEATURES, TOPIC KNOWLEDGE, AND TASK PRODUCTS

by

JINGJING LIU

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Communication and Information

written under the direction of

Nicholas J. Belkin, Ph.D.

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2010

ABSTRACT OF THE DISSERTATION

Personalizing Information Retrieval Using Task Features, Topic Knowledge, and Task Products

By JINGJING LIU

Dissertation Director:
Nicholas J. Belkin, Ph.D.

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 and contextual factors. The current study looks particularly at users' dwelling behavior, measured by the time that they spend on documents; and several contextual factors: the stage of users' work tasks, task type, users' knowledge of task topics, to explore whether or not taking account of task stage, task type, and topic knowledge could help predict document usefulness from the time that users spend on the documents. This study also investigates whether or not expanding queries with important terms extracted from task products and useful pages improves search performance. To these ends, a controlled lab experiment was conducted with 24 student participants, each coming three times in a two-week period to work on three sub-tasks in a general work task. Data were collected by logging software that recorded user-system interaction and questionnaires that elicited users' background information and perceptions on a number of aspects. Observations in the study and examinations of the data found that the time users spent on documents could have three different types: total display time, total dwell time, and decision time,

which had different roles in working as a reliable indicator of document usefulness. Task stage was found to help interpret certain types of time as reliable indicators of document usefulness in certain task types, so was topic knowledge, and the latter played a more significant role when both were available. This study contributes to a better understanding of how information seeking behaviors, specifically, time that users spend 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 interpret time as an indicator of usefulness. These findings have theoretical and practical implications for using behaviors and contextual factors in the development of personalization systems. Future studies are suggested on making use of these findings as well as research on related issues.

ACKNOWLEDGMENTS

I cannot imagine this dissertation being finished without the support from many people. It is a great pleasure to thank them all.

My deepest appreciation goes to Nick Belkin. I am most grateful for his supervision, guidance, and support for both the intellectual pursuit of my doctorate and the building of a career. He is an exceptional person as well as researcher.

My thanks go also to my other committee members, Xiangmin Zhang and Jacek Gwizdka, for their continued advice and support throughout my doctoral study. I thank my outside member, Diane Kelly, for her contribution and support to my dissertation research, and for her mentoring, encouragement and advice in thinking about my future research and career.

I would like to thank Dan O'Connor for his always generous help and support, including supervision on several research practica. I thank Gretchen Chapman for her supervision of my Cognitive Science research practicum and research article collaboration. I thank Tefko Saracevic for his invaluable advice and encouragement at times.

I thank other professors at the SC&I who have provided intellectual and other support: Claire McInerney, Nina Wacholder, Stew Mohr, Maria Dalbello, Paul Kantor, Harty Mokros, Marie Radford, and Craig Scott. I thank Joan Chabrak, who has helped me on various issues throughout the years. Others at the SC&I have also provided various kinds of support including: Jon Oliver, Karen Knovick, Louise Forman, Elizabeth Ciccone, Cecilia Gal, Nick Diakopoulos, Marsha Bergman, Loretta Reda, and the IT staff.

My dissertation research has benefited from the generous and insightful comments of researchers at various doctoral colloquia and conferences. They include: Bruce Croft, Susan Dumais, Gary Marchionini, Barbara Wildemuth, Doug Oard, Michael Nelson, Miles Efron, Cathy Blake, Michael Twidale, Pertti Vakkari, Leif Azzopardi, Eugene Agichtein, Soo Young Rieh, and four anonymous reviewers from the SIGIR 2010 conference. Other departments at Rutgers and other institutions have contributed to my intellectual growth: Eviatar Zerubavel, Georghe Muresan, Zenon Pylyshon, Alvin Goldman, and Ken Shan.

I also want to thank many of my friends and colleagues for their intellectual, emotional, and technical support and collaboration: Yuelin Li, Xiaojun Yuan, Chang Liu, Michael Cole, Ying Sun, Ying Zhang, Jing Ning, Ralf Bierig, Jun Zhang, Irene Lopatovska, Catherine Smith, Teresa Keeler, Sung Un Kim, and Colleen Cool.

My heartfelt thanks go to my family and relatives who have provided invaluable and irreplaceable support to the accomplishment of my dissertation. Thank you very much, my husband and son, parents, parents-in-law, brother, cousin-in-law and his wife, and uncle- and aunt-in-law.

The LIS department and the SC&I have generously offered support through a fellowship and a graduate assistantship. The Institution of Museum and Library Services supported my dissertation research under grant #LG-06-07-0105-07.

Last, but certainly not the least, I would like to thank all my research participants. This research cannot be done without their help.

DEDICATION

To

My husband & son, Guocan & Andy Wang

My parents, Shiliang Liu & Xinti Hou

My parents-in-law, Huaqing Wang & Lianyi Chen

And my brother, Xiangyi Liu

Thank you all for your love, understanding, and whole-hearted support.

TABLE OF CONTENTS

Abstract of dissertation ii

Acknowledgements iv

Dedication vi

List of tables xii

List of Illustrations xv

Chapter 1. Introduction 1

Chapter 2. Literature Review 4

2.1 Problematic situation and knowledge acquisition 4

2.2 Contextual/situational factors in IR studies 6

2.2.1 Task and personalization 8

2.2.2 Knowledge and personalization 22

2.2.3 Desktop repository as a source of implicit relevance feedback 28

Chapter 3. Theoretical Stance 33

3.1 Research model 33

3.2 Situational variables considered in the current study 37

3.2.1 Task type classified by task structure 38

3.2.2 Stage of task 38

3.2.3 Topic knowledge 39

3.2.4 Task product 40

3.3 Research Questions 41

Chapter 4. Methodology 44

4.1 Controlled lab experiment for data collection 44
4.2 General Study Design 45
4.2.1 Operationalization of task stage 45
4.2.2 Experimental design 46
4.2.3 Data needed and collection methods 47
4.3 Experiment components 48
4.3.1 Tasks 49
4.3.2 Task orders 51
4.3.3 Participants 53
4.3.4 Study location and computer equipment 54
4.3.5 Search systems 54
4.3.6 Logging software 58
4.3.7 Consent form 59
4.3.8 Other data collection supporting materials 59
4.4 Experiment procedure 62
Chapter 5. Results 65
5.1 Characteristics of the participants 65
5.2 Characteristics of the data 67
5.3 Various types of time 69
5.3.1 Dwell time 71
5.3.2 Total dwell time 71
5.3.3 Display time 71
5.3.4 Total display time 71

5.3.5 Decision time 72
5.4 Results of RQ1 72
5.4.1 Sub-question 1a 73
5.4.2 Sub-question 1b 80
5.4.3 Sub-question 1c 84
5.4.4 Summary of results of RQ1 88
5.5 Results of RQ2 91
5.5.1 Sub-question 2a 92
5.5.2 Sub-question 2b 100
5.5.3 Sub-question 2c 107
5.5.4 Sub-question 2d 112
5.5.5 Summary of the above results 119
5.5.6 Sub-question 2e 121
5.5.7 Answers to RQ2 and sub-RQs 125
5.6 Results of RQ3 127
Chapter 6. Discussion 129
6.1 Three Types of Time and Their Usefulness in Modeling Users and Personalizing Search 129
6.2 Time, task stage, and document usefulness 130
6.2.1 Time as an indicator of Usefulness in the Stage Model 130
6.2.2 Stage as a Helpful Contextual Factor in Inferring Usefulness 133
6.2.3 Task Type as a Helpful Contextual Factor in Inferring Usefulness 136
6.3 Time, topic knowledge, and usefulness 137

6.3.1 Topic knowledge patterns across 3 stages 137
6.3.2 Time as an Indicator of Usefulness in the Topic Knowledge Model 140
6.3.3 Topic Knowledge as a Helpful Contextual Factor in Inferring Usefulness 142
6.3.4 Task Type as a Helpful Contextual Factor in Inferring Usefulness 144
6.3.5 The relation between time and topic knowledge 146
6.3.6 Comparison of the roles of task stage and topic knowledge in interpreting time as an indicator of usefulness 146
6.3.7 Implications of these findings 150
6.4 How Findings Relate to the Theoretical Model 154
Chapter 7. Conclusions 157
7.1 Summary of this Dissertation Research 157
7.2 Implications for System Design 161
7.3 Limitations and Future Studies 162
APPENDICES 166
A. Recruitment Notice 166
B. Consent Form 167
a. General consent form 167
b. Consent for audio and video recording 169
c. Consent for Further Use of Recorded Data 170
C. General Instructions 171
D. Session Instructions 172
E. General Task 173
F. Note sheet 175

G. *Background questionnaire* 176

H. *Pre-session task questionnaire* 179

I. *Pre-session sub-task questionnaire* 180

J. *Usefulness Evaluation Questionnaire* 181

K. *Post-session sub-task questionnaire* 182

L. *Post-session task questionnaire* 184

M. *Exit Interview* 185

N. *Receipt* 187

O. *Codes for left-side panel in QE interface* 188

Bibliography 191

Curriculum Vitae 198

LIST OF TABLES

Table 4.1 Comparison of lab experiment vs. naturalistic & longitudinal study 45

Table 4.2 Study design..... 47

Table 4.3 Data collection summary 48

Table 4.4 Participants' condition assignment 52

Table 5.1 Summary of saved documents in all sessions..... 68

Table 5.2 Summary of mean pages per session in each stage 69

Table 5.3 GLM Univariate results of the effects of usefulness, task stage, and their interaction on mean log(10) of total display time in both tasks..... 77

Table 5.4 GLM Univariate results of the effects of usefulness, task stage, and their interaction on mean log(10) of total dwell time in both tasks combined..... 78

Table 5.5 GLM Univariate results of the effects of usefulness, task stage, and their interaction on mean log(10) of decision time in both tasks combined 79

Table 5.6 GLM Univariate results of the effects of usefulness, task stage, and their interaction on mean log(10) of total display time in the dependent task 81

Table 5.7 GLM Univariate results of the effects of usefulness, task stage, and their interaction on mean log(10) of total dwell time in the dependent task..... 82

Table 5.8 GLM Univariate results of the effects of usefulness, task stage, and their interaction on mean log(10) of decision time in the dependent task 83

Table 5.9 GLM Univariate results of the effects of usefulness, task stage, and their interaction on mean log(10) of total dwell time in the parallel task 85

Table 5.10 GLM Univariate results of the effects of usefulness, task stage, and their interaction on mean log(10) of total dwell time in the parallel task 86

Table 5.11 GLM Univariate results of the effects of usefulness, task stage, and their interaction on mean log(10) of decision time in the parallel task 87
Table 5.12 Summary of the *F*(p) values of factors (obtained from GLM analyses) 88
Table 5.13 Mean and standard deviation of four types of topic knowledge at 3 stages ... 92
Table 5.14 Mean and Standard Deviation of four types of knowledge in two types of tasks at 3 stages 95
Table 5.15 GLM Univariate results of the effects of usefulness, topic knowledge, and their interaction on mean log(10) of total display time in both tasks combined. 102
Table 5.16 GLM Univariate results of the effects of usefulness, topic knowledge, and their interaction on mean log(10) of total dwell time in both tasks combined ... 104
Table 5.17 GLM Univariate results of the effects of usefulness, topic knowledge, and their interaction on mean log(10) of decision time in both tasks combined 105
Table 5.18 GLM Univariate results of the effects of usefulness, topic knowledge, and their interaction on mean log(10) of total display time in the dependent task 109
Table 5.19 GLM Univariate results of the effects of usefulness, topic knowledge, and their interaction on mean log(10) of total dwell time in the dependent task 110
Table 5.20 GLM Univariate results of the effects of usefulness, topic knowledge, and their interaction on mean log(10) of decision time in the dependent task 111
Table 5.21 GLM Univariate results of the effects of usefulness, topic knowledge, and their interaction on mean log(10) of total display time in the parallel task 114
Table 5.22 GLM Univariate results of the effects of usefulness, topic knowledge, and their interaction on mean log(10) of total dwell time in the parallel task 116

Table 5.23 GLM Univariate results of the effects of usefulness, topic knowledge, and their interaction on mean log(10) of decision time in the parallel task 117

Table 5.24 Summary of the *F*(p) values of factors (results of GLM analyses) 119

Table 5.25 GLM results when both task stage and topic knowledge was considered.... 125

Table 6.1 Summary of indicators of usefulness..... 133

Table 6.2 Summary of indicators of document usefulness in the topic knowledge model 142

Table 6.3 Frequency of knowledge levels by task stages in the parallel task..... 147

Table 6.4 Frequency of knowledge levels by task stages in both tasks combined 147

LIST OF ILLUSTRATIONS

Figure 3.1 A research model: factors and relations in an IR episode 35

Figure 4.1 Search interface for the 2nd and 3rd sessions in QE condition 56

Figure 4.2 Experiment procedure 64

Figure 5.1 All 24 participants' computer and searching expertise (x-axis: participant number; y-axis: level) 66

Figure 5.2 All 24 participants' online searching experience (in years) (x-axis: participant number; y-axis: year) 67

Figure 5.3 Summary of the number of viewed pages (with usefulness scores) for all participants in 3 sessions (x-axis: participant number; y-axis: number of viewed pages) 69

Figure 5.4 Different types of time 71

Figure 5.5 The distribution of total display time in both tasks 75

Figure 5.6 The distribution of total display time in the dependent task 75

Figure 5.7 The distribution of total display time in the parallel task 75

Figure 5.8 The distribution of log(10) total display time in both tasks 75

Figure 5.9 The distribution of log(10) total display time in the dependent task 75

Figure 5.10 The distribution of log(10) of total display time in the parallel task 76

Figure 5.11 The distribution of original usefulness data in both tasks 76

Figure 5.12 The distribution of combined usefulness data in both tasks 76

Figure 5.13 Relations between usefulness, task stage, and log(10) of total display time in both tasks combined 77

Figure 5.14 Relations between usefulness, task stage, and log(10) of total dwell time in both tasks combined 78
Figure 5.15 Relations between usefulness, task stage, and log(10) of decision time in both tasks combined 79
Figure 5.16 The distribution of original usefulness data in the dependent task 81
Figure 5.17 The distribution of combined usefulness data in the dependent task 81
Figure 5.18 Relations between usefulness, task stage, and log(10) of total display time in the dependent task 82
Figure 5.19 Relations between usefulness, task stage, and log(10) of total dwell time in the dependent task 83
Figure 5.20 Relations between usefulness, task stage, and log(10) of decision time in the dependent task 84
Figure 5.21 The distribution of original usefulness data in the parallel task 85
Figure 5.22 The distribution of combined usefulness data in the parallel task 85
Figure 5.23 Relations between usefulness, task stage, and log(10) of total display time in the parallel task 85
Figure 5.24 Relations between usefulness, task stage, and log(10) of total dwell time in the parallel task 86
Figure 5.25 Relations between usefulness, task stage, and log(10) of decision time in the parallel task 87
Figure 5.26 Relations of time, usefulness, and stage 89
Figure 5.27 Pre- and post-session general task topic knowledge across 3 stages 93
Figure 5.28 Pre- and post-session sub-task topic knowledge across 3 stages 94

Figure 5.29 Pre- and post-session general task topic knowledge across 3 stages in both tasks..... 96
Figure 5.30 Pre- and post-session sub-task topic knowledge across 3 stages in both tasks 96
Figure 5.31 Distribution of the original usefulness data..... 101
Figure 5.32. Distribution of the original usefulness data..... 101
Figure 5.33 Distribution of the combined topic knowledge data..... 102
Figure 5.34 Distribution of the original topic knowledge data..... 102
Figure 5.35 Relations between usefulness, topic knowledge, and log(10) of total display time in both tasks combined 103
Figure 5.36 Relations between usefulness, topic knowledge, and log(10) of total dwell time in both tasks combined 104
Figure 5.37 Relations between usefulness, topic knowledge, and log(10) of decision time in both tasks combined..... 106
Figure 5.38 Distribution of the original usefulness data in the dependent task..... 108
Figure 5.39 Distribution of the combined usefulness data in the dependent task..... 108
Figure 5.42 Relations between usefulness, topic knowledge, and log(10) of total display time in the dependent task..... 109
Figure 5.40 Distribution of the combined topic knowledge data in the dependent task. 109
Figure 5.41 Distribution of the original topic knowledge data in the dependent task.... 109
Figure 5.43 Relations between usefulness, topic knowledge, and log(10) of total dwell time in the dependent task..... 111

Figure 5.44 Relations between usefulness, topic knowledge, and log(10) of decision time in the dependent task 112
Figure 5.45 Distribution of the combined usefulness data in the parallel task 113
Figure 5.46 Distribution of the original usefulness data in the parallel task 113
Figure 5.49 Relations between usefulness, topic knowledge, and log(10) of total display time in the parallel task 114
Figure 5.47 Distribution of the combined topic knowledge data in the parallel task 114
Figure 5.48 Distribution of the original topic knowledge data in the parallel task 114
Figure 5.50 Relations between usefulness, topic knowledge, and log(10) of total dwell time in the parallel task 116
Figure 5.51 Relations between usefulness, topic knowledge, and log(10) of decision time in the parallel task 118
Figure 5.52 Relations between time, topic knowledge, and usefulness 120
Figure 5.53 Relations between usefulness, stage, and decision time in both tasks 122
Figure 5.54 Relations between usefulness, stage, and decision time in the parallel task 122
Figure 5.55 Relations between usefulness, topic knowledge and decision time in both tasks 122
Figure 5.56 Relations between usefulness, topic knowledge and decision time in the parallel task 122
Figure 6.1 Frequency of knowledge levels by task stages 148
Figure 6.2 The revised IR model showing the relations between factors 156

Chapter 1. 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. During search, people usually issue short queries with an average of only 2 to 3 words (Spink et al., 2001; Beitzel et al., 2007), which often give only a partial expression of the searcher's information need. Many keyword terms are inherently ambiguous, but traditional search engines cannot effectively disambiguate them. They typically provide the same search results to the query no matter who submits it, and under what circumstances. The documents that are most relevant to a specific user are often not in the top ranks. Spending time looking for the desired document beyond the first result page or to issue new queries requires additional time and effort, which often frustrates users. Furthermore, people are often not able to articulate appropriate queries, especially when they have little or no knowledge about the topics that they are working with, namely, when they are in an Anomalous State of Knowledge (ASK) (Belkin, Oddy, & Brooks, 1982).

Such challenges call for a solution that can "personalize" search results for each user. It has been long agreed in the information retrieval (IR) research community that major improvement of IR search performance can only be accomplished by taking account of the users and their contexts, rather than through proposing new retrieval algorithms which have reached a plateau (cf., Sparck-Jones, 1995, 2000; Keenoy & Levene, 2005). The area of interactive IR has over several decades seen proposals of models that consider the context of the users and embed the idea of personalization. They include Taylor (1968) and his reference interview, Belkin, Oddy, & Brooks (1982) and their ASK model, and Dervin (1992) and her sense-making, to name

a few. However, it is not until the late 1990s that significant major research efforts arise on applying personalized search in operational IR system design. The ultimate goal of personalization for user's interaction with information is to make the interaction "as effective and pleasurable as possible" by taking account of differences among users and adapting the systems for individuals (or groups) rather than providing homogeneous access for all (Belkin, 2006). Personalized search systems gather, store, and utilize additional information about the users and/or their contexts beyond the queries that they submit to the system. They make the search results better adapted to particular users, by putting those documents users may desire on the top ranks. The focus on users and their contexts is what makes personalized search a compelling area (Pitkow et al., 2002).

Personalization in information search can have two main dimensions or facets. One involves what is being personalized, and the other addresses the types of evidence that can be used for personalization. On the former dimension, personalization can be applied to search result content with regard to the document's relevance/usefulness, interface presentation in terms of query and/or result page display means, and interaction mode being user-initiated, system-initiated, or mixed (Belkin, 2006). On the latter dimension, the various evidence sources, i.e., the additional information about user and/or context, include explicit user preferences, user behaviors from which implicit user interests can be inferred, and contextual factors that help elicit information about the user's interests, such as the user's desktop repository, surfing history, tasks that drive the users to do the search, their topic knowledge, and personal characteristics, etc.

Previous personalization studies have mostly focused on search result content, and they have adopted the implicit approach to infer user interests which has the advantage of not bothering and interrupting the users from their search. The following behavioral and contextual

factors have been used to implicitly obtain additional information about the user's interests: dwell time, browsing history, query history, and desktop repository information. Many other facets listed above remain unexplored. The current study hopes to contribute to search result content personalization by looking particularly at three contextual factors that have not yet been much studied in search personalization systems but are very likely to implicitly provide additional information on the user's interests. They are: 1) the stage in a work task, 2) the knowledge that the user has on task topic, and 3) the documents that the user has generated for the task as part of the desktop repository information. To be more specific, this study is aimed at exploring how these factors may help implicitly inferring document usefulness, in particular examining the possible interaction effects of these different characteristics and the users' behaviors on performing personalization.

Chapter 2. Literature Review

This chapter provides a review of the related literature. It first introduces the problem situation and knowledge acquisition thoughts from a phenomenological point of view, which provides a theoretical background for information seeking behavior and IR research. The main part of this chapter is a thorough review of the contextual/situational factors in IR from a personalization perspective. This includes discussions on task and personalization, knowledge and personalization, and desktop repository as a source of implicit relevance feedback. Research on predicting document usefulness from user behaviors is also discussed.

2.1 Problematic situation and knowledge acquisition

In their far-reaching work from the phenomenological perspective, Schutz & Luckmann (1973) note that human individuals have their life-plans and their knowledge accumulates during the process of reaching this overall goal. In the natural attitude, one simply takes for granted the world surrounding oneself. Every state of affairs remains unproblematic "until further notice" (Schutz & Luckmann 1973, p. 4). One's stock of knowledge consists of solutions to problems of previous experiences, the "reference schema" (Schutz & Luckmann 1973, p. 10). If a new experience fits the reference schema of an individual, it confirms the validity of the stock of his/her knowledge. Meanwhile, when an experience does not fit in one's reference schema or when the novel experience is not compatible with the reference schema, an individual's stock of knowledge becomes deficient. Thus, what has been taken for granted is brought into the problematic, and the person is in a problematic situation, or in other terms, in an ASK (Belkin, 1980) or in a situation where there is an information gap (Dervin, 1980).

In a problematic situation, the activities by which one attempts to keep one's work or life moving on are often called "tasks" (Li, 2008). Various types of tasks being frequently examined

in the information science field include search tasks and work tasks. A search task usually refers to a user's activity to 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 (Li, 2008). The relationship between the two types of tasks is that work tasks are often motivations of search tasks. It should be noted that tasks that drive people to information seeking are not restricted to those which a strictly work-related, but needs to be understood in a much broader sense, to include various sorts of non-work information seeking activities in human individuals' everyday lives. The everyday life information seeking (ELIS) has been attracting increasing research attention. Previous studies in this area have investigated aspects on 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 understand 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 life, which may be more suitable in the situation of interactive IR (IIR) (Belkin, Cole, & Liu, 2009).

Schutz & Luckmann (1973) also points out that the human individual is always in a concrete situation, and knowledge is built on sedimentations of former experiences bound to situations. Human individuals' subjective experiences of the life-world are spatially, temporally, and socially arranged, and each individual's experience is unique due to its biographical characteristic. Schutz & Luckmann's (1973) theory is of great value to information seeking, the essential activity of problem solving in one's daily life. We can interpret that work tasks constitute unique situations through which individual knowledge accumulates, and the subjective experience and behaviors vary in various work tasks and perhaps even in the work sub-tasks. We

can also interpret that an individual's knowledge in each situation is different from that in other situations. For instance, generally and roughly speaking, one's knowledge state changes after solving a problem, or even in each individual session, stage or episode during the process of seeking information to solve the problem. With the change of one's knowledge, his/her behaviors in performing the task are likely to change, as well. It is appealing to conduct empirical studies to explore how people's tasks, as well as knowledge (changes) may affect their behaviors in seeking for information to fill the information gaps in order to ultimately accomplish their tasks.

2.2 Contextual/situational factors in 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 defined so variously" (Dervin, 2003, 112), and the concepts of context and situation has been used interchangeably (Cool, 2001). Based on a thorough review of the concept of situation across 6 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 the clear description of disambiguating situation from context, the following nearly a decade so far 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. While it is not the focus here

and we by no means intend to argue whether it is a good way or not, using context as an umbrella term to refer to a variety of factors could simplify things for a certain purpose. Due to the above reasons, the following literature review of the current thesis does not attempt to differentiate the two concepts of context and situation, but uses the term context in general[1].

Contextual factors have been addressed by many researchers as important and should be taking into account in IR research and system design (e.g., Belkin, 1993; Ingwersen & Järvelin, 2005; Belkin, Muresan, & Zhang, 2004; Dumais, 2007). Task as a contextual factor has attracted fairly extensive research efforts in terms of the effect of different task features on information search and use (e.g., Bystrom & Järvelin, 1995; Vakkari, 1999; Vakkari, 2001). User knowledge, including domain knowledge and topic knowledge (i.e., topic familiarity; these two terms are used interchangeably in this thesis) have also gained some research attention with respect to their effects on information seeking behaviors (e.g., Wildemuth, 2004; Kelly & Cool, 2002). However, it is only in recent years that research findings on all these aspects are more and more effectively used, or being proposed to be used in the design of operational systems that tailor search results toward individual users. Another source of context that provides information about the user is the personal desktop repository, which has been used in the design of some recent personalization systems (e.g., Chirita et al., 2006; 2007). This section reviews related studies in three aspects of context: on the user's task, on the user's topic knowledge, and on the recent approaches that utilize the user's local desktop information in personalization. All studies reviewed here belong to the "objectified context" camp (Talja et al., 1999), treating context as an objective reality which provides a background for the study of individuals' or individual groups' behaviors.

[1] In Chapter 3, context and situation are differentiated with each other in the proposed theoretical model.

2.2.1 Task and personalization

2.2.1.1.1 Task types and user behaviors

Researchers have spent rather extensive efforts on examining the effects of different tasks on information searchers' behaviors and performance. A commonly seen basis of this stream of research is to classify user tasks into different types along some task feature(s).

There are different types of tasks based on different task features, which generated various task type definitions, for example, closed vs. open-ended tasks (Marchionini, 1989), specific tasks vs. general tasks (Qiu, 1993), fact-finding vs. information gathering (Toms et al., 2007), to name just a few. The various standards and definitions of task type classification make it difficult to compare findings across studies. Therefore, it seems necessary to have some standard classification scheme. A recent and rather extensive classification scheme is provided by Li (2008). By examining the relationships between work task and search task, the author refined a faceted classification of tasks she previously constructed (Li, 2004). Li (2008) conducted semi-structured in-depth interviews with 12 student, faculty, and staff members in a university community whose work heavily depended on information systems. The data suggested that some existing task facets were not appropriate and should accordingly be dropped from the classification, while some new task facets emerged and should be added to the classification. The refined task classification scheme includes a number of dimensions of task product, objective complexity, subjective complexity, difficulty, urgency, to name a few.

It is necessary to introduce the definitions of several dimensions in Li's (2008) classification scheme as well as in other researchers' works since there are many studies involving such task types that are to be introduced in the following part. Along the dimension of task product, Li (2008) classifies tasks into several categories including: intellectual (a task producing new ideas or findings), decision/solution (a task making a decision or solving a

problem), or the factual (a task locating facts, data, etc.), and so on. Freund (2008) classified information task types in a similar way, though she did not specify the facet or dimension in her classification scheme. The five task types include: 1) learning about a topic: trying to learn about an unfamiliar topic, which is similar to "intellectual" in Li (2008); 2) making a decision: trying to make a decision, which is similar to "decision/solution" in Li (2008); 3) finding out how to: trying to find out how to do something (there is no type in Li (2008) exactly corresponding to this type, but it could be, under some circumstances, matched to "intellectual"); 4) finding facts: trying to find specific factual information about products or technologies, which is similar to "factual" in Li (2008); and 5) finding a solution: trying to solve a problem or fix a malfunction, which is similar to "decision/solution" in Li (2008). Another approach to classifying task types is Kellar, Watters, & Shepherd (2007). They had two types of task which involve searching[2]: 1) Fact finding, in which users look for specific facts or pieces of information; and 2) Information gathering, involves the collection of information, often from multiple sources.

Another approach to task type classification is in Kim (2006), in which tasks are classified into four types: 1) factual task: task to seek for specific precise data, which is similar to Li (2008) and Freund (2008); 2) descriptive task: task to define/describe thing, event, reason, means, etc.; 3) instrumental task: task to determine what to do or how to do it, which is similar to the "finding out how to" in Freund (2008); and 4) exploratory task: task to require generalization related to facts in meaningful patterns (there is not a type in Li (2008) or Freund (2008) that can be exactly matched to this one, and it is somewhat similar to "intellectual" in Li (2008)). Knowing these definitions help understand the various uses of task types in related research.

[2] Note that they had four types of information seeking tasks. The other two, namely, browsing and transaction, are not searching tasks toward a specific goal. Being outside the focus of the current paper (i.e., searching toward a specific goal), they are not discussed here.

Li (2008) defines the degree of objective task complexity according to the number of activities in a work task or the number of information source types in a search task. Another definition of task complexity frequently seen is by Byström and her colleagues (e.g., Byström & Järvelin, 1995; Byström, 2002). In their research, task complexity was defined from the worker's point of view based on "a priori determinability of, or uncertainty about, task outcomes, process, and information requirements" (Byström & Järvelin, 1995, p. 194). Their definition is somewhat more subjective than that of Li (2008), meanwhile, it is also different from the subjective complexity in Li (2008), which is defined as user's subjective opinions on task complexity and difficulty. The following sections review previous studies that involve task types classified along the above mentioned classification dimensions.

2.2.1.1.2 Task product

Based on her classification scheme, Li (2008) investigated the relationship between work tasks and interactive information searching behavior. One task facet she examined in the classification scheme is task product (another is objective task complexity, which will be introduced in the next section). Along the task product dimension, Li (2008) designed the intellectual and the decision/solution tasks. A controlled experiment was conducted, with 24 university students who had various academic levels and backgrounds as participants. The results showed that work tasks are important factors in shaping users' interaction with information systems. The facet „Product' significantly affected the number of IR systems consulted and result pages viewed, the number of search engines consulted and web result pages viewed, the number of library resources consulted and library result pages viewed, mean query length, and success. This study demonstrated that a faceted approach to conceptualizing tasks in IR research is feasible and effective.

Liu et al. (in press) investigated user behaviors associated with different task types classified based on dimensions in Li (2008), including task product, task complexity, search task goal (quality), and an extended dimension, level. They conducted a lab experiment with 22 participants who were upper division of journalism/media studies undergraduate students. The researchers followed the simulated task scenario (Borlund, 2000) and designed 4 journalism assignments, varying according to the different classification dimensions. As for task product, there were two levels: factual and mixed (of factual and intellectual). Two factual tasks were copy editing (collect sources that verify the correctness of 3 facts) and advance obituary (collect sources needed to write an advance obituary for the artist Trevor Weekes) assignments, and the two mixed tasks were background information checking (collect all published news stories in major newspapers about the effect of 9/11 on international student visa application) and interview preparation (find two appropriate people to interview for high education budget cut and save their contact information) assignments. It was found that users spent significantly longer time to complete mixed-product tasks than factual tasks. They visited significantly more pages and more sources in mixed-product tasks than in factual tasks. However, the number of queries they issued did not show differences, nor did the number of search sources they used.

A similar approach was by Kellar, Watters, & Shepherd (2007), who conducted a field study with 21 participants and examined how users navigated and interacted with Web browsers across different information-seeking tasks. The study logged some implicit measures of users' behaviors: dwell time, number of pages viewed, and the use of specific browser navigation mechanisms. It was found that the Information Gathering task was the most complex one: participants spent more time completing it, viewed more pages, and used the Web browser functions most heavily. This study, however, did not further examine how these implicit

measures can be used to predict document usefulness.

There are some works which classify tasks, or items that are connected to tasks, based on the similar dimensions as "product" in Li (2008), and these works are also discussed here. Kelly et al. (2002) and Murdock et al. (2007) examined the relationship between IR question tasks and document features. The two types of task questions they considered were classified based on the orientation of user's information need: one was task- (or procedure-) oriented while the other was fact-oriented. Task/procedural-oriented questions usually require the users to find out how to accomplish a task, which is similar to the "finding out how to" type in Freund (2008). Fact-oriented questions require the users to simply find a fact, which corresponds to the "factual" category in Kim (2006), Li (2008), the "fact-finding" type in Kellar, Watters, & Shepherd (2007), and "finding facts" in Freund (2008). Since the two types of tasks in Kelly et al. (2002) and Murdock et al. (2007) roughly match the types classified on the dimension of task products in Li (2008), these works are put into this section in the current thesis. They found that the types of relevant documents with certain features vary according to different types of search tasks. FAQs and list occurs in more documents judged relevant to task-oriented than those judged relevant to fact-oriented. In other words, the types of search tasks may predict which types of document are relevant. This study, however, considers only the document features in inferring document usefulness but leaves out the user's behaviors, such as clicking through, saving, and printing, etc., which usually provide richer information about the particular user who is doing the search.

2.2.1.1.3 Task complexity

Besides task product, another task facet examined in Li (2008) is objective task complexity. According to the IR system and type of sources used in accomplishing the work tasks, her study designed tasks with three levels of task complexity: high, medium, and low. The

results showed that objective task complexity affected the greatest number of aspects of interactive information searching behavior, including: the number of IR systems consulted, the number of result pages viewed and items viewed, the number of search engines consulted, portals visited, web result pages and items viewed, users' interaction with library resources, all query-related interactive behavior, success, satisfaction, and time.

Liu et al. (in press) also examined user behaviors in tasks of different levels of complexity. Their 4 tasks had two levels of complexity: copy editing and interview preparation being low complexity, and background information checking and advance obituary being high complexity. It was found that users spent significantly longer time to complete high complexity tasks than low complexity tasks. They visited significantly more pages and more sources, issued significantly more queries, and used significantly more search sources in high complexity tasks than low complexity tasks. The findings were consistent with those in Li (2008).

White, Ruthven, & Jose (2005) examined the influence of some factors, including task complexity, on the utility of relevance feedback, especially implicit relevance feedback (IRF), as opposed to explicit relevance feedback (ERF). They defined task complexity according to the number of potential information sources and type of information required to complete a task. They designed tasks with three levels of complexity: high, middle, and low. It was found that the users preferred IRF for more complex tasks, but they preferred ERF for less complex tasks. This study implies that in order to avoid task bias, task complexity should be taken into account when designing systems involving IRF or ERF. Since different types of RF are appropriate for tasks with different levels of complexity, it would be beneficial to use both types of RF simultaneously in a system which can automatically detect task complexity and switch between the two modes of RF. Such automatic detection and adaptation will be supported by observations of users'

behaviors interacting with IR systems and sets of criteria or parameters to model users' behaviors, which call for more research.

Other researchers have also looked at task complexity, although they employed a different definition of task complexity. Byström and her colleagues conducted a series of studies analyzing the effect of task complexity on human information behaviors, specifically, the relationships between task complexity, information types, and information sources (e.g., Byström & Järvelin, 1995; Byström, 2002). In their research, task complexity was defined from the worker's point of view based on "a priori determinability of, or uncertainty about, task outcomes, process, and information requirements" (Byström & Järvelin, 1995, p. 194). In other words, the more familiar a task worker is with the task requirement, the less complex a task is perceived. This definition is rather subjective as opposed to Li (2008) and White, Ruthven, & Jose (2005), both of which attempted to take a rather "objective" way to define complexity by operationalize it into some measurable variables. It should be noted that although being subjective in nature, the complexity following Byström's definition is still different from task difficulty (e.g., Gwizdka & Spence, 2006; Gwizdka, 2008), another subjective measure frequently used to assess users' perceptions to the tasks.

The tasks in their study were categorized into several groups based on the different levels of complexity. With the increasing complexity, they are: automatic information processing tasks, normal information processing tasks, normal decision tasks, known, genuine decision tasks, and genuine decision tasks. Information types were classified into three categories: problem information, domain information, and problem-solving information. They collected data in municipal administration settings by means of both questionnaires and diaries. Though the definitions in these studies are a bit different from Li (2008), they obtained similar findings. The

findings indicate that task complexity is related to information types and information sources selection. As people's task complexity increases, they need more types and more sources of information, need more domain information and more problem solving information, are less likely to predict the types of information they need, and are more dependent upon experts to provide useful information. In summary, these studies all show that task complexity as a facet of task type have impact on user's searching behaviors.

2.2.1.1.4 Stage of task

Stage of task is another general feature of the task that has received careful investigation regarding the information seeker's affective, emotional, and physical action changes during the information seeking process. Li (2008) does not include this clearly as one dimension, and this is a reasonable expansion to Li's (2008) classification scheme.

Kelly's (1963) construct theory depicts the process of construction as occurring in six different phases when the 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 ISP (Information Seeking Process) model, Kulthau (1991) proposes a model of the information search process as proceeding in 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 indicated that stage of task may be an important factor that relates to the user's judgment of document usefulness.

Vakkari and colleagues have a series of papers (e.g., Vakkari, 2000; Vakkari, 2001; Vakkari & Hakala, 2000) describing their research on exploring 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. The participants were 11 master students preparing a research proposal. They engaged in an IR search three times during the course, in the beginning, middle, and ending points respectively. The study used pre-search interviews to elicit the users' stages and post-search interviews to elicit their relevance judgments and the reasons for their judgments. The users' interactions with the system were logged. It was observed that the user's vocabulary changed from broader to narrower terms. As for the search tactics, it was found that when the task stage progresses, 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 are suggestive in designing personalization systems that could provide tools helping users build their conceptual structure in the initial stage of tasks, although it should be noted that the authors did not show the statistical significances of the changes in relevance criteria.

Similarly, Taylor et al. (2007) believed that the results of the user's cognitive changes during the IR process are partially revealed through the changes in relevance criteria choices over the ISP, and they also examined the relationships between the users' relevance assignments on the retrieved documents and the stages in the process of completing their tasks. The data came from a convenience sample consisting of 40 undergraduate students in an introductory computer science course at a US university, who were required to write an essay on a topic of general

computer science interest using at least 5 sources. For each of the sources they found, they were asked to fill out a questionnaire eliciting their opinions on whether or not they would use it and why, as well as how far along they were (i.e, which stage they were) in doing the essay assignment. The reasons leading to their usefulness judgments were open ended and were coded into multiple categories. The stages were pre-set choices in the questionnaire, including picking a specific topic, learning about the topic, formulating thoughts, and writing the paper. These stages corresponded to those of selection, exploration, formulation, and presentation in Kulthau's (1993) ISP model. The results showed statistically significant relationships, which Vakkari & Hakala (2000) lacks, between the users' stages in the search process and relevance categories chosen. The finding concurs with and adds details to previous studies such as Vakkari & Hakala (2000). This branch of research demonstrated the differences in user's relevance judgment in different stages of the task, however, it leaves as an open issue how such differences could be modeled through the user's behaviors.

Another issue is that it is not always easy to split these stages exactly and accurately in empirical research because such stages do not often or necessarily have apparent borders or lines, especially when the user is not involved in explicitly expressing such stages. In their study analyzing the effect of relevance feedback, White, Ruthven, & 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 ERF is used more toward the end. They further found that task complexity affects when the user gets to the most interactive point in using the IRF based system. In other words, for the more complex tasks, users may spend more time initially interpreting search results before they interact with them.

Tombros, Ruthven, & Jose (2004) used a similar posterior way to determine the initial and ending sessions of the user's progress. Under the general idea to investigate the criteria used by online searchers when assessing the relevance of Web pages for information seeking tasks, they looked specifically into how searchers' criteria evolved during the different stages of tasks. They invited 24 participants, each being asked to search on three information-seeking tasks. They identified the stages in the users' task progress by identifying the first and last sets of Web documents that the users visited. The data showed that the users' criteria along the duration of a task displayed a certain degree of variation, especially for the tasks that the users had a higher perception of task completion.

Another way to operationalize the stage of task is Lin (2001), which manipulates the user's task with different sub-tasks to be completed in different search sessions. In his study, a task scenario is designed which requires the participants to finish a task that requires making a vacation plan. This plan is accomplished through three steps/sessions: to identify candidate places for the trip, to compare the different candidate places and decide on one place to go, and to make a plan for the trip. Both the ways in the White, Ruthven, & Jose (2005) and the Lin (2001) experiments that operationalize task stage are arbitrary to some extent, but the latter is more close to the situation in people's daily life solving complex tasks.

2.2.1.1.5 Task structure

If the work task consists of multiple sub-tasks, the relationship between the sub-tasks seems necessary to be taken into account because the task-doer could take different orders of these sub-tasks during the process of accomplishing the work task. This dimension is not included in the original work of Li (2008), but a similar idea can be found in some works. For example, Toms et al. (2007) classifies 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 (2008) by adding to its classification scheme a new dimension of task structure.

2.2.1.1.6 Summary

As can be seen from the above review, there is a rich body of literature on task as a contextual factor in terms of how it affects users' information search behavior and performance. The works reviewed in this section classifying tasks along the different dimensions build foundations for the current research looking at the stage of task and task structure in personalization. However, further research efforts are needed. For one thing, the previous works typically concluded by finding the behavioral or performance differences among users performing different types of tasks, and apparently further approaches are needed to make use of these findings in developing systems that utilize the information that a user's task can provide for personalization. This includes approaches on how a system can predict task types from observed users' behaviors, and how the system can provide better search results based on the prediction of task types.

On another direction, a majority of the previous studies also 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. All these variables cannot be obtained until the end of a session. While these results can be used to predict task type in general posteriori, it is not easy to make use of these findings into adaptive search. Lower level behavioral variables that can be captured and used real-time are needed, for example,

document dwell time, number of pages per query, etc.

In addition, the above reviewed studies mostly concern a two-way relationship between tasks and user behaviors, or between tasks and search performance. No consideration was taken with respect to the usefulness of the documents that the users interacted with. Document usefulness is an important element to the search system. Not only do systems want to return useful documents in 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 quickly. Being able to predict document usefulness based on user behaviors would benefit the system in personalizing search. There have been previous studies which looked 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). Previous studies that had different experiment settings have generated seemingly conflicting findings of the relationship between document reading time and preference/relevance judgment. While attempting to design a filtering system based on monitoring user behaviors, Morita & Shinoda (1994) found a strong tendency for users to spend a greater length of time reading those articles rated as interesting than those rated as not interesting. In a different setting which was interactive in nature, asking the users to perform search tasks, Kelly & 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 & Belkin (2004) further suggests that contextual factors should be taken into account, which leads to a three-way relationship examination among document usefulness, user behavior, and contextual factors. The following Section 2.2.1.2 will introduce this three-way relationship when the contextual factor is specified as task.

2.2.1.2 Task, search behavior, and document usefulness

Generally speaking, task has been found in previous research 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 & 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 differs 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/or users. This study, however, did not examine how to incorporate the contextual factors and what the actual effectiveness would be.

This pending problem was addressed by White & Kelly (2006). They explored the interactions between dwell time and the two factors of user and task. They examined if additional information from the user and/or the task helps reliably to establish a dwell time threshold to predict document usefulness, and how effective this method would be. Their results showed that tailoring display time threshold based on task information improved implicit relevance feedback algorithm performance. In other words, display time was proved to be able to successfully predict document usefulness when the task information is considered. This study is a successful case 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 & Kelly's (2006) approach to classifying tasks, they only collapsed the different everyday life tasks identified by their 7 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 work by Li (2008) which provides an extensive task classification scheme.

2.2.2 Knowledge and personalization

Another contextual factor which may be helpful in providing additional user interest information for personalization is the user's familiarity with search topic, or with search topic domain. The former is called topic knowledge, and the latter domain knowledge. There have been quite some studies examining the effect of topic or domain knowledge in IR. Some of them focus on how such knowledge influence user's search tactics, often operationalized as the search term use (e.g., Hsieh-Yee, 1993; Vakkari, 2002; Wildemuth, 2004), and others look at how it affects search performance, usually measured by precision and recall (e.g., Marchionini, 1989; Allen, 1991). The latter branch of research hardly detected relationships between domain knowledge and search performance, and no conclusions can be drawn regarding personalization of IR. Meanwhile, the former branch of studies did find behavioral differences among users with different levels of domain knowledge, and these findings often have implications in various aspects of design of personalization systems.

In the works that explore how user's domain knowledge or topic knowledge affects searching behaviors, some focused on search tactics: query formulation and reformulation, and others focused on behaviors related to the search results: document features, dwell time, etc. These two groups of works are introduced respectively.

2.2.2.1 Domain knowledge and search tactics

Hsieh-Yee (1993) reports on a study investigating the effects of subject knowledge and search experience on the tactics in online searches of novice and experienced users. The study

was conducted by using the ERIC database and the DIALOG system. Participants were two groups: 30 novice users who had little search experience and 32 experienced users who had at least one year of search experience and attended courses or workshops on searching skills. Each was given one search task inside their field (with high familiarity) and one task outside their field (with low familiarity). Results show that when users have had a certain searching experience, subject knowledge affected their searching tactics. When they work with a less familiar topic, they used the thesaurus more for term suggestion, made more effort in preparing for the search, monitored the search more closely, included more synonyms, and tried out more term combinations than when they searched a familiar subject area. The findings have two points calling for attention. First, with the increasing popularity of the Internet and search engines, the majority of web users nowadays are experienced, hence, it seems that subject knowledge should have an influence on searching behaviors in a wider user population than before, and should become a more important factor that is worth further investigation. Second, further research is needed for how to make use of such behavioral differences of the users to personalize search, for example, to infer the user's knowledge from their search tactics, or to provide them suitable system features according to their knowledge level to benefit their search.

Wildemuth (2004) sought to understand the effect of domain knowledge on search tactic formulation. In terms of search tactics, she looked at the "moves", i.e., the changes of search terms, concepts represented by the terms, and operators. This research invited 77 medical students to work with questions in the domain of microbiology. Each participant had three search occasions, one before entering a medical course, one right after the end of the course, and one six months after they finished the course. The three assessment occasions in the research design represent different levels of domain knowledge among the participants. Their search logs were

coded into moves, and it was found that low domain knowledge was associated with less efficient selection of concepts to include in the search and with more errors in the reformulation of search tactics. Again, such differences implied that it would be beneficial to provide the users who have different levels of domain knowledge with different systems or system features that support search tactic formulation and reformulation.

Vakkari, Pennanen, & Serola (2003) examined psychology students' searches at two points in their development of research proposals. Twenty-two students of psychology attending a 3-month seminar made searches in PsychINFO for preparing a research proposal both in the beginning and the end of the seminar. The study found that as students learned more about their research topics, the "clearest change in students' searching was the use of a wider and more specific vocabulary" (p. 459). It was indicated that domain knowledge affects people's ability to choose appropriate search terms.

Sihvonen & Vakkari (2004) further found that domain knowledge improves interactive query expansion assisted by a thesaurus. They conducted a study exploring how domain experts and novice in pedagogics expanded queries supported by the ERIC thesaurus. The expert group consisted of 15 undergraduates in pedagogy and the novice group of 15 students with no previous studies in this field. The results showed that the number and type of terms selected from the thesaurus for expansion by experts improved search effectiveness, whereas there were no connections between the use of thesaurus and improvement of effectiveness among novices. The differences in making use of thesaurus found in this study may offer support for personalizing search.

2.2.2.2 Topic knowledge and search result reading behaviors

The above section introduced the relationship between a user's domain knowledge and his/her search tactics, which often includes query formulation and reformulation. Some other studies look at the users' familiarity with the topics[3] and their behaviors with regard to the searching results. Such behaviors often include document features related to reading behaviors, dwell time, the ratio of saved to all viewed documents, etc.

In their participation in the TREC (Text Retrieval Conference) 2004 HARD (High Accuracy Retrieval from Documents) track, Belkin and colleagues (Belkin et al., 2004) considered the searcher's familiarity with the topic as one contextual factor that may be useful in tailoring retrieval to an individual and the individual situation. They formed three hypotheses regarding how to take account of topic familiarity from the aspects of a document's readability, concreteness score, and average number of syllables per word. The searchers with low familiarity with a topic are, they hypothesized, more likely to find the documents with high readability, high concreteness score, and low average number of syllables per word as relevant, and those with high familiarity, vice versa. Unfortunately, while they successfully prepared the groundwork, due to the time limitation, they were unable to complete the experimental process testing these hypotheses.

Kumaran, Jones, & Madani (2005) attempted to differentiate documents that match different levels of topic familiarity by document features. They defined two types of web pages: the 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 the advanced web pages which assume their readers to have sufficient background knowledge and familiarity with the key technical/important terms in the topic. A classifier was built to classify the

[3] In this dissertation, user's familiarity with task topics is treated as an operationalization of his/her "topic knowledge".

documents according to different features (e.g., stop-word, line-length) that could be predictive of assumed topic familiarity. An experiment to re-rank search results for people with lower topic familiarity showed that the classifier was effective: the portion of introductory pages at top 5 and top 10 result sets using this method were significantly higher than that in the baseline run using default search engine ranking. Their method can be effective in biasing result ranking for topic familiarity when it is known, meanwhile, this study indicated that certain features of the document could be predictive of the document being introductory or advanced, and that a user who read this document having high or low familiarity with this topic. This could be useful in implicitly inferring one's topic familiarity, which will accordingly help in personalization system design that takes account of the user's topic familiarity when it is not explicitly known.

In a different approach, Kelly & Cool (2002) sought to understand if topic familiarity can be inferred through user behaviors. They investigated the relationship between topic familiarity and two types of search behaviors: reading time and efficacy. Efficacy was measured by the ratio of the number of saved documents to the total number of viewed documents. They found that with the increase of one's familiarity with topics, his/her reading time tends to decrease and the efficacy increases. This indicated that it may be possible to infer topic familiarity implicitly from searching behavior. However, further efforts are needed in order to tell which specific documents may be predicted as useful based on topic knowledge and reading time, and/or the user's saving, viewing, and other behaviors.

While the above 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 & Belkin (2002) addressed the user's topic

familiarity. They pointed out that topic familiarity may affect the types of information search and behaviors exhibited by the user. They illustrated the likely way that topic familiarity may affect user's reading time on a document: the relationship between reading time of relevant and non-relevant documents is not simply linear, rather, it could vary in two very different ways according to topic familiarity. For those with low degree of familiarity, reading time for both relevant and non-relevant documents may be similar, but for those with high degree of familiarity, reading time for relevant and non-relevant documents may be very different. Their concept is intuitively sensible, but there has been no further research hypothesis developed or effort spent on verifying this type of relationship in a systematic way.

2.2.2.3 Assessment of knowledge

There is a need to discuss the assessment of topic and domain knowledge when talking about studies on knowledge because it is not easy to obtain such knowledge both accurately and efficiently (quickly), especially with the restriction of time limitation in a user study setting. Generally speaking, there are two main approaches in measuring such types of knowledge. Those studies that look at topic familiarity often obtain it from the users by asking them to self-report their degree of familiarity with search topics based on a Likert scale. For example, Kelly & Cool (2002) asked users in a post-search questionnaire to rate their familiarity with the topics on a 5-point scale, where 1 was for not at all and 5 was for extremely familiar. User's self-assessment of their familiarity in this way is quite subjective.

In a different approach, those looking at domain knowledge have tried relative objective means to test one's knowledge. Some of them had the users to work on tests or solve problems. For instance, Wildemuth (2004) asked the participants to answer clinical problems in microbiology and judge the correctness of the answers to obtain their knowledge on the

microbiology domain. Some other studies tried using thesauri to measure a person's domain knowledge. For example, by defining domain knowledge as concepts and relationships between concepts in a thesaurus, Croft & Das (1990) acquired users' domain knowledge using two approaches: one was to ask users to specify concepts that were related to query concepts and the type of relationship, and the other was to ask users to suggest possible related concepts, clarify relation types and validate the relationship.

There are pros and cons in both assessment methods. The self-reported familiarity with search topic or a domain is easy to conduct, but it may carry bias or be inaccurate in that it is quite subjective. There is not a unique criterion for different users to make judgments on their degrees of familiarity, and people over or under estimate themselves at times. On the other hand, to measure people's knowledge using tests is typically time consuming, and the test reliability also needs to be assessed.

2.2.3 Desktop repository as a source of implicit relevance feedback

Recent years have seen an increase in both the practice and the research on PIM (Personal Information Management), which concerns how people "acquire, organize, maintain, retrieve, use, and control information items" for everyday use to accomplish their tasks, both work related and non-work related (Jones & Teevan, 2007, p.3). Many of the information items that people deal with in their personal information space are information objects processed and stored on their personal computers. Such a collection of personal information on one's computer, also called desktop repository, usually consists of web pages and other documents that a user saves, the browser history, and email messages, etc. Typically, PIM research and practice helps people better manage their personal information, easily locating "Stuff I've Seen" (Dumais et al., 2003) and "keeping found things found" (Jones, 2007). Moreover, this branch of research and

practice can also contribute to regular IR in the Web. On the one hand, Komlodi and colleagues have done research on how to use search histories in assisting IR research through building supporting interface tools (Komlodi, Soergel, & Marchionini, 2006; Komlodi, Marchionini, & Soergel, 2007). On the other hand, being an archive recording one's activities in everyday life, the desktop repository provides rich information implying the user's interests and therefore is a good source of evidence for personalization for the user's regular information search, for example, to help people generate (c.f., Gwizdka, 2006) or refine queries (e.g., Teevan, Dumais, & Horvitz, 2005; Chirita, Firan, & Nejdl, 2006; Chirita, Firan, & Nejdl, 2007).

Komlodi and colleagues (Komlodi, Soergel, & Marchionini, 2006; Komlodi, Marchionini, & Soergel, 2007) investigated the use of search history in legal information seeking. They observed and interviewed attorneys or law librarians for their search and use of search history. They found that searchers need historical information in information seeking, and indeed, searchers rely heavily on their memory and external memory aids when searching for information. The findings encouraged the design of user interface tools and guidelines building on search history information, for example, the inclusion of a structured and editable search history display in the search screen. Other recommendations included integrating search history information in other search system displays such as result lists, creating tools for planning, managing search histories and results, and saving relevance judgments, all of which are obviously based on individuals' specific situations and are approaches to search personalization.

Along another line, a number of recent studies have attempted to use desktop repository for search personalization from a different perspective. They use the repository as a source of personal interest and extract useful terms for query expansion. In Teevan, Dumais, & Horvitz's (2005) study personalizing a user's current Web search, they considered the user's prior

interaction with a wide variety of content and used this prior interaction as source of implicit feedback. The source included previously issued queries and previously visited Web pages, as well as some desktop repository information such as documents and emails the users have read and created. A user study was conducted to evaluate the usefulness of the personalized search system. Fifteen participants evaluated the top 50 Web search results from MSN Search for approximately 10 self-selected queries each. The results showed that the personalized system improved search performance compared with non-personalized baseline system. In terms of the desktop information, specifically, it was found that using the user's entire desktop index led to the best performance, followed by using the recently indexed content (within the last month) and then the indexed Web page content only. The authors discussed that the richness of the information used for representing user interest is important in achieving the best performance. In addition, this study revealed that the combination of the Web ranking and the personalized ranking yielded a significant improvement over either individual ranking method.

Chirita, Firan, & Nejdl (2006) also attempted to utilize the PC desktop to capture the user's interests. They investigated the opportunities to select personalized query expansion terms for Web search using three different desktop oriented approaches: summarizing the entire desktop data, summarizing only the desktop documents relevant to each user query, and applying natural language processing techniques to extract compounds from relevant desktop resources. An evaluation study was conducted with 15 participants involved. It was found that personalized query expansion terms by summarizing only those relevant to each user query are more effective than summarizing the entire desktop data. This seems to be inconsistent with Teevan, Dumais, & Horvitz's (2005) finding that the richer the representation, the better the performance. A closer review showed that the two studies used different sets of partial desktop information. Teevan,

Dumais, & Horvitz (2005) considered the desktop information in the last month and the viewed Web pages only, which can be viewed as only "partially" desktop repository but not necessarily "relevant"; in comparison, Chirita, Firan, & Nejdl (2006) looked at only "relevant" resources to the user's query, therefore, it is not surprising that they found the relevant fraction outperformed the entire desktop repository as IRF sources.

Chirita, Firan, & Nejdl (2007) conducted another study using the personal collection of text documents, emails, etc., to expand user-issued queries. This study investigated five techniques of various granularity levels for generating the keywords used for expanding queries. Three of the five techniques were local analysis techniques focused only on those desktop documents best matching the user's query, and the extracted expansion keywords were the most relevant terms, compounds, and sentence summaries from these documents. The other two techniques were global desktop analysis considering the entire desktop repository, for which term co-occurrences and the external thesauri were used in the query expansion process. The evaluation study invited 18 participants, each installing on their machine the experimental personalization system which indexed all their locally stored content. Each participant was asked to choose 4 queries related to their everyday activities, one of which is ambiguous, i.e., they thought to have at least three meanings. For each query, the researcher collected the top-5 URLs generated by the above mentioned 5 techniques and shuffled them into a set, and then asked the participants to judge each URL's relevancy. The results demonstrated performance improvement of some techniques, especially on ambiguous terms. In many cases, the simple desktop term frequency and lexical compounds with the local desktop analysis performed best. This is consistent with Chirita, Firan, & Nejdl (2006) that the most relevant desktop documents are more effective in providing user interests than the general approach.

While these personalization approaches receive promising results overall, it was also found that personalization may hurt search performance in some cases, especially when the queries are less ambiguous or more navigational (e.g., Dou, Song, & Wen, 2007; Teevan, Dumais, & Liebling, 2008). One approach to solving such problems is to detect under what circumstances personalization should be used, yet, it should also be useful to study what personalization algorithms are less vulnerable to circumstances, or more robust to improve search performance. Therefore, further research efforts are needed to identify how desktop information can be better utilized to obtain better personalization performance.

Chapter 3. Theoretical Stance

This chapter presents the theoretical framework for this dissertation. A research model for personalization in IR is proposed which addresses the relationships among user behaviors, contextual/situational factors, and document usefulness. This is followed by discussion of the situational variables considered in the current study, including task type, stage of task, topic knowledge, and task product. This chapter ends with the description of 3 research questions, each of which includes several sub-questions.

3.1 Research model

Personalizing search results for a specific user requires that the IR system understands the user's information need, interest, or preference from possible alternative sources beyond the user-issued query(ies). Predicting which documents are useful for a user often involves understanding the user's goal, general contexts and/or specific situations, as well as the use's search behaviors within the current search episode. This requires understanding of a three-way relationship among document usefulness, user behaviors, as well as the contextual factors. The following model (Figure 3.1) illustrates various factors that are key elements in an IR episode which can provide significant evidence for personalization, as well as the basic relationships and interactions among these factors.

Five sets of elements are included in this model, which are life-world goal, general contextual factors, specific situational variables, user behaviors, and document usefulness. From the personalization perspective, the five sets of elements belong to three major categories which constitute a three-way relationship: on one side is document usefulness, which is the core value that a personalization system tries to learn, predict or infer; on a second side is user behavioral

information, which is what users do and which can be observed by systems; and on a third side are goal, general contextual factors, and specific situational variables, which set and convey the background and environmental information of the users who are conducting information search. This third category of elements can be learned by the system, too. In the third category, life-world goal is the ground on which the IR activity is conducted, contextual factors convey framework meanings of a user's environment, and situational variables are the concrete factors that can be used to study the three-way relationship by statistical methods.

As mentioned previously in Section 2.1, Schutz & Luckmann (1973) note that a human individual has his/her life plan. When an individual's knowledge is insufficient to solve the new experience, he/she is in a problematic situation, and therefore has a "task" for which the information seeking activity is originated. The model represented in Figure 3.1 describes the case of a single information seeking activity. Goal in this model is not necessarily the overall life-world plan, but can also be very specific, on any level and in any aspect in life, either work-related or non-work related. Clearly, this goal that drives the user to seek information is closely related to one's knowledge and task, as shown in Figure 3.1.

The concepts of *context* and *situation* are not used interchangeably in this research as they are sometimes used in the Information Science literature, but they are differentiated from each other. Context is a more general concept describing the abstract meaning of a users' environment, while situation is a more specific concept and is "the dynamic aspect of context" (Cool, 2001). A context may consist a variety of situations (Sonnenwald, 1999), and situation is used as "an important unit variable" to specify context as a framework of meaning (Cool, 2001). For example, the general concept of the contextual factor "knowledge" can have several specific situations: topic knowledge, domain knowledge, search knowledge, etc. Likewise, the general

concept of the contextual factor "task" can also have a number of situational variables, for instance, according to the various task features: complexity, task product, task stage, etc.

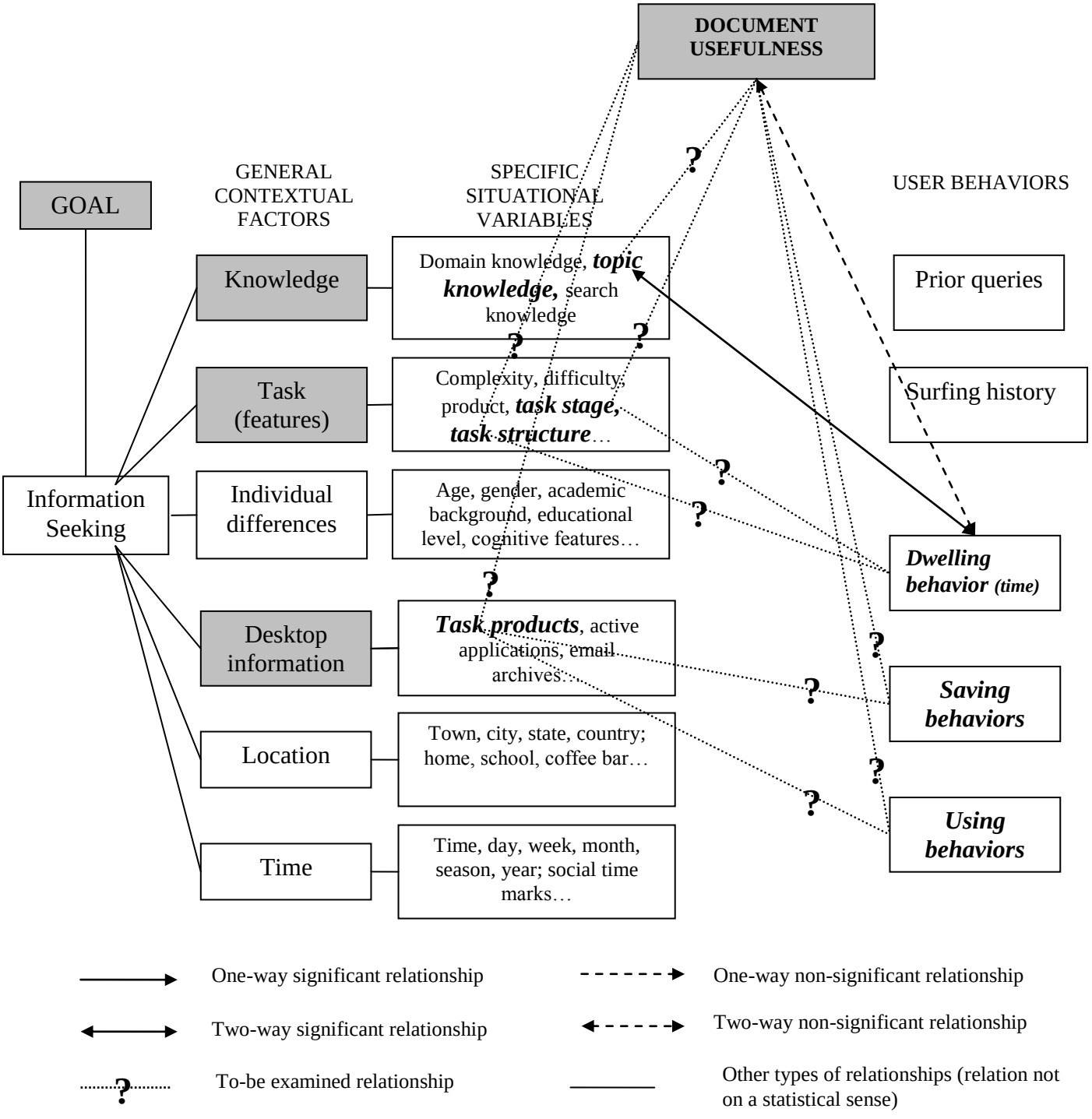


Figure 3.1 A research model: factors and relations in an IR episode

As shown in Figure 3.1, the factors in the general context column indicate an individual's general contexts. Schutz & Luckmann (1973) noted that a human individual is in a problematic situation when his/her stock of knowledge is deficient to cope with the new experience in the everyday life. This point indicates that knowledge and task (what brings an individual to information seeking) are crucial factors to describe an individual's general context. Schutz & Luckmann (1973) also point out that each individual's experience of the life-world is unique and is spatially, temporally, and socially arranged. This implies that individual characteristics, time, and location should also be important factors defining one's context. Moreover, as is known, in the digital environment, one's desktop repository keeps records of one's activities on their computer, which is certainly a part of their life-world and therefore is also included in the model as a valuable source describing one's general context.

At a lower level are dimensions defining an individual's specific situation. These dimensions are detailed components that specify the general contextual factors. Values of these dimensions on this level are able to identify an individual's specific situation in an IR episode. Those of particular interest to this research include domain knowledge, task stage, sub-task structure, and task products generated in the previous stages of the current search episode. These variables, as well as the interaction between/among them, are possible sources from which the system can learn about a user's preference, in other words, a document's usefulness to the user. More details on these dimensions will be introduced in Section 3.2.

Another set of components in this model are user behaviors. 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, 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 specific situation in the IR episode. For example, a user's dwell time on a retrieved information object may be the result of his level of domain knowledge, and/or his 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, i.e., 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 hopefully tell how this document is useful to him/her. Such interaction effects have rarely been studied in personalization, but it could be important even in the intuitional sense, therefore, research is needed to examine such interaction effects.

3.2 Situational variables considered in the current study

There are four situational variables in Figure 3.1 that were examined in this research: task structure, stage of task, domain knowledge, and task product. These factors have been studied, to some extent in previous research, with regard to their relations with search performance and/or user behaviors. Nevertheless, there has been no research, to our knowledge, that considers the interaction among these factors on user's search behaviors. The current study attempts to explore such interaction effects, if there are any, on user's behaviors, and how such interaction effects may help personalization for IR. This section introduces the rationales of these four situational variables.

3.2.1 Task type classified by task structure

This study considers the different types of tasks classified by sub-task relationships, i.e., task structure. Using a similar way to Toms et al. (2007), two basic types of tasks are conceptualized: the parallel and the dependent. Sub-tasks could be in parallel to each other, and the accomplishment of one is not necessarily based on that of others. The knowledge needed for, and acquired after, one sub-task is not necessarily based on that of others. Accordingly, the order of the sub-tasks is not fixed but rather can vary. On the other hand, some sub-tasks could be dependent upon others, and the accomplishment of one is usually based on that of others. In this case, the knowledge needed for, and acquired after, one search sub-task is usually built on that of the previous ones. The order of the sub-tasks in such a task is usually fixed.

Obviously, there are other relatively more complicated relationships among sub-tasks. For example, the sub-tasks could have some relations combining both the parallel and the dependent: some sub-tasks are in parallel but are meanwhile dependent on others in the same task. This study, however, considers only the two simple types of sub-task relationships, i.e., the parallel and the dependent. Doing so is not only because this can easily detect the effects (or differences of the effects) of these two types of tasks on the users' behaviors, but also because these two are in fact the most basic types. Other types of tasks could be examined in future studies.

3.2.2 Stage of task

Based on the review of the related literature on tasks, it turns out that task stage has been well studied in terms of its relationship with the user's search tactic and other behaviors, and previous research did find that user's behaviors and cognitive status (e.g., relevance judgment criteria) vary along different stages in the search task (e.g., Vakkari, 2000; Taylor et al., 2007). It would be interesting to see how the information of task stage can be used for personalizing

search results. One question of interest is to examine whether or not task stage, as a contextual factor, can provide useful information for implicitly inferring the document's usefulness from some behavioral data. To be more specific, this research analyzes if stage of task can help in predicting document usefulness by dwell time. In other words, this study looks to see if there is an interaction effect of stage of task and the user's dwell time on predicting document usefulness. Considerations on such an interaction effect have been lacking in the related literature.

It should also be noted that although task stage is often represented as a temporal variable, it is as such a logical variable. As time passes, a task can stay in a certain stage. For instance, it can often be seen in people's life that a task is suspended for a certain period of time. In this case, the task stage can hardly be said to have any changes although the time does pass by. Therefore, although task stage is sometimes divided by time (points), for example, as beginning, middle and end stages (e.g., White, Ruthven, & Jose 2005), to view the task stage as a purely temporal factor has its limitations.

3.2.3 Topic knowledge

The third situational variable of interest to this research is topic knowledge. A user's topic knowledge usually changes along the searching process (c.f., Kulthau'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. The literature has seen much research effort spent on how topic knowledge influences either user behaviors (e.g., Hsieh-Yee, 1993; Vakkari, 2002; Wildemuth, 2004), or search performance (precision or recall; to some extent, document usefulness also belongs to this domain) (e.g., Allen, 1991; Marchionini, 1989), or both, but it has not yet seen how topic knowledge and user behavior may have to do with document usefulness. In addition to the effect of topic knowledge on user behaviors, as has been done in the literature, this study also looks at

the interaction effect of 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.

Furthermore, although previous studies have looked at the relationship between task stage and topic familiarity, as well as how this relationship and the two factors influence user's behaviors (e.g., Vakkari, 2002; Kelly & Cool, 2002), an approach has been absent to detecting how this relationship, as well as the users' behaviors, may be utilized to implicitly infer document usefulness. The current study is such an approach to extend the literature. Specifically, this study aims at exploring the role of topic knowledge, as well as comparing the effects of topic knowledge and task stage, in helping infer document usefulness from time that users spend on the documents.

3.2.4 Task product

Another facet of interest to this study as a potential source of evidence for personalization was the products that the user creates during the searching and task accomplishment process. Such products could be viewed as part of the desktop repository. As mentioned in Section 2.2.3, the desktop repository has been found in previous studies to be a good source of personalizing search results that match the specific person's information need (e.g., Chirita, Firan, & Nejdl, 2006, 2007; Teevan et al., 2005), but it calls for further research on how desktop repository can be better used for IR personalization. Specifically, utilizing the entire desktop is not as effective as using only those relevant to the current search (Chirita, Firan, & Nejdl, 2006; 2007). The product(s) generated in the process of completing a task is part of the desktop repository specifically related to the *work* task, and is very probably related to the *search* task to some extent because both the product(s) and the search task are aimed at accomplishing the same work

task. Therefore, we would like to see if using the product(s) generated for the current work task only, instead of the whole desktop, can result in improved personalization performance.

3.3 Research Questions

In general, this dissertation is aimed at exploring the effects of the user's stage toward accomplishing the task, the user's topic knowledge, as well as the product(s) generated during the course of the search, on personalizing search results. The study also takes into consideration the two types of tasks classified along the dimension of sub-task structure. Specifically, this study is conducted through examining the relationships among the above mentioned factors. Three general research questions are developed in order to reach the goals of this research, each consisting of several sub-questions, as follows.

RQ1: Does the stage of the user's task help in interpreting time as an indicator of document usefulness?

This RQ was in fact aimed at examining if there is an interaction effect between stage of task and dwell time on document usefulness. This RQ involves examining the relationships among the variables in two different types of tasks, as well as in both tasks combined. Therefore, three sub-questions were developed:

Sub-question 1a: In general, i.e., in both the parallel and the dependent tasks, does the stage of the user's task help in interpreting dwell time as an indicator of document usefulness?

Sub-question 1b: In the parallel task, does the stage of the user's task help in interpreting dwell time as an indicator of document usefulness?

Sub-question 1c: In the dependent task, does the stage of the user's task help in interpreting dwell time as an indicator of document usefulness?

RQ2. Does the user's topic knowledge help in interpreting time as an indicator of document usefulness?

Similarly to RQ1, RQ2 was in fact asked to examine if there is an interaction effect between topic knowledge and dwell time on document usefulness. RQ2 also involves examining the relationships among the variables in two types of tasks. In addition, since there could be pre-task and post-task topic knowledge, the patterns of pre-task and post-task topic knowledge should be examined. Further, the possible role that topic knowledge plays in helping interpret time as an indicator of document usefulness should be compared with the possible role that task stage plays. Therefore, RQ2 consists of the following 5 sub-questions.

Sub-question 2a: What are the patterns of the users' pre- and post-task topic knowledge?

Sub-question 2b: In general, i.e., in both the parallel and the dependent tasks, does the user's topic knowledge help in interpreting time as an indicator of document usefulness?

Sub-question 2c: In the dependent task, does the user's topic knowledge help in interpreting time as an indicator of document usefulness?

Sub-question 2d: In the parallel task, does the user's topic knowledge help in interpreting time as an indicator of document usefulness?

Sub-question 2e: How does topic knowledge compare with stage in terms of their roles in helping interpreting time as an indicator of document usefulness?

RQ3. Do the user's work task product(s) and saving and using behaviors help with query disambiguation?

This RQ examines the effect of a personalization technique that extracts terms for query expansion from combining work task product(s) and previously viewed and saved documents. As there were two types of tasks designed in the study, RQ3 looked at the performance of this

technique both on a general level ignoring task type and on a specific level considering each task type respectively. Sub-question 3a was developed for the general level; sub-questions 3b and 3c are developed for the specific level considering task types:

Sub-question 3a: Do the user's work task product(s) and saving and using behaviors help in query disambiguation in general?

Sub-question 3b: Do the user's work task product(s) and saving and using behaviors help in query disambiguation in the parallel task?

Sub-question 3c: Do the user's work task product(s) and saving and using behaviors help in query disambiguation in the dependent task?

Chapter 4. Methodology

This chapter describes the method of the study. It starts with an introduction of the data needed to answer the research questions and the determination of a controlled lab experiment as the general method to collect required data. The general study design is then presented. This is followed by detailed descriptions of the various components of the study including participants and instruments and materials used in the study. Finally the experiment procedures are described.

4.1 Controlled lab experiment for data collection

This study is aimed at exploring if there are interaction effects between/among the examined situational variables, user behaviors, and document usefulness judgments. To answer the three research questions, the study needs data that are able to determine task stages, discriminate between two different types of tasks, characterize the searcher's topic knowledge at different stages of the task, access and use products created by the searcher during the task process, capture the searcher's behaviors including time stamp, querying, document reading, using and saving, and identify the retrieved document usefulness.

Generally speaking, there are at least two ways to collect the required data: 1) a multi-staged controlled lab experiment, and 2) a longitudinal and naturalistic study. Both methods have their own advantages and disadvantages, as listed in Table 4.1. A naturalistic and longitudinal study is able to capture real users' activities with their real information needs, however, due to the uncontrolled settings, a main limitation would be the difficulty to attribute the results to certain independent variable(s), making it not easy to answer RQs 1 and 2. Moreover, in order to elicit the user's topic knowledge and evaluation of document usefulness, some type of post-task interviewing is still needed in such a natural setting. On the other hand, in a controlled lab

experiment, task stage and of task type can be rather carefully and accurately designed and easily controlled. Other variables, such as topic knowledge and usefulness judgment, can also be easily obtained through questionnaires during the lab experiment process. As for practical issues, a lab experiment is much more cost-effective and it is easier to maintain the computer and the logging tool by conducting a lab experiment than by providing desktop computers to the participants over a rather long period of time. Although data collected in a lab experiment does not completely represent the real users' real information needs, this weakness is hoped to be reduced to some extent through careful design of simulated task scenario suggested by Borlund (2000). Based on the above considerations, the study chose to use a controlled lab experiment to collect data.

Table 4.1 Comparison of lab experiment vs. naturalistic & longitudinal study

	Advantages	Disadvantages
Controlled lab experiment	• Can easily control variables such as the stage of task • Can easily elicit variables such as topic knowledge • Cost-effective and easier to manage the system and logging facility	• Not real user's real information need
Naturalistic & longitudinal study	• Real users with real information need	• Variables are not controlled • Evaluation on document usefulness & topic knowledge elicitation require post-task interviewing

4.2 General Study Design

4.2.1 Operationalization of task stage

Operationalization of the stage of a task (information seeking related) could be done in two ways with different degrees of difficulty for dividing task activities. One 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. The other case applies in some tasks with clear sub-task boundaries which can be easily divided into stages. A "complex" work task that includes sub-tasks 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, and/or efforts, and/or knowledge. Therefore, such tasks may often be conducted in sessions. In a natural setting, the sessions could either be divided by the sub-task boundaries, or by the user's time or effort limitations.

Due to the ease of identifying task boundaries, the current study employed the second approach. In this study, tasks were designed to have sub-tasks with clear boundaries, and participants were invited to the experiment lab, working on assigned tasks on different days, and finishing one sub-task at one time.

It should be noted this approach does not necessarily contradict with the ISP. In this case, for each sub-task or the overall task, the user may still have gone through the six-stage ISP. The two operationalization methods just differ in that they are from different perspectives with different criteria for separating task stages.

4.2.2 Experimental design

In general, this study had four conditions along two dimensions (see Table 4.2 Study design). One dimension was task type, being parallel or dependent (task details will be introduced later in this chapter). This dimension addressed the research questions and sub-questions regarding comparisons of these two types of tasks. The other dimension was system version, being either with query-expansion (QE) or with non-query expansion (NQE). The QE version of the system provides a list of terms for the participants to use, as desired, to formulate or reformulate their queries (details of the QE techniques are introduced later in this chapter),

while the NQE version of the system provides a regular Internet Explorer screen and does not have term suggestions. This dimension was designed to answer RQ3 regarding the performance of query expansion using task products.

In addition, to manipulate the independent variable task stage, each participant was asked to come three times for the experiment, where the three times are regarded as task stages 1, 2, and 3 respectively. Participants in the QE group started to have the QE version from stage 2.

Table 4.2 Study design

Condition	Task	System version		
		Session (stage) 1	Session (stage) 2	Session (stage) 3
1	Dependent	NQE	NQE	NQE
2	Parallel	NQE	NQE	NQE
3	Dependent	NQE	QE	QE
4	Parallel	NQE	QE	QE

4.2.3 Data needed and collection methods

With such a study design as introduced in Section 4.2, the data/variables needed and their collection methods can be made more specific, as is shown in Table 4.3. While details of the data are described later in this chapter, they are briefly introduced here. Some of them were controlled by experiment design. For example, stage of task had three values, namely, 1, 2, and 3, according to which session a sub-task was worked on. Type of task was dependent or parallel, which were assigned to different participants. Task order given to the participants was assigned, varying in 6 ways for each type of task. Meanwhile, the sub-task order that the participants actually took was noted down in the beginning of each session. Some were obtained by the logging software Morae, including time, pages viewed and saved, products generated, and the rankings of the retrieved results, etc. Some data were collected by questionnaires, such as topic knowledge, usefulness judgment.

Table 4.3 Data collection summary

	Data/variables needed	RQ (sub-RQ)	Variable measure	Collection method
1	Stage of task	1a, 1b, 1c, 2e	Stage 1, 2, and 3	Task control
2	Type of task	1b, 1c, 2c, 2d, 3b, 3c	Two types: independent & parallel	Task control
3	Topic knowledge	2a, 2b, 2c, 2d, 2e	7-point scale	Pre- & Post-session task questionnaires; Pre- & Post-session sub-task questionnaires
4	Dwell time	1a, 1b, 1c, 2b, 2c, 2d, 2e	Interval	Morae logging software
5	Usefulness judgment	All RQs & sub-RQs	7-point scale	Usefulness Evaluation Questionnaire
6	Saving and using	3a, 3b, 3c	Nominal	Morae logging software
7	Task product	3a, 3b, 3c	for QE processing	Morae logging software
8	Queries	3a, 3b, 3c	for QE processing	Morae logging software
9	Result ranking	3a, 3b, 3c	Ordinal	Morae logging software
10	Difficulty level	All RQs & sub-RQs	7-point scale	Pre- & Post-session task questionnaires; Pre- & Post-session sub-task questionnaires
11	Assigned task order	All RQs & sub-RQs	1 - 6	Task control
12	Actual task order	All RQs & sub-RQs	Varying	Note down
13	Age	All RQs & sub-RQs	Ratio or Ordinal	Background questionnaire
14	Gender	All RQs & sub-RQs	Categorical (male, female)	Background questionnaire

(Note: IV: independent variable; DV: dependent variable)

4.3 Experiment components

This part describes the various experimental components of this study, including participants and instruments and materials used in the study. A note is that these instruments and materials were what were used in the formal experiment, after various revisions based on a pilot study conducted beforehand. The pilot study had 3 participants, one of whom worked with the NQE version and two of them worked with the QE version of the system. The 1st session of the

pilot study started on January 7, 2009, and the last session of it ended on February 9, 2009. Since there were 3 sessions in which each participants came to the experiment, the formal experiment started before the last session of the pilot study. The 1st session of the formal experiment started on January 27, 2009, and the last session of it ended on February 27, 2009.

4.3.1 Tasks

This study adopted the definitions of work task and search task by Li (2008). A work task is an activity that one perform to fulfill one's responsibility for the work, and a search task is an activity to search for information through interaction with information systems. This study used work tasks that were not equal to search tasks and had at least one associated search task.

This study used journalists' assignments as tasks. The major reason was that they could be relatively easily set as realistic tasks in different search domains. As mentioned before, 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 being a parallel task and the other a dependent task. They both had three sub-tasks, each of which was worked on by the participant during a 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 the two significant facets that have been demonstrated (e.g., Li, 2008) to influence user's search behaviors: product and objective task complexity. First, the task product was set as intellectual for all three sub-tasks, specifically, each sub-task asked the participants to submit a report, which by its nature embedded new ideas or findings (Li, 2008). Second, the objective complexity of the two tasks was roughly the same, both being low complex, using Li's (2008) definition of task complexity. This meant that each sub-task 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 (also see Appendices E & F).

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 3 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.

In the experiment, the users were also asked to include the citations (e.g., URLs) in their reports. This was not included in the task description though.

4.3.2 Task orders

The research assumption underlying the task description is that the sub-task order in the parallel task is not fixed while that in the dependent task is at least to some extent fixed. To maintain consistency, the experiment did not control the order of sub-tasks, but rather chose to let the participants determine the sub-task orders that they wanted to follow. Sub-task orders that appeared in task description were rotated. The rotation followed a Latin Square design, as follows:

The parallel task:

1. Honda, Toyota, Nissan
2. Toyota, Nissan, Honda
3. Nissan, Honda, Toyota
4. Honda, Nissan, Toyota

5. Nissan, Toyota, Honda
6. Toyota, Honda, Nissan

The dependent task:

1. Collect information and make a list – Select 3 models – Compare 3 models
2. Select 3 models – Compare 3 models – Collect information and make a list
3. Compare 3 models – Collect information and make a list – Select 3 models
4. Collect information and make a list – Compare 3 models – Select 3 models
5. Compare 3 models – Select 3 models – Collect information and make a list
6. Select 3 models – Collect information and make a list – Compare 3 models

One round of the order involves 6 participants, and the 24 participants constituted 4 round of such rotation. Table 4.4 shows the basic assignment of the tasks to the participants from the perspective of task order balancing.

Table 4.4 Participants' condition assignment

Participants	Task	Sub-task order	System version		
			Session (stage) 1	Session (stage) 2	Session (stage) 3
s01	Dependent	1	NQE	NQE	NQE
s02	Dependent	2	NQE	NQE	NQE
s03	Dependent	3	NQE	NQE	NQE
s04	Dependent	4	NQE	NQE	NQE
s05	Dependent	5	NQE	NQE	NQE
s06	Dependent	6	NQE	NQE	NQE
s07	Parallel	1	NQE	NQE	NQE
s08	Parallel	2	NQE	NQE	NQE
s09	Parallel	3	NQE	NQE	NQE
s10	Parallel	4	NQE	NQE	NQE
s11	Parallel	5	NQE	NQE	NQE
s12	Parallel	6	NQE	NQE	NQE
s13	Dependent	1	NQE	QE	QE
s14	Dependent	2	NQE	QE	QE
s15	Dependent	3	NQE	QE	QE
s16	Dependent	4	NQE	QE	QE
s17	Dependent	5	NQE	QE	QE
s18	Dependent	6	NQE	QE	QE

s19	Parallel	1	NQE	QE	QE
s20	Parallel	2	NQE	QE	QE
s21	Parallel	3	NQE	QE	QE
s22	Parallel	4	NQE	QE	QE
s23	Parallel	5	NQE	QE	QE
s24	Parallel	6	NQE	QE	QE

4.3.3 Participants

As previously discussed, this study used journalists' assignments as work tasks, and accordingly the participants were also recruited from those who have certain knowledge and skills to deal with such kinds of assignments. This study invited participants from Journalism/Media studies undergraduate student community in the School of Communication and Information (SC&I) at Rutgers University. There were three main reasons to use this group of students as participants. First considered when choosing students as participants rather than real journalists was the limitation of the resources: budget, and chance to find available journalists, etc. The second reason was the convenience of recruitment. These students were available in the same school as the author. Third, the educational background and level of this group of students were roughly the same, which should have helped, at least to some degree, avoid the potential individual effects on users' task performance.

The recruitment was conducted through the following means: sending recruitment emails to student listservs and posting recruitment ads on post-boards in the SCILS building. The recruitment notice is attached in Appendix A.

Each participant was invited to come and work on the assigned task three times within a two-week interval, at their convenient time slots. This time span was determined based on the consideration of coordinating participants' schedule, minimizing their work load, and ensuring that the experiment could be finished in a timely manner. Participants received remuneration for their assistance in the study. Upon completing the whole assignment, each of them received $30.

In order to encourage them to work on the assignment in a serious manner, the study employed an incentive system. They were told in advance that the 6 participants who submitted the most detailed reports on the assignment would each receive an additional $20.

4.3.4 Study location and computer equipment

The study was conducted in an interaction lab in the SC&I building at Rutgers University. The participants were provided a desktop computer to work on the tasks, with high-speed Internet connection. The participants were allowed to freely choose whatever online systems that they wanted to search in and also to freely choose online sources that they wanted to use in order to accomplish their tasks. For technical reason to make a side-panel work (see more details in Section 4.3.5), the browser was restricted to Internet Explorer (IE) version 6.0.

4.3.5 Search systems

For those participants assigned to the NQE condition, the regular IE was used throughout all three sessions. For those assigned in the QE condition, the regular IE was used in their first sessions, while a different interface, the QE version, was provided for the second and the third sessions to support query formulation. This different interface consisted of two parts: the right side panel was the regular IE browser, and the user could search the open Web using whatever search engine they liked, but they were not allowed to open more than one browser window, or more than one browser tab (IE 6.0 was used because it does not support multi-tab) in order to make the experiment system and the logging software Morae run more smoothly; the left-side panel displayed a list of terms extracted from their previous session(s) for the users to select to include in their queries. A screenshot of this interface is shown in Figure 4.1, and the script to support the left-side panel can be found in Appendix O. The script functions correctly only in IE.

In terms of query expansion, there could be at least three approaches: 1) the opaque way,

in which the queries are expanded in a hidden way, and the users do not know what terms are added, 2) the transparent way, in which the users are shown the added terms to their queries after the search, but the users are not able to use those terms for expansion, and 3) the penetrable way, in which the users are provided the lists of expansion terms and they can select which terms they want to use to expand their queries (c.f., Koenemann & Belkin, 1996). The current study used the penetrable approach. This decision was based on the following two considerations. First, the penetrable way in Koenemann & Belkin (1996) demonstrated the best improvement on performance among the three ways. Second, we wanted to balance the variety of searching systems and the restriction of technical support. On the one hand, we wanted to enable users to search in all types of online systems; on the other hand, it was not easy to make this completely hidden because we used the dynamic WWW, not a self-built system with a consistent collection. This may to some extent cause a sort of bias in that users could perhaps think using the words to expand their query will lead to better system performance. Using an opaque way to expand query will be conducted in a future study to compare its performance with the penetrable way.

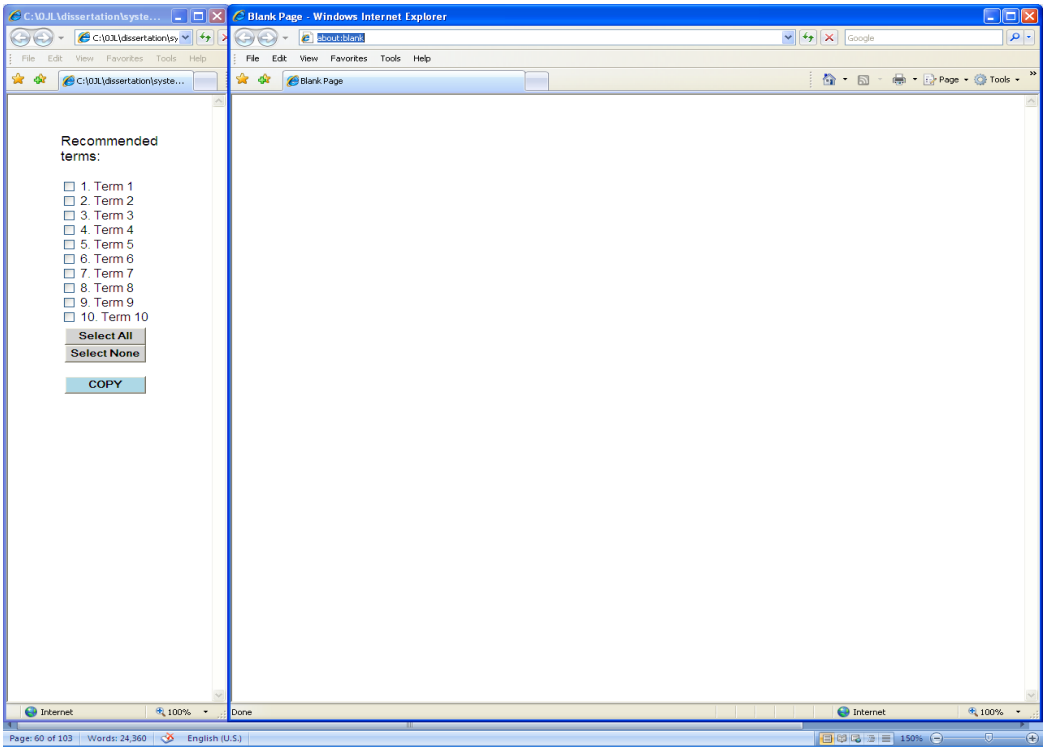


Figure 4.1 Search interface for the 2nd and 3rd sessions in QE condition

In the QE condition interface, the terms shown in the left-side list were extracted by two approaches from two different sources. One approach treated the documents that were saved and/or used for the assignment in previous session(s) as relevant to the query, and extracted significant terms from these documents. The other approach was to extract significant terms from the user-generated product(s), including the assignment report(s), and any notes that users may have taken. The assumption here was that the generated product was relevant to the query since it was aimed at solving the task that involved the user's information need. Significant terms were selected by the value of term frequency, i.e., how many times a term appeared in the collection of saved and/or used documents, as well as the user-generated product. The term frequency method for extracting query expansion terms was also used in Chirita, Firan, & Nejdl (2006), and the researchers obtained improved performance by expanding queries using terms with top frequencies in the user's desktop repository collection.

Term frequency was obtained using the index function in the Indri[4] toolkit. The terms identified by the two approaches were merged into a single list ranked by term frequency from the most to the least, and the top *n* terms were selected to be shown in the list on the search interface's left-side panel. The number *n* was determined by the following rules. First, those terms with a frequency of 3 and greater were selected. We hypothesized that terms with a frequency of less than 3 in both sources would not be good candidates for query expansion. Secondly, the maximum number of terms we selected would be 20. This maximum number of 20 was in consistent with previous studies. In Koenemann & Belkin's (1996) study, the average number of terms that the user chose from the suggested term list to expand their query was about 17. In Belkin et al. (2005) (TREC HARD track project), there were two sources for extracting query expansion terms, each contributing 10 terms or phrases, and the total number of terms in a clarification form were from 10 to 20 depending on if there were duplicate terms or phrases extracted from the two sources. In addition, this maximum number of terms to display to the users, i.e., 20, was also based on the consideration of minimizing user effort in using these terms. All terms should have been displayed in one single page without requiring scroll-down. These rules of determining the number *n* were tested in pilot experiments and *n* fell roughly within the scope of 14-20.

The way to select terms from the list and add the selected terms to the query was easy. Users could select any terms that they like by clicking the checkbox, and deselect the terms by clicking a checked checkbox. They could click the "Select All" button if they wanted to select all terms, and by clicking the "Unselect All" button, all terms would be unselected. After selecting the terms, the user could click the "COPY" button, and the selected terms would be copied to the clipboard. To expand the query or simply use the selected terms as query keywords, the user

[4] http://www.lemurproject.org/indri/

could simply paste the copied terms in the clipboard in the query box in the main panel.

4.3.6 Logging software

The Morae logging software was installed on the experimental machine in order to log participants' search activities. The following information was logged:

1 Time

Morae recorded the time stamp of each single event. According to such time stamps, the starting and ending time of each search episode were determined and the duration of each search episode could be computed. Therefore, the time that users spent on each page could be obtained from the logged time. This was used to address RQs 1 and 2.

2 Pages that the users viewed

Each page that the participant had ever viewed was recorded. These pages were replayed at the end of each session for the participant to judge their usefulness to the task. The usefulness values were necessary to address all three RQs.

3 Pages that the users saved

The participants were asked to work on the assignments as naturally as they work on any assignments in their everyday lives, including saving pages (or other information objects) as they normally do when doing search. Possible ways of saving consist of bookmarking, saving web pages, and so on. Saving retrieved pages was one type of user behaviors that was frequently seen in seeking information and accomplishing one's tasks. As discussed earlier, the saved pages were treated as part of the user's desktop repository and were for extracting important terms for query expansion. The saved web pages were particularly helpful for solving RQ3. However, it turned out that users rarely saved documents, as is described in Chapter 5.

4 Ranks of the retrieved results

Morae was able to record the results page, from where the ranks of the retrieved results can be obtained. The ranks, especially those evaluated by the participants, were helpful in computing the performance of the QE and NQE versions in order to solve RQ3.

4.3.7 Consent form

This study was approved by Rutgers University Institutional Review Board (IRB) for the protection of human subjects, and it started after the approval was granted. Consent form was a must-have material to be included in the IRB application, and each participant was supposed to be informed of the consent form before the experiment. In particular, after each participant arrived in the first session, before starting the real experiment, he/she was asked to read and sign an informed consent form.

The consent form described the following issues: a) general purpose of the study; b) times of sessions and time span of the study; c) basic procedure of the study; d) monitoring and recording of all computer activities; d) compensation for participation; and e) how the collected data would be analyzed, stored and used in the future. A copy of the consent form can be found in Appendix B.

4.3.8 Other data collection supporting materials

Other than the items that Morae recorded, data were collected by some other supporting materials. These include: Background questionnaire, Pre- and Post-session task questionnaires, Pre- and Post-session sub-task questionnaires, Usefulness Evaluation Questionnaire, and finally the Exit Interview. All these questionnaires/interview were provided online.

4.3.8.1 Background questionnaire

A Background questionnaire was prepared to gather participants' demographic information and background, including information about age, gender, and academic levels. The

Background questionnaire also asked participants to indicate their previous computer and searching experience. A copy of the background questionnaire can be found in Appendix G. The information obtained from this questionnaire was used to characterize the participants.

4.3.8.2 Pre-session Task Questionnaire

A Pre-session Task Questionnaire was administered right after the participants were shown the general task. This questionnaire asked the participants to indicate their familiarity with the work task in general, how much experience they had with this kind of assignment, as well as the how difficult they expected the task to be (pre-task difficulty). A copy of the Pre-session task questionnaire can be found in Appendix H. The information gathered by the Pre-session task questionnaire was used to address RQ2 regarding the effect of topic knowledge on usefulness judgment.

4.3.8.3 Post-session Task Questionnaire

A Post-session task questionnaire was administered after the participant finished each session. This questionnaire asked the participants to indicate their familiarity with the work task at that specific point after they completed each session. A copy of the Post-session task questionnaire can be found in Appendix L. The information gathered by the Post-session task questionnaire was used to address RQ2 regarding the effect of topic knowledge and the interaction effect of topic knowledge and search stage on usefulness judgment.

4.3.8.4 Pre-session sub-task questionnaire

A Pre-session sub-task questionnaire was administered before each participant started the sub-task in each session. This questionnaire asked the participants to indicate their familiarity with the sub-task at that point, how much experience they had with this kind of sub-task, as well as the sub-task difficulty that they expect. A copy of the Pre-session sub-task questionnaire can

be found in Appendix I. The information gathered by the Pre-session sub-task questionnaire was used to address RQ2 regarding the effect of topic knowledge and the interaction effect of topic knowledge and search stage on usefulness judgment.

4.3.8.5 Post-session sub-task questionnaire

A Post-session sub-task questionnaire was administered after each participant finished the sub-task in each session. This questionnaire asked the participants to indicate their familiarity with the sub-task at that point, as well as the sub-task difficulty that they felt from their experience. A copy of the Post-session sub-task questionnaire can be found in Appendix K. The information gathered by the Post-session sub-task questionnaire was used to address RQ2 regarding the effect of topic knowledge and the interaction effect of topic knowledge and search stage on usefulness judgment.

4.3.8.6 Usefulness Evaluation Questionnaire

The purpose of this questionnaire was to elicit the participants' usefulness judgments of their viewed and/or saved information objects during this session. This questionnaire was administered to the participants at each session right after they finished the sub-task and before the post-session sub-task questionnaire. This was shown to the participants simultaneously with a replay of the viewed and/or saved information objects during this session. This questionnaire asked the users to give their usefulness judgments and confidence levels based on a 7-point Likert scale. It also asked if they had seen the page before, and if so, how familiar they were with the page, based on a 7-point scale. A copy of the Usefulness Evaluation Questionnaire is available in Appendix J. The information gathered by this questionnaire was used to address RQs 1 & 2.

The replay used the Morae Manager software. Each page users had viewed was manually

marked by the experimenter in Morae observer at the same time when the page was opened by the users in their search, and the marked data was recorded in the video and was used to locate the pages in the recorded file to show to the participants. When replaying the recorded file, each marked page was clicked and evaluated one by one. To link the evaluation with the pages correctly, the usefulness rating was manually noted down by the experimenter in the comments area in each marked page, once the participant gave their rating scores.

4.3.8.7 Exit Interview

After each participant had completed the whole work task in session 3, an exit interview involving several open questions was conducted. This interview focused on the participants' general perceptions regarding how they felt about their experience of this study, their overall knowledge acquisition on the task, their perceptions and interpretations about the scales and procedures used during this study, and any other comments in general that they may have had. A copy of the interview questions can be found in Appendix M.

This interview offered an opportunity for the participants to express their opinions that may not have been covered in the data collection materials but which might be useful to the analysis in the other data. The interviews were recorded.

4.4 Experiment procedure

This section describes the general procedure of this experiment. Figure 4.2 illustrates this procedure. Totally there were three sessions for each participant. The same steps in all three sessions as well as those steps unique in each session are marked differently.

Upon arrival for the first session, each participant was first asked to read and sign a paper version of the informed consent form. During this process, the experimenter also briefly highlighted and explained what the participant would be asked to do in the study. After this, the

participant moved to the computer. He/she was first shown the general task instructions (see Appendix C). This was followed by a background questionnaire which elicited the participant's demographic information, background, and previous search experience. The participant was then presented the work task assigned to him/her, followed by a pre-session task questionnaire. The participant was then asked to choose one sub-task to work on in the current session. Following this was a pre-session sub-task questionnaire. After that came the sub-task instructions (Appendix D), showing what the participants could and could not do. Then the participant started to work with the sub-task, searching and writing the required report. The whole process was limited to 40 minutes. During the session, the participant was allowed to freely view, save, and use any information objects that they retrieved to accomplish the sub-task. After having completed the sub-task, or reached the time limitation of 40 minutes, the participant was asked to save the product(s) that have (has) been generated for the session. Then the viewed pages were displayed, together with the Usefulness Evaluation Questionnaire. After the evaluation, the participant completed a Post-session sub-task questionnaire to indicate his/her post-session sub-task topic knowledge and difficulty of the sub-task, followed by a post-session task questionnaire to indicate his/her post-session task topic knowledge.

Except for the consent form and the background questionnaire, the above described process was repeated in sessions 2 and 3 for other sub-tasks. For those who were assigned to use the QE version of system in sessions 2 and 3, a step showing instructions and demonstrating use of the QE function was administered before the participants started to search and write reports.

In the final session, after the Post-session task questionnaire, the participant was given the exit interview that asked the participant about his/her overall knowledge acquisition on the task, comments on the instruments and scales used in the experiment, and others, etc.

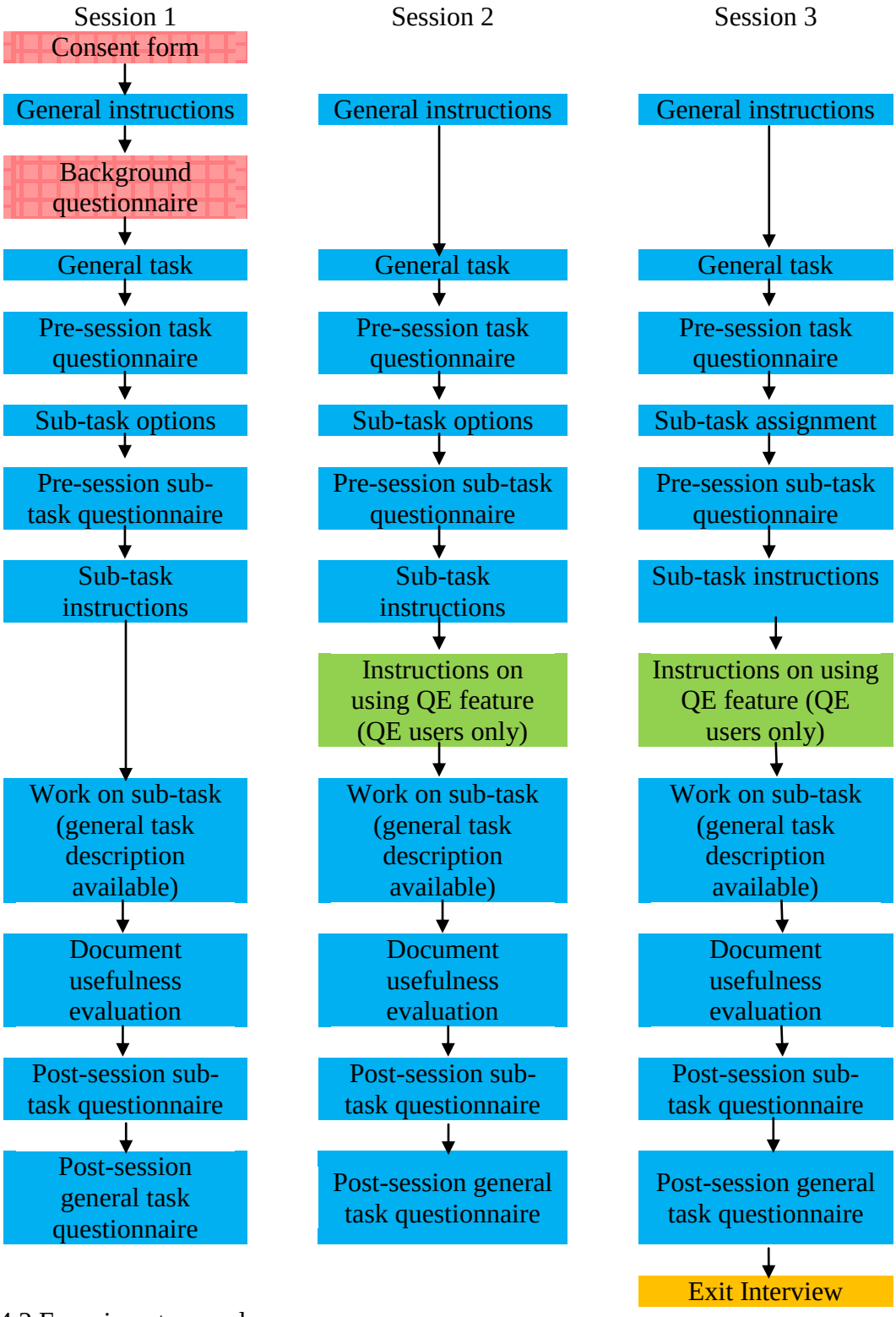


Figure 4.2 Experiment procedure

Chapter 5. Results

This chapter reports the results of the study. The characteristics of the participants are first reported, including their demographic information as well as previous computer and search skills. The characteristics of the dataset are then described. Various types of time that users spent on retrieved document are then defined, three of which are important for the current study. The results for the three research questions are then presented. For research questions 1 and 2, results are reported for both tasks combined and for the parallel and the dependent task individually, which are different sub-questions. In each sub-question, results are reported when each of three different types of time is examined. Research question 3 is not able to be addressed and the reason for this is explained.

5.1 Characteristics of the participants

Twenty-three Journalism/Media studies and one communication undergraduate students participated in the study. They all finished all 3 sessions of the experiments. Of the 24 participants, 21 were female and 3 were male. Ten of them were seniors, 6 of them were juniors, and 8 of them were sophomores. Their ages varied between 18 and 23, with an average of 20.4 (standard deviation, simplified as SD in the following, was 1.3) years.

Participants' computer and search experience and expertise were obtained through the Entry Questionnaire. They were asked to indicate their levels of computer expertise and of searching expertise on a seven point scale, where 1 = novice and 7 = expert. Figure 5.1 lists the results of all 24 participants' respondents. On average, participants self-assessed their levels of computer expertise as 4.6 (SD=1) on a 7 point scale. One rated himself as level 2, one as level 3, two as level 6, one as level 7, and the rest of the 19 participants rated themselves as level 4 or 5. They rated their levels of searching expertise as 5.4 (SD=0.9) on the 7-point scale. One rated

himself as level 3, one as level 7, and the other 22 participants rated themselves as level 4, 5, or 6. Participants were also asked how many years they had been doing online searching. The results are displayed in Figure 5.2. On average, they had 8.4 years (SD=2.9) of online searching experience. All had at least 4 years of searching experience, and the maximum was 15 years.

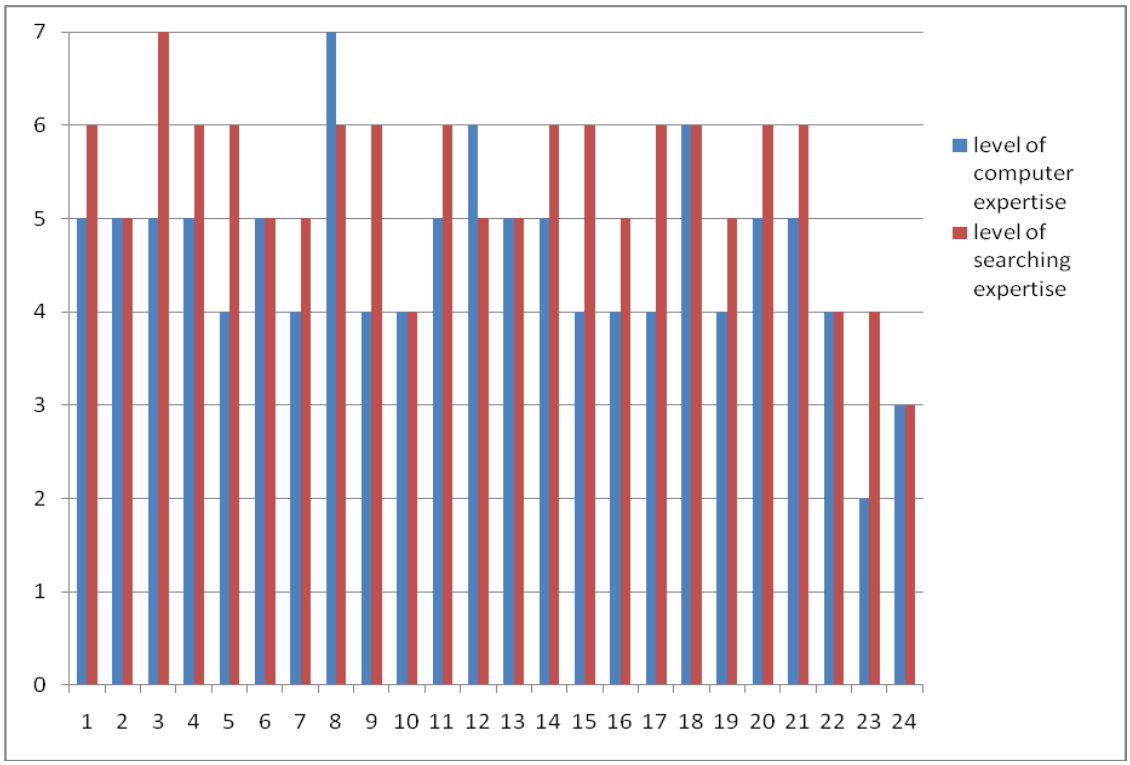


Figure 5.1 All 24 participants' computer and searching expertise (x-axis: participant number; y-axis: level)

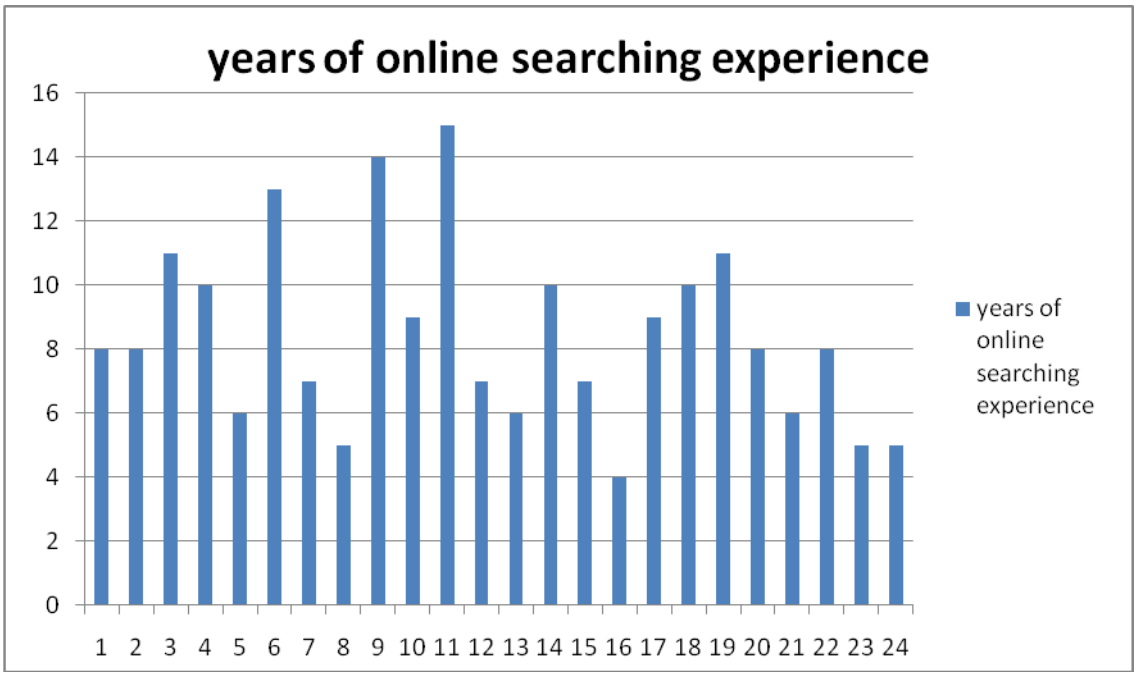


Figure 5.2 All 24 participants' online searching experience (in years) (x-axis: participant number; y-axis: year)

5.2 Characteristics of the data

Recall that in the experiment, participants were asked to freely save and use any sources that they can find to write their reports. Data of users' interaction with the computer was logged by Morae software. It was observed that users rarely saved documents. In all 72 sessions (24 participants * 3 sessions), saving behavior happened only with two users at 3 sessions, and a total of only 7 Web pages were saved (Table 5.1). Instead, it was observed that participants used the Web pages a lot, and writing their reports in parallel with searching on the Web and reading the located information was frequently seen. In addition, many participants created a different MS WORD document to make notes than the one that they submitted for their reports.

In each session of the experiment, after the participant submitted the report, those documents (mainly Web pages) that were ever opened by the participant were assessed for their usefulness in accomplishing the whole task, based on a 7-point scale. The documents were

displayed in the same order that the participant opened them. In general, for the pages that were opened multiple times in a session, most of the times, users judged them consistently for each session, i.e., there was only one usefulness score associated with each unique Web page. In the rare cases when users assigned different usefulness scores to the same web page in a session, the last assessment score was used.

Table 5.1 Summary of saved documents in all sessions

	User id	Session number	# of saved pages
1	S09	1	2
2	S09	2	3
3	S12	1	2

The experiment was aimed at asking the users to evaluate each page that they had viewed, however, some pages missed the evaluation score due to the following reasons. The task topics were about cars, and a large number of car websites were relatively dynamic and lively, including in different pages of their websites many flashes, videos, images, tabs, and frames. Sometimes the same URL could have different displays. At other times, the same display could have different URLs. Sometimes, the website had pop-up windows or text boxes to display some texts about the features of a car. All these phenomena affected the experimenter in instantly judging in the experiment whether different displays of pages should be treated as the same or different page, as well as marking pages for evaluation purposes. Due to these reasons, totally, for the 72 sessions, there were 179 unique URLs and pages (including the different tabbed/framed pages in a single URL) which missed usefulness scores. There were 993 unique URLs which had usefulness scores. The missing ratio was 15.2% (179/(179+993)). Since the pages were missing in a random way, it was supposed that this does not affect the interpretation of the data using the collected 993 unique Web pages.

In general, for all 72 sessions, 993 Web pages were collected with usefulness scores. The number of pages in each session was displayed in Figure 5.3. The mean pages per session in general was 14 (SD=7.3). The mean pages per session for each stage are shown by Table 5.2.

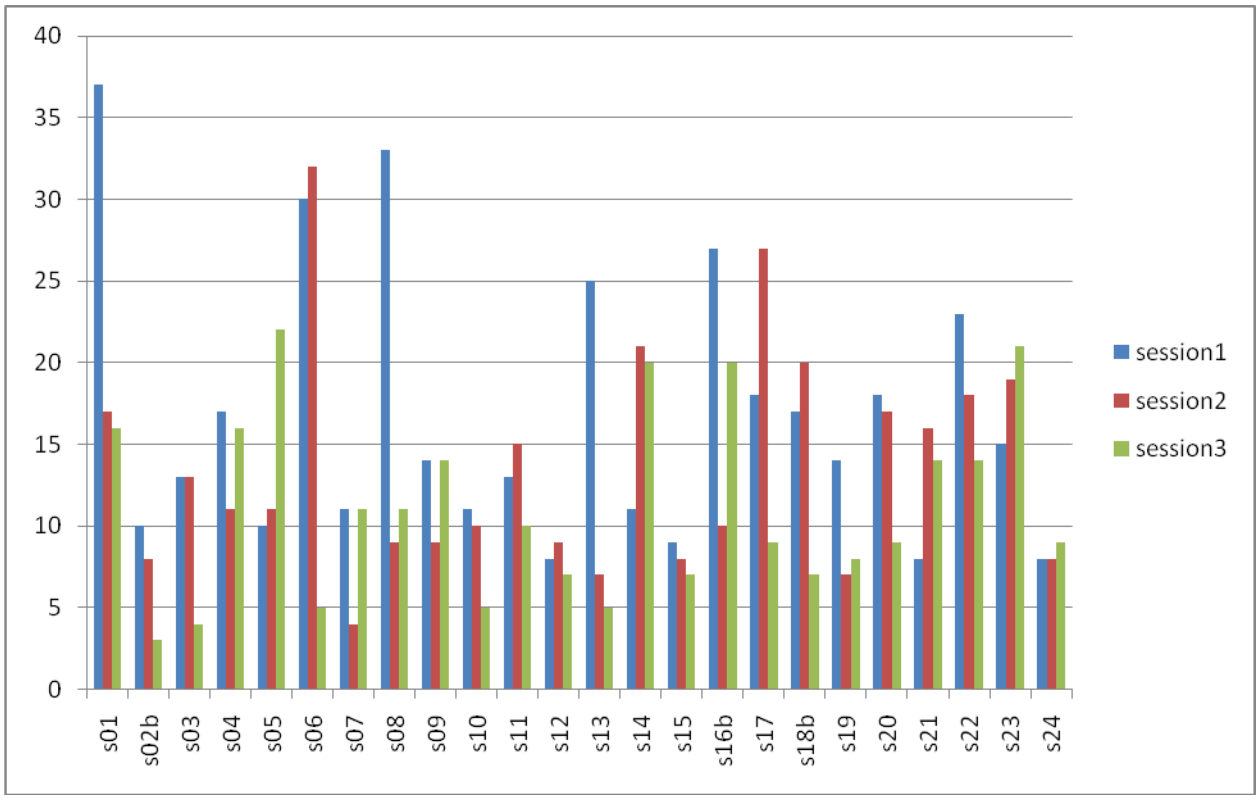


Figure 5.3 Summary of the number of viewed pages (with usefulness scores) for all participants in 3 sessions (x-axis: participant number; y-axis: number of viewed pages)

Table 5.2 Summary of mean pages per session in each stage

	Stage 1	Stage 2	Stage 3
Mean	17	14	11
SD	8.3	6.8	5.7

5.3 Various types of time

In the related literature, the time variable in interactive IR studies can refer to two general things according to the variable level: task completion time (task level) and dwell time (document level). Task completion time is the total time that a user spends completing the IR task. Dwell time (sometimes named display time or reading time) measures the time that a user

spends on a retrieved document. In the current study, only the document level time is relevant and considered.

Names, definitions, and measures of document level time in previous research have variations. Kelly (2004) used "display time" to measure "the length of time that a document was displayed in the subject's active Web browser window" (p. 91). For identical pages viewed at different times, Kelly (2004) summed all display times of this page to arrive at a total elapsed time. Gwizdka (2008) used a different variable, "time-per-click", to measure time which was also, roughly speaking, on a document level. Time-per-click is the average duration of all events that are separated by mouse click activities, including Web page display, query input, and so on. Despite these variations in the use of time in previous studies, in general, document level time in any single study had only one definition and did not have detailed categorization.

In the current study, it turned out that becasue users were asked to accomplish a work task, i.e., to generate some task products based on their information searching, users often wrote in MS WORD in parallel with searching for information on the Web. Retrieved documents could be open for a long time but not always active, especially when the users were focusing on writing. There is a need to differentiate several types of time on the document level, which could be named as dwell time, display time, and decision time (Figure 5.4). The definitions of these times are described in detail below.

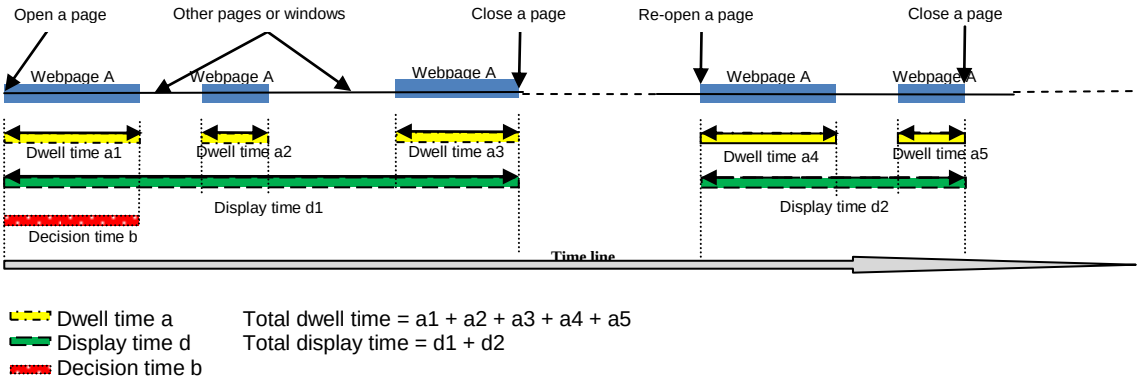


Figure 5.4 Different types of time

5.3.1 Dwell time

Dwell time (denoted as *a* in Figure 5.4) is defined as the time duration from each point when the user starts viewing a document (usually when a document is opened) to 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 other applications[5]). Each dwell time is the time that a user dwells on the document[6], or in other words, that a document is active for the user to read.

5.3.2 Total dwell time

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

5.3.3 Display time

Display time (denoted as *d* in Figure 5.4) refers to the total duration of a document between when it is opened to when it is closed. This is also the total time that the document remains open, no matter if it is active, i.e., if the user views it or not, after it is opened. It is possible that a document was opened for multiple times in an experiment session, so one document could have multiple display times at different points. If the user dwells with a webpage in the entire duration of its opening time, i.e., the user does not go to other applications before he/she closes this webpage, the display time of this page is equal to the dwell time.

5.3.4 Total display time

Total display time is the sum of all display times for a document that is revisited during a session. For those webpages opened multiple times, if each time, the user dwells with a webpage

[5] The current study asked the users to keep only one Web browser (IE window) open for the purpose of logging, therefore, it is not possible that two web pages or other retrieved documents were open at the same time.
[6] In this study, when a window was active, it is assumed the user was active working with applications in this window. For example, when the IE window was active, we think the user was reading, or at least, dwelling on the Web pages or documents. When the MS WORD window was active, we assume that the user was working on the report writing.

in the entire duration of its opening time, i.e., the user does not go to other applications before he/she closes this webpage, then the total display time of this page is equal to the total dwell time.

5.3.5 Decision time

Another type of time we identify and define is the first dwell time, denoted as b in Figure 5.4. This time is called decision time 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, going to an MS WORD document and starting writing most likely meant that the Web page that the user has just viewed was useful; leaving a Web page and going back to search result page (to refine queries or open another search result) perhaps meant that the page just viewed was not useful. In the current study, the end point of this decision time has two markers: going to another document (usually another Webpage), or going to another application (e.g., MS WORD, etc.).

Among all the above mentioned types of time, total display time, total dwell time, and decision time best represent the features (for example, usefulness) of a certain document across the whole session (see more detailed explanation later in Sections 5.4.1.1, and 5.4.1.2), and they are used for analysis.

5.4 Results of RQ1

All statistical analyses of all three RQs were conducted in SPSS 17.0. General Linear Model (GLM) was used for statistical examination for RQs 1 and 2 because it can detect the interaction effects between/among variables in these two RQs.

RQ1: Does the stage of the user's task help in interpreting time as an indicator of document usefulness?

This RQ looks at the relationships between task stage, document usefulness, and time. Since there are two types of tasks, these relationships should be examined with different types of tasks, as well. As mentioned in the above section, there were three types of time. They are examined respectively with respect to their relationships with other factors. In addition, this RQ had 3 sub-questions according to the different task types that the above mentioned relation was examined in.

5.4.1 Sub-question 1a

Sub-question 1a: In general, i.e., in both the parallel and the dependent tasks combined, does the stage of the user's task help in interpreting time as an indicator of document usefulness?

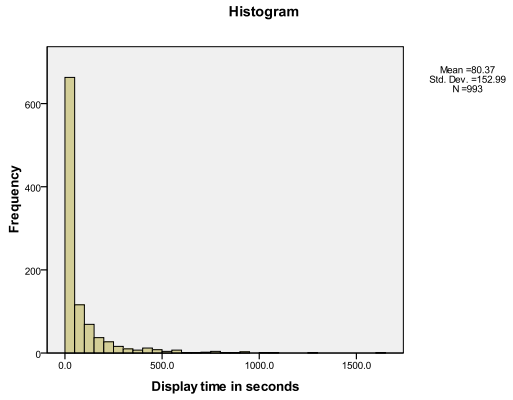
5.4.1.1 Total display time

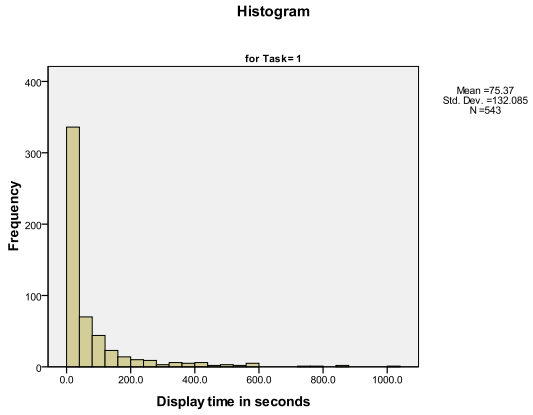
As mentioned before, in this study, some retrieved documents were opened multiple times in a session and there is a display time for each time the same retrieved document was open. In the evaluation process in each experiment session, most of the times, the same document get the same evaluation score, and in rare times, they were given different scores, but the last one was used. So in the data, each document was given a unique usefulness score no matter how many times it was opened. It is therefore reasonable to consider the total display time of each document, rather than the individual display time for each time it was open, in a session, for its relationship with this document's usefulness.

5.4.1.1.1 Transformation of time

It is often reported in the literature (e.g., Kelly, 2004) that dwell or reading time is not normally distributed, so it is necessary to examine the distribution of total display time before running any statistical data analysis. Figure 5.5, Figure 5.6, and Figure 5.7 show the distribution of total display time in both tasks and in individual tasks, respectively. As can be seen, the time

distributions were far away from normal. This was consistent with what was found in the literature (e.g., Kelly, 2004). In order to adjust these distributions and to improve the interpretability of results on relationships between factors, a logarithm transformation was performed using the log base of 10[7,8]. Figures Figure 5.8, Figure 5.9, and Figure 5.10 show the distributions of total display time in both tasks combined and in each individual task after data transformation, which were much more bell-shaped, even though some were not perfectly normal).





[7] With regard to transformation using logarithm, both the natural log and log base 10 would be fine, and the difference is that they used different scale. The current study chose the log base of 10 based on the consideration that it may be a bit easier to estimate the original time (for those who look at the data in analysis) in seconds (which conveys the real time information) when seeing the transformed data.

[8] In this study, all other times, including total dwell time and decision time had the same characteristic; i.e., the distributions of the original data were far away from normal, but the transformed data (using a log base of 10) had more bell-shaped distributions. So all the time examined in RQs 1 and 2 in this dissertation used log(10) transformation.

Figure 5.5 The distribution of total display time in both tasks

Figure 5.6 The distribution of total display time in the dependent task

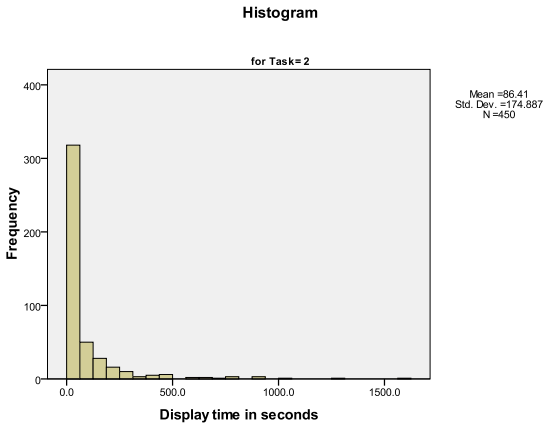
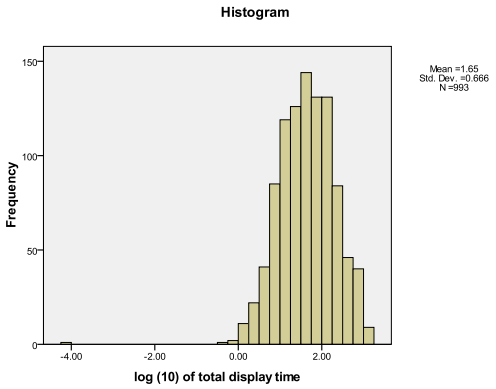


Figure 5.7 The distribution of total display time in the parallel task



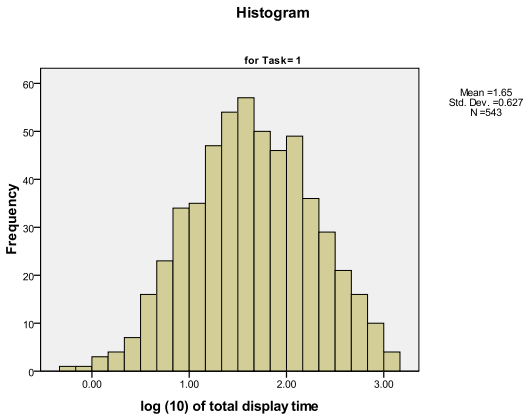


Figure 5.8 The distribution of log(10) total display time in both tasks

Figure 5.9 The distribution of log(10) total display time in the dependent task

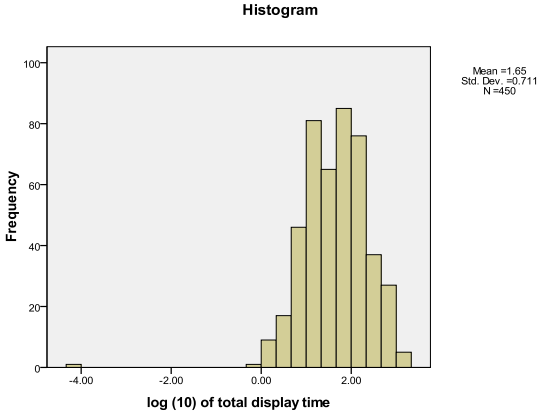


Figure 5.10 The distribution of log(10) of total display time in the parallel task

5.4.1.1.2 Grouping usefulness scores

Usefulness scores for each viewed document in this study were explicitly extracted. They were obtained at the end of each session by asking the users to rate, based on a 7-point Likert scale, how useful the document was for accomplishing the task. The 7-point Likert scale was appropriate for collecting user assessments (Tang, Shaw, & Vevea, 1999), but could be too fine-grained for a system to differentiate. In their study, White & Kelly (2006) collapsed the original 7-point scores elicited from user ratings into binary scores to represent document usefulness: not relevant and relevant. Sometimes binary categorization of relevance is not enough, and in some TREC practices, 3-point relevance assessment scale was used: not relevant, somewhat relevant, and very relevant. In the current study, document usefulness was collapsed into three groups: not useful, somewhat useful, and very useful.

To this end, the distribution of usefulness was first examined for each RQ and sub-RQ. For example, for the total dwell time in RQ1a, Figure 5.11 shows the original usefulness distribution:

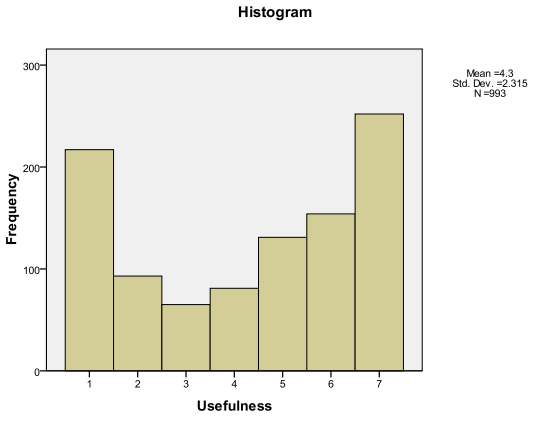


Figure 5.11 The distribution of original usefulness data in both tasks

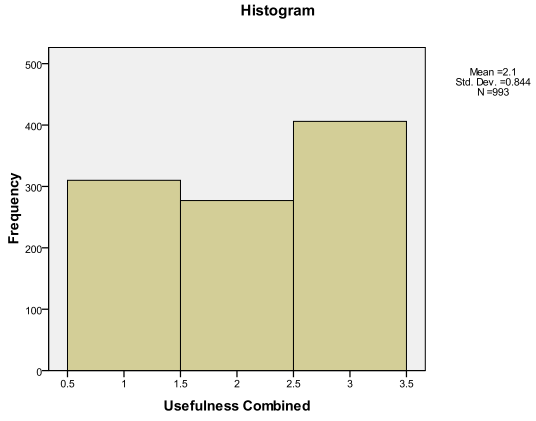


Figure 5.12 The distribution of combined usefulness data in both tasks

From this distribution, it is reasonable to combine scores 1- 2 into a low-useful group, 3-5 into a mid-useful group, and 6-7 into a high-useful group. Figure 5.12 shows the distribution after grouping, where the 3 groups were quite balanced. In the following part of this subsection (Section 5.4.1.1), unless specified, usefulness is denoted as the combined usefulness.

5.4.1.1.3 Data analysis for total display time in RQ1a

Table 5.3 and Figure 5.13 show the relationship between usefulness, stage, and log(10) of total display time.

Table 5.3 GLM Univariate results of the effects of usefulness, task stage, and their interaction on mean log(10) of total display time in both tasks

Effect source (factor or interaction)	F	p
Stage	4.150	.016
Usefulness	123.779	.000
Stage*Usefulness	2.658	.032

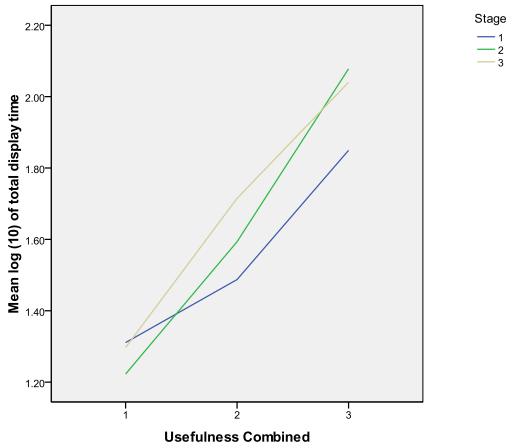


Figure 5.13 Relations between usefulness, task stage, and log(10) of total display time in both tasks combined

Results show that there was a significant effect of usefulness, $F(2, 990)=123.779$, $p<.001$, 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 effect of stage, $F(2, 990) = 4.15$, $p<.05$. In addition, there is a significant interaction effect between stage and usefulness, $F(4, 988)=2.658$, $p<.05$, meaning that the patterns of the

relation between usefulness and total display time varied across stages. Specifically, in stage 2, non-useful documents' display time was a bit lower than in stage 1 and 3, and the very useful documents' display time was almost the same as that in stage 3, both higher than the display time in stage 1.

5.4.1.2 Total dwell time

Table 5.4 and Figure 5.14 report results when total dwell time was considered. Just as what was done with the total display time in the previous subsection, it is also reasonable to consider the total dwell time of each document, rather than the individual dwell time for each time it was open, in a session, for its relationship with this document's usefulness. Also as mentioned above, total dwell time was transformed into log(10) before conducting statistical analysis. Further, the usefulness scores used the collapsed 3-point scale.

Table 5.4 GLM Univariate results of the effects of usefulness, task stage, and their interaction on mean log(10) of total dwell time in both tasks combined

Effect source (factor or interaction)	F	p
Stage	1.682	.187
Usefulness	75.402	.000
Stage*Usefulness	.817	.514

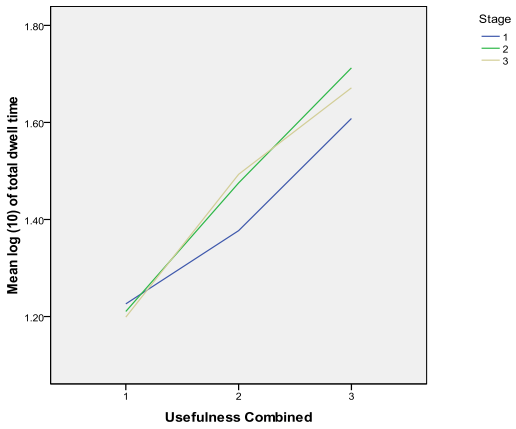


Figure 5.14 Relations between usefulness, task stage, and log(10) of total dwell time in both tasks combined

As can be seen from Table 5.4, usefulness had a significant main effect, $F(2, 990)=75.402$, $p<.001$, 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.

5.4.1.3 Decision time

Table 5.5 and Figure 5.15 report results when decision time was considered. Again, the transformed log(10) and the combined 3-group usefulness scores were used.

Table 5.5 GLM Univariate results of the effects of usefulness, task stage, and their interaction on mean log(10) of decision time in both tasks combined

Effect source (factor or interaction)	F	p
Stage	.326	.722
Usefulness	2.158	.116
Stage*Usefulness	3.619	.006

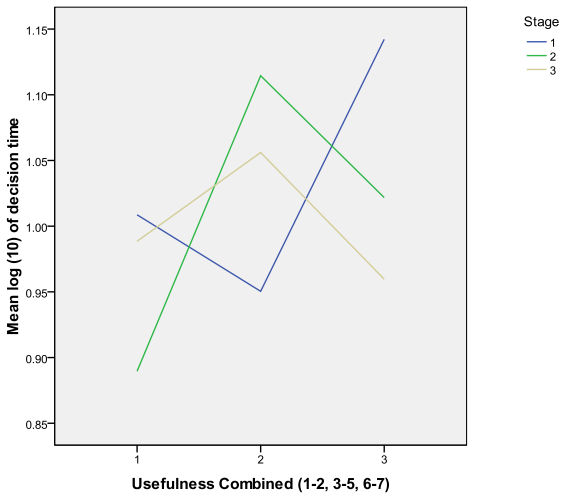


Figure 5.15 Relations between usefulness, task stage, and log(10) of decision time in both tasks combined

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, $F(4, 988)=3.619$, $p<.01$, between usefulness and stage on the log(10) of decision time. The patterns of decision time in the 3 stages were very different. Specifically, in stage 1, decision time for somewhat useful documents was the lowest, but users spent more decision time on non-useful documents, and even more time on very useful documents. However, in stage 2, the pattern was exactly reversed. The decision time for somewhat useful documents was the longest, followed by that for very useful documents, and then the non-useful documents. This showed that in stage 2, users did not spend as long a time for very useful documents as they did for somewhat useful documents. In stage 3, the decision time for not useful and very useful documents was both shorter than that for somewhat useful documents.

5.4.2 Sub-question 1b

Sub-question 1b: In the dependent task, does the stage of the user's task help in interpreting time as an indicator of document usefulness?

The previous sub-question RQ1a considered both types of task when they were combined together. The current sub-question RQ1b looks particularly at the dependent task. Again, all three types of time were examined, and all relationships were examined using the transformed logarithm (base of 10) of time.

5.4.2.1 Total display time

Just as in RQ1a, in RQ1b, it is also reasonable to consider the total display time of each document, rather than the individual display time for each time it was open, in a session, for its relationship with this document's usefulness. The following reports data analysis results when total display time was considered.

In order to collapse the usefulness scores from the 7-point scale to 3 groups, an examination of usefulness data's distribution in the dependent task was conducted, as shown in Figure 5.16.

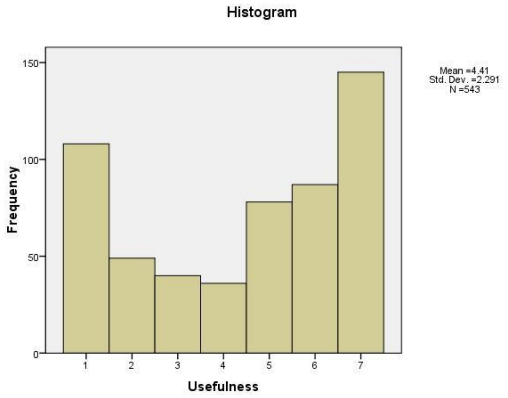


Figure 5.16 The distribution of original usefulness data in the dependent task

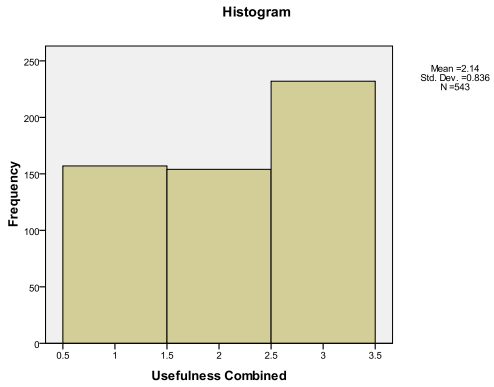


Figure 5.17 The distribution of combined usefulness data in the dependent task

From this distribution, it seems reasonable to combine scores 1- 2 into a little useful group, 3-5 into a somewhat (or medium) useful group, and 6-7 into a very useful group. Figure 5.17 showed the distribution after grouping, where the 3 groups were quite balanced.

Table 5.6 and Figure 5.18 showed the relation between stage, the combined usefulness (in the following of this subsection, Section 5.4.2, usefulness means combined usefulness unless specified), as well their interaction effects on log(10) of total display time.

Table 5.6 GLM Univariate results of the effects of usefulness, task stage, and their interaction on mean log(10) of total display time in the dependent task

Effect source (factor or interaction)	F	p
Stage	1.959	.142
Usefulness	63.404	.000
Stage*Usefulness	1.905	.108

Results show that usefulness had a main effect, $F(2, 540)=63.404$, $p<.001$, meaning that in the dependent task, the relation between the combined usefulness and log(10) of total display

time was significant. Total display time could be a reliable indicator of document usefulness. Stage did not have significant main effect, meaning that the relation between stage and log(10) of total display time was not significant. There was no significant interaction effect between stage and usefulness on log(10) of total display time as well.

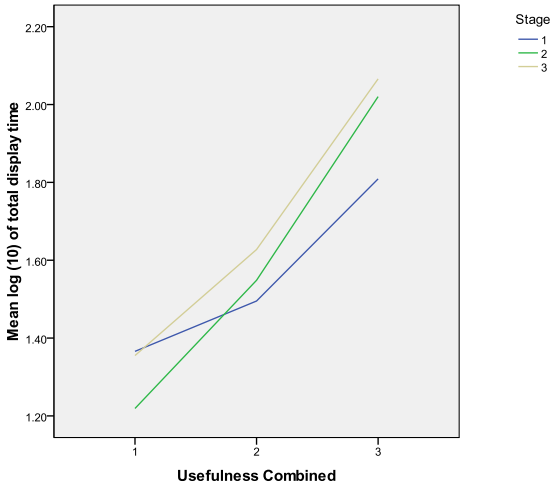


Figure 5.18 Relations between usefulness, task stage, and log(10) of total display time in the dependent task

5.4.2.2 Total dwell time

Table 5.7 and Figure 5.19 report results when total dwell time, i.e., reading time (or viewing time) was considered. Just as was done with the total display time in the previous subsection, it is also reasonable to consider the total dwell time of each document, rather than the individual dwell time for each time it was open, in a session, for its relationship with this document's usefulness.

Table 5.7 GLM Univariate results of the effects of usefulness, task stage, and their interaction on mean log(10) of total dwell time in the dependent task

Effect source (factor or interaction)	F	p
Stage	1.290	.276
Usefulness	35.225	.000
Stage*Usefulness	.829	.507

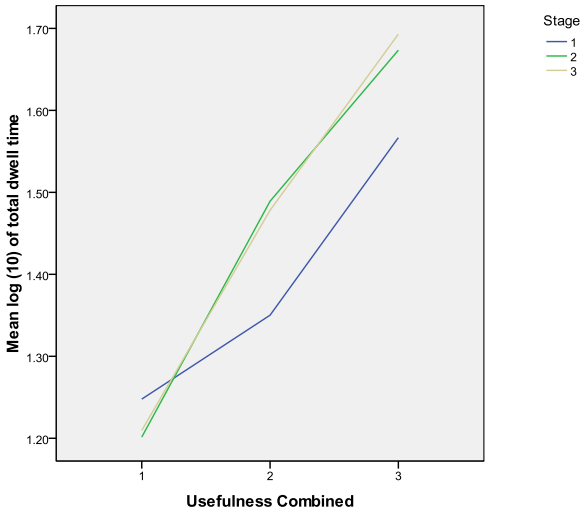


Figure 5.19 Relations between usefulness, task stage, and log(10) of total dwell time in the dependent task

Results show that in the dependent task, usefulness had a significant main effect on log(10) of total display time, $F(2, 540)=35.225$, $p<.001$, meaning that the relations between usefulness and log(10) of total display time was significant. Total dwell time could be a reliable indicator of document usefulness. Stage did not have a significant main effect, nor was there a significant interaction effect between stage and usefulness with regard to time. Figure 5.19 shows that the total display time patterns in stage 2 and stage 3 were almost identical.

5.4.2.3 Decision time

When decision time was considered, the following results (Table 5.8 and Figure 5.20) were obtained.

Table 5.8 GLM Univariate results of the effects of usefulness, task stage, and their interaction on mean log(10) of decision time in the dependent task

Effect source (factor or interaction)	F	p
Stage	.790	.454
Usefulness	3.336	.036
Stage*Usefulness	1.572	.180

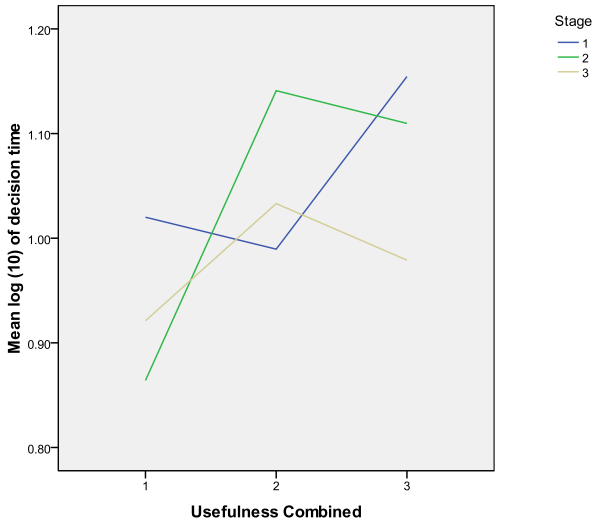


Figure 5.20 Relations between usefulness, task stage, and log(10) of decision time in the dependent task

In the dependent task, usefulness showed a significant main effect on log(10) of decision time, $F(2, 540)=3.336$, $p<.05$, meaning that the relation between usefulness and log(10) of decision time was significant. This indicated that longer decision time would indicate that the documents were more useful. However, stage did not have a significant main effect, nor was there a significant interaction effect between stage and usefulness on decision time.

5.4.3 Sub-question 1c

Sub-question 1c: In the parallel task, does the stage of the user's task help in interpreting time as an indicator of document usefulness?

5.4.3.1 Total display time

Again, the total display time of each document instead of the individual display time for each time it was open was considered for its relationship with this document's usefulness. The following reports data analysis results when total display time was considered in the parallel task.

Usefulness was again examined when it was collapsed into smaller groups. Figure 5.21 shows its original distribution, and Figure 5.22 shows the distribution after usefulness scores

were combined into 3 groups: 1-2 into a little useful group, 3-5 into a somewhat useful group, and 6-7 into a very useful group. In the following of this subsection, Section 5.4.3, usefulness refers to the combined usefulness unless specified.

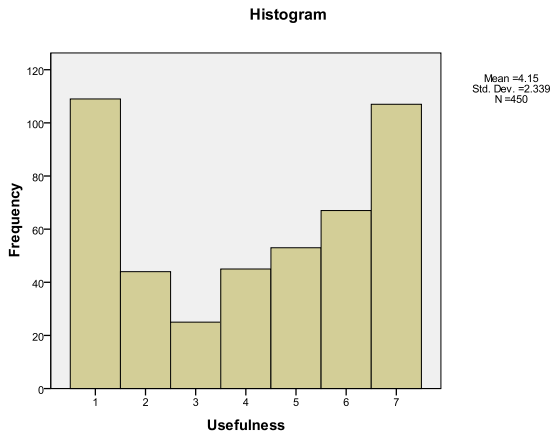


Figure 5.21 The distribution of original usefulness data in the parallel task

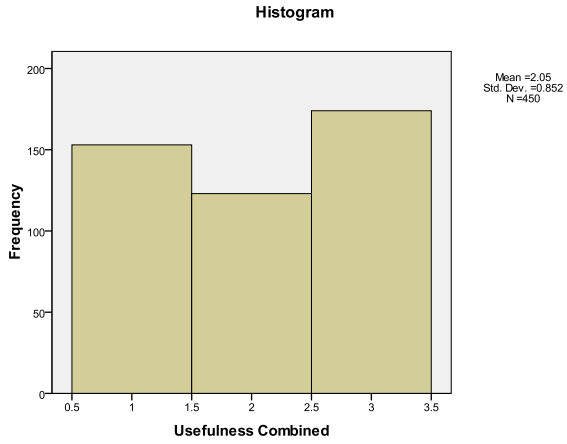


Figure 5.22 The distribution of combined usefulness data in the parallel task

Table 5.9 GLM Univariate results of the effects of usefulness, task stage, and their interaction on mean log(10) of total dwell time in the parallel task

Effect source (factor or interaction)	F	p
Stage	2.402	.092
Usefulness	61.110	.000
Stage*Usefulness	1.393	.236

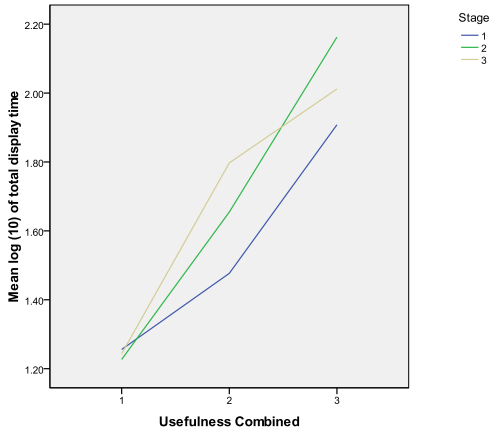


Figure 5.23 Relations between usefulness, task stage, and log(10) of total display time in the parallel task

In the parallel task, results (Table 5.9 and Figure 5.23) show that usefulness had a significant main effect on log(10) of total display time, $F(2, 447)=61.110$, $p<.001$, meaning that the relation between usefulness and log(10) of total display time was significant. Stage did not have a significant main effect, nor was there a significant interaction effect between stage and usefulness on total display time.

5.4.3.2 Total dwell time

The following reports results when total dwell time was considered. Just as was done in the previous subsections, it is also reasonable to consider the total dwell time of each document, rather than the individual dwell time for each time it was open, in a session, for its relationship with this document's usefulness. Table 5.10 and Figure 5.24 show the effects of usefulness and stage on log(10) of total dwell time in the parallel task.

Table 5.10 GLM Univariate results of the effects of usefulness, task stage, and their interaction on mean log(10) of total dwell time in the parallel task

Effect source (factor or interaction)	F	P
Stage	.477	.621
Usefulness	40.781	.000
Stage*Usefulness	.425	.791

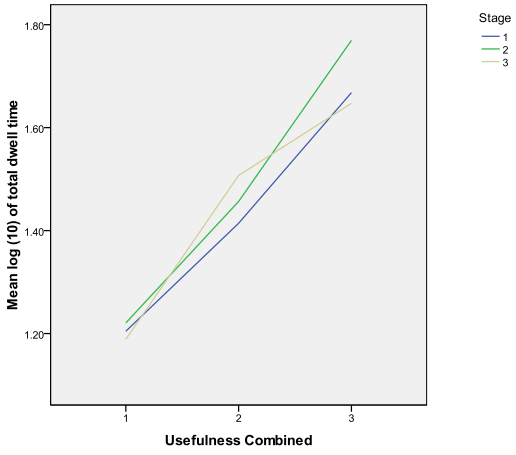


Figure 5.24 Relations between usefulness, task stage, and log(10) of total dwell time in the parallel task

In the parallel task, results show that usefulness had a significant main effect on time, $F(2, 447)=40.781$, $p<.001$, meaning that the relation between usefulness and log(10) of total dwell time was significant. Stage did not have a main effect, nor was there a significant interaction effect between stage and usefulness on total dwell time.

5.4.3.3 Decision time

Table 5.11 and Figure 5.25 show the relation between stage, the combined usefulness, as well their interaction effects on decision time in the parallel task.

Table 5.11 GLM Univariate results of the effects of usefulness, task stage, and their interaction on mean log(10) of decision time in the parallel task

Effect source (factor or interaction)	F	p
Stage	.449	.639
Usefulness	.140	.869
Stage*Usefulness	2.478	.043

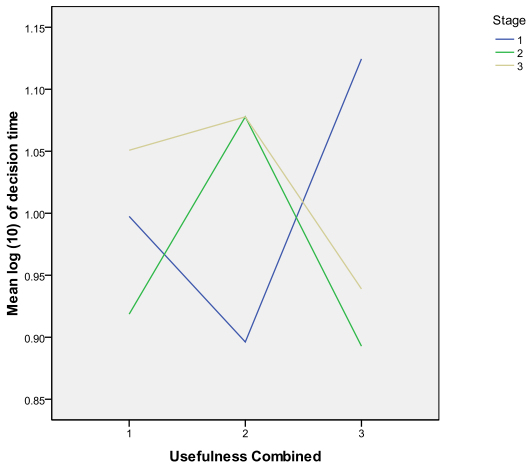


Figure 5.25 Relations between usefulness, task stage, and log(10) of decision time in the parallel task

Results show that in the parallel task, neither usefulness nor stage had a significant main effect on log(10) of decision time, meaning that neither stage nor usefulness had a significant relation with decision time. However, the interaction between stage and the combined usefulness was significant, $F(4, 445)=2.478$, $p<.05$, on decision time. The relation between usefulness and

time varies across different stages. Specifically, in stage 1, decision time for somewhat useful documents was the lowest, but users spent more decision time on non-useful documents, and even more time on very useful documents. However, in stage 2, the pattern was exactly reversed. The decision time for somewhat useful documents was the longest, followed by that for very useful documents, and then the non-useful documents. This shows that in stage 2, users did not spend as long a time for very useful documents as they did for somewhat useful documents. In stage 3, the decision time for very useful documents was very short, shorter than that for little useful documents, which was shorter than somewhat useful documents.

5.4.4 Summary of results of RQ1

This section summarizes the above results for RQ1.

Table 5.12 shows the summary of the main effect of Usefulness (combined), the main effect of stage, and the interaction effect of stage and usefulness (combined). Data are F and p (in parenthesis) values obtained in GLM analyses.

Table 5.12 Summary of the *F*(p) values of factors (obtained from GLM analyses)

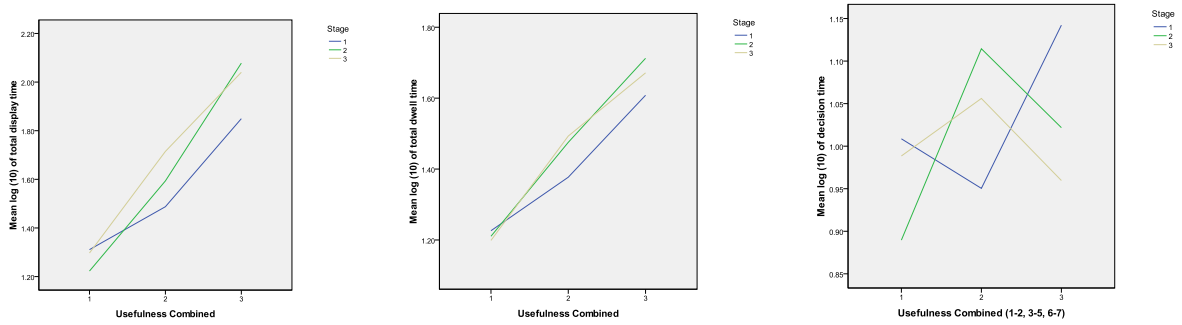
Task	Type of time	Stage	Usefulness	Stage*Usefulness
In both tasks combined	Log(10) total display time	4.150(.016)	123.779(.000)	2.658(.032)
	Log(10) total dwell time	1.682(.187)	75.402(.000)	.817(.514)
	Log(10) decision time	.326(.722)	2.158(.116)	3.619(.006)
Dependent task	Log(10) total display time	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)

(Those in bold were statistically significant. The p values are in the parentheses.)

The following (Figure 5.26) show the figures again showing the relations between time, stage, and usefulness:

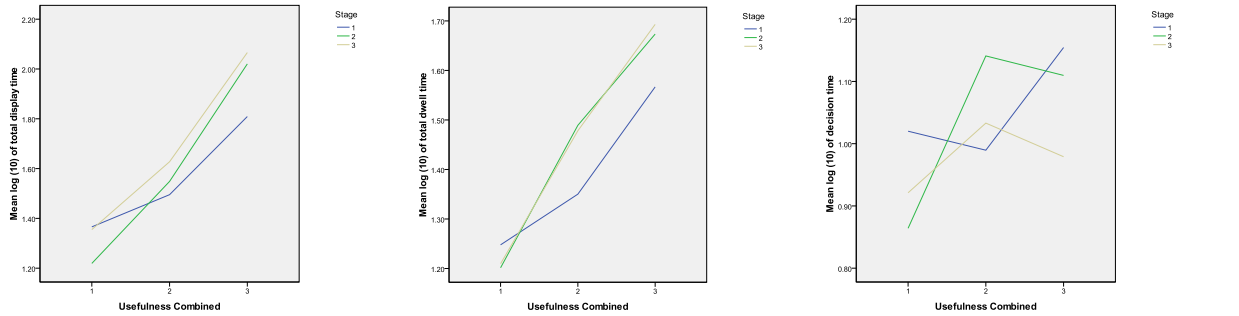
In both tasks combined

Log(10) total display time Log(10) total dwell time Log(10) decision time



In the dependent task:

Log(10) total display time Log(10) total dwell time Log(10) decision time



In the parallel task:

Log(10) total display time Log(10) total dwell time Log(10) decision time

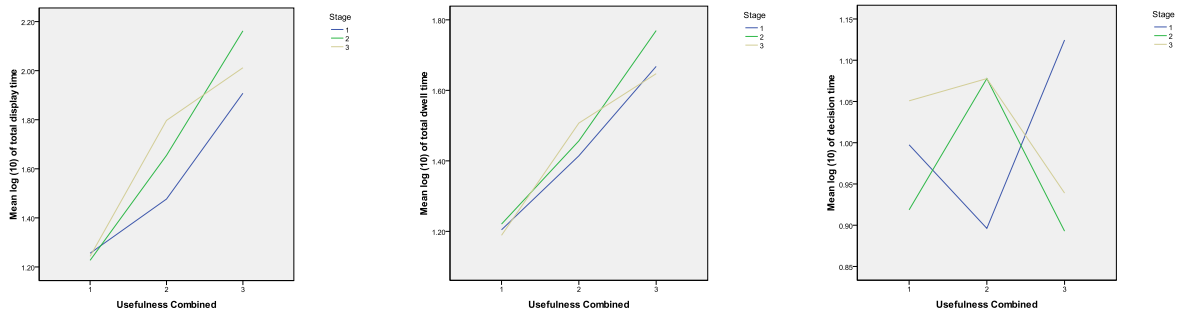


Figure 5.26 Relations of time, usefulness, and stage

From the figures, one can see that decision time had very different patterns than the other two types of time, with which usefulness had positive correlation. From Table 5.12, one can further see that the effects of stage and usefulness also varied for different types of times, as well as 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.

Based on the analysis, RQ 1 is answered as follows:

RQ1: Does the stage of the user's task help in interpreting time as an indicator of document usefulness?

Based on the findings in the current study, the brief answer to RQ1 is: yes, when total display time or decision time was considered; but no, when total dwell time was considered.

Sub-question 1a: In general, i.e., in both the parallel and the dependent tasks, does the stage of the user's task help in interpreting time as an indicator of document usefulness?

The answer to this sub-question is: yes, when total display time or decision time was considered (a significant 3-way relationship was found among stage, usefulness, and total display time, specifically, there was a significant interaction effect between stage and usefulness on total display time; so was stage, usefulness, and decision time, i.e., there was a significant interaction effect between stage and usefulness on decision time); but no, when total dwell time was considered (no significant 3-way relationship was found among stage, usefulness, and total dwell time.

Sub-question 1b: In the dependent task, does the stage of the user's task help in interpreting time as an indicator of document usefulness?

The answer to this sub-question is: no, when any of the three types of time: total display time, total dwell time, and decision time, was considered. No significant 3-way relationships were found among stage, usefulness, and any type of time.

Sub-question 1c: In the parallel task, does the stage of the user's task help in interpreting time as an indicator of document usefulness?

The answer to this sub-question is: yes, when decision time was considered (a significant 3-way relationship was found among stage, usefulness, and decision time, i.e., there was a significant interaction effect between stage and usefulness on decision time); but no, when total display time or total dwell time was considered (no significant 3-way relationship was found among stage, usefulness, and total display time, or among stage, usefulness, and total dwell time).

5.5 Results of RQ2

RQ2: Does the user's topic knowledge help in interpreting time as an indicator of document usefulness?

This RQ looks at the relationships among the user's knowledge of the task topic, document usefulness, task type, and time. Topic knowledge here is 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, so there were two scores for general task topic knowledge in each session: pre- and post-session general task topic knowledge. In addition, users were also asked to assess their familiarity degrees with the sub-tasks both before and after each session. Therefore, there were also two scores for sub-task topic knowledge in each session: pre- and post-session sub-task topic knowledge. In total, there

were four types of topic knowledge: pre- and post-session general task topic knowledge, and pre- and post-session sub-task topic knowledge.

In this RQ, we first looked at the patterns of these different types of topic knowledge, which was aimed at answering sub-RQ2a. Then, similarly as RQ1, for RQ2, we also looked at sub-questions according to the different task types, and examined each of them respectively, which aimed at answering sub-RQs 2b, 2c, and 2d. Finally, topic knowledge was compared with task stage (the factor examined in RQ1) in terms of their roles in helping interpret time as an indicator of document usefulness, which was sub-RQ2e. In sum, RQ2 consists of 5 sub-questions.

5.5.1 Sub-question 2a

Sub-question 2a: What are the patterns of the users' pre- and post-session general task topic knowledge and pre- and post-session sub-task topic knowledge?

5.5.1.1 Descriptive data

Examination of the normality and distribution found that all the above four types of topic knowledge were not normally distributed. Nevertheless, it is still intuitively helpful to look at the mean and standard deviation of them. Table 5.13 showed the descriptive data of the four types of topic knowledge across 3 stages. Figure 5.27 shows the change of the pre- and post-session task topic knowledge across 3 stages.

5.5.1.1.1 Pre-session task topic knowledge ratings in both tasks in general

Table 5.13 Mean and standard deviation of four types of topic knowledge at 3 stages

	Mean (Standard Deviation)			
Stage	Pre-session general task topic knowledge	Post-session general task topic knowledge	Pre-session sub-task topic knowledge	Post-session sub-task topic knowledge
1	2.75 (1.51)	4.25(1.03)	2.75(1.51)	4.29(1.46)
2	3.79(1.53)	4.96(1.12)	2.75(1.70)	5.00(.93)
3	4.75(1.80)	5.42(1.02)	3.00(1.91)	5.38(1.28)

	Mean (Standard Deviation)			
Stage	Pre-session general task topic knowledge	Post-session general task topic knowledge	Pre-session sub-task topic knowledge	Post-session sub-task topic knowledge
1	2.75 (1.51)	4.25(1.03)	2.75(1.51)	4.29(1.46)
2	3.79(1.53)	4.96(1.12)	2.75(1.70)	5.00(.93)
3	4.75(1.80)	5.42(1.02)	3.00(1.91)	5.38(1.28)
Total	3.76(1.80)	4.88(1.15)	2.83 (1.70)	4.89(1.31)

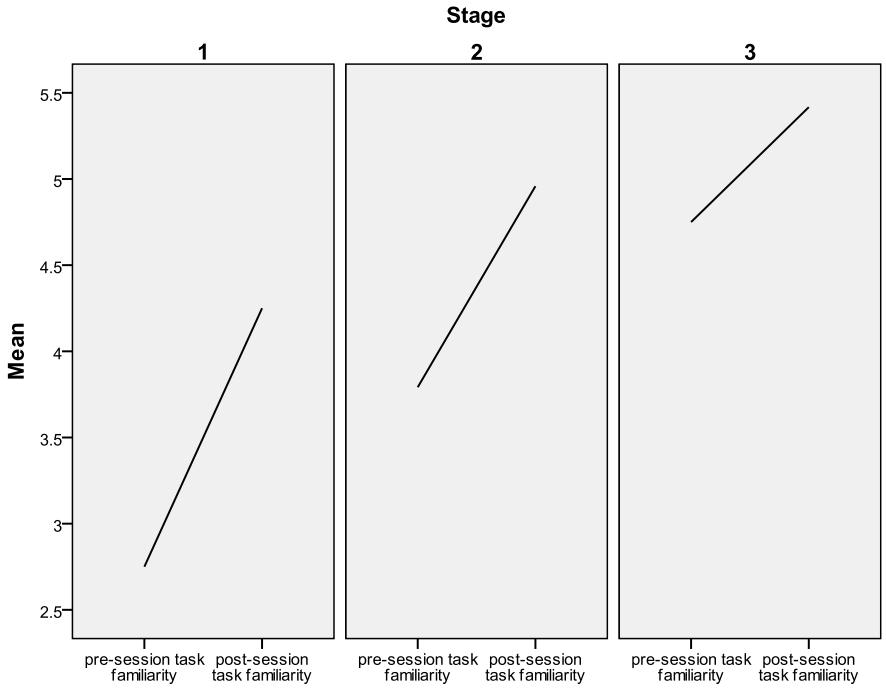


Figure 5.27 Pre- and post-session general task topic knowledge across 3 stages

Figure 5.27 shows the tendency of the changes of pre- and post-session general task topic knowledge across the 3 stages. As can been seen, in general, post-session general task topic knowledge was higher than pre-session, and it was reasonable that the pre-session task topic knowledge was a bit lower than the previous session's post-session task topic knowledge since in the beginning of a session, participants may have forgotten some of what they had learned in the previous session.

Figure 5.28 shows the tendency of the changes of pre- and post-session sub-task topic knowledge across the 3 stages. As can been seen, in general, post-session knowledge was higher than pre-session, however, pre-session sub-task topic knowledge did not change across stages, meaning that the users did not gain knowledge on sub-tasks, which was reasonable considering that the sub-tasks were different across stages.

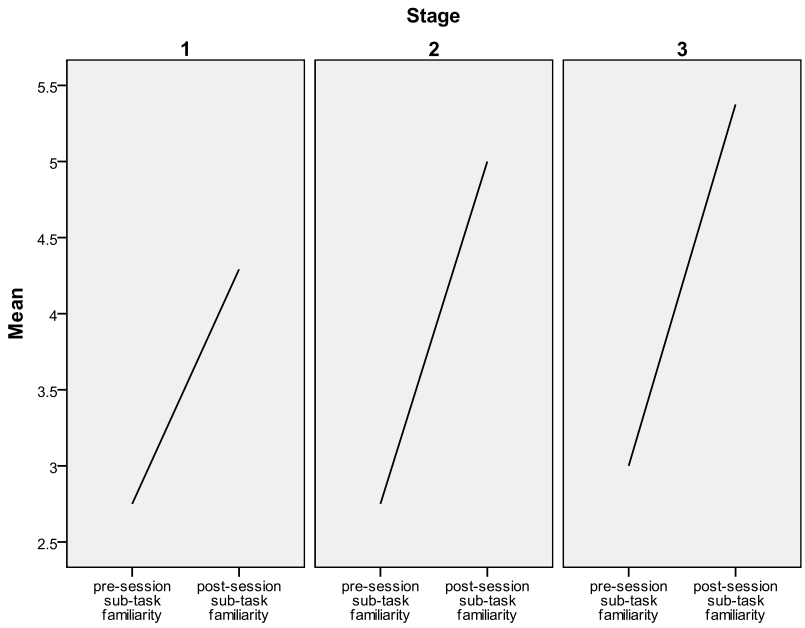


Figure 5.28 Pre- and post-session sub-task topic knowledge across 3 stages

5.5.1.1.2 Topic knowledge ratings in the two types of tasks

Table 5.14 shows the mean and standard deviation of the four types of topic knowledge in the 3 stages, as well as in the dependent (denoted as task 1) and the parallel (denoted as task 2) tasks.

Figure 5.29 shows the tendency of the changes of pre- and post-session general task topic knowledge in two tasks across the 3 stages. As can been seen, in general, post-session general task topic knowledge is higher than pre-session, except for stage 3 in the parallel task (task 2).

Pre-session task topic knowledge in later stages was higher than that in the earlier stages. In addition, topic knowledge of the general task (task 2) in the parallel task was higher than that in the dependent task (task 1).

Table 5.14 Mean and Standard Deviation of four types of knowledge in two types of tasks at 3 stages

Stage	Task	Mean (Standard Deviation)			
		Pre-session general task topic knowledge	Post-session general task topic knowledge	Pre-session sub-task topic knowledge	Post-session sub-task topic knowledge
1	dependent	2.33 (1.073)	2.50 (1.382)	4.33 (1.155)	4.17 (0.389)
	parallel	3.17 (1.801)	3.00 (1.651)	4.25 (1.765)	4.33 (1.435)
	Total	2.75 (1.511)	2.75 (1.511)	4.29 (1.459)	4.25 (1.032)
2	dependent	3.58 (1.443)	2.50 (1.508)	4.67 (0.651)	4.75 (0.622)
	parallel	4.00 (1.651)	3.00 (1.907)	5.33 (1.073)	5.17 (1.467)
	Total	3.79 (1.532)	2.75 (1.700)	5.00 (0.933)	4.96 (1.122)
3	dependent	4.08 (1.676)	2.75 (1.765)	5.25 (1.357)	5.50 (1.000)
	parallel	5.42 (1.730)	3.25 (2.094)	5.50 (1.243)	5.33 (1.073)
	Total	4.75 (1.800)	3.00 (1.911)	5.38 (1.279)	5.42 (1.018)
	dependent	3.33 (1.568)	2.58 (1.519)	4.75 (1.131)	4.81 (0.889)
	parallel	4.19 (1.925)	3.08 (1.842)	5.03 (1.464)	4.94 (1.372)
Total	Total	3.76 (1.796)	2.83 (1.695)	4.89 (1.306)	4.88 (1.150)

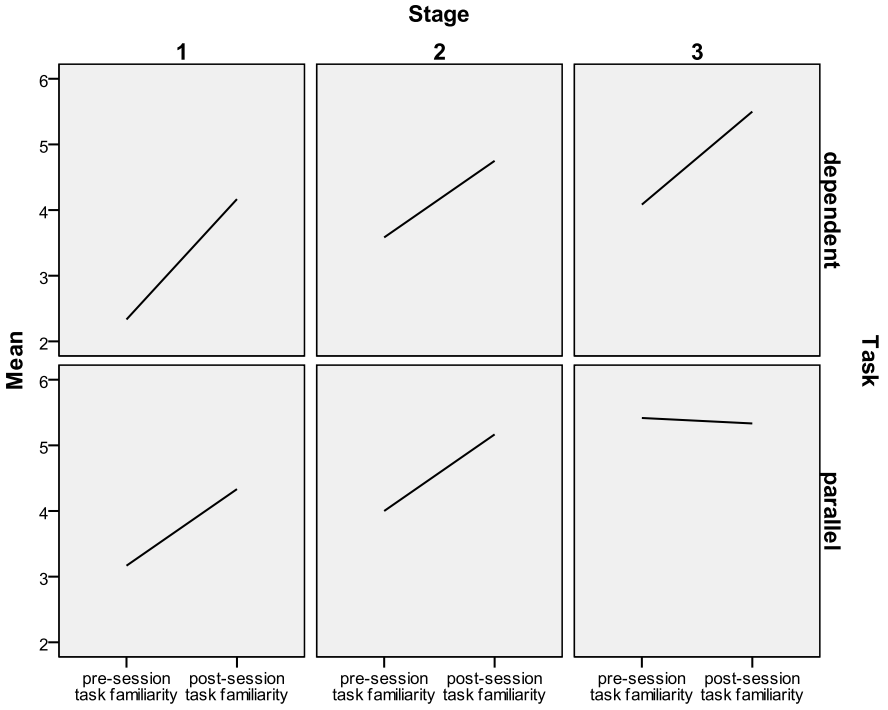


Figure 5.29 Pre- and post-session general task topic knowledge across 3 stages in both tasks

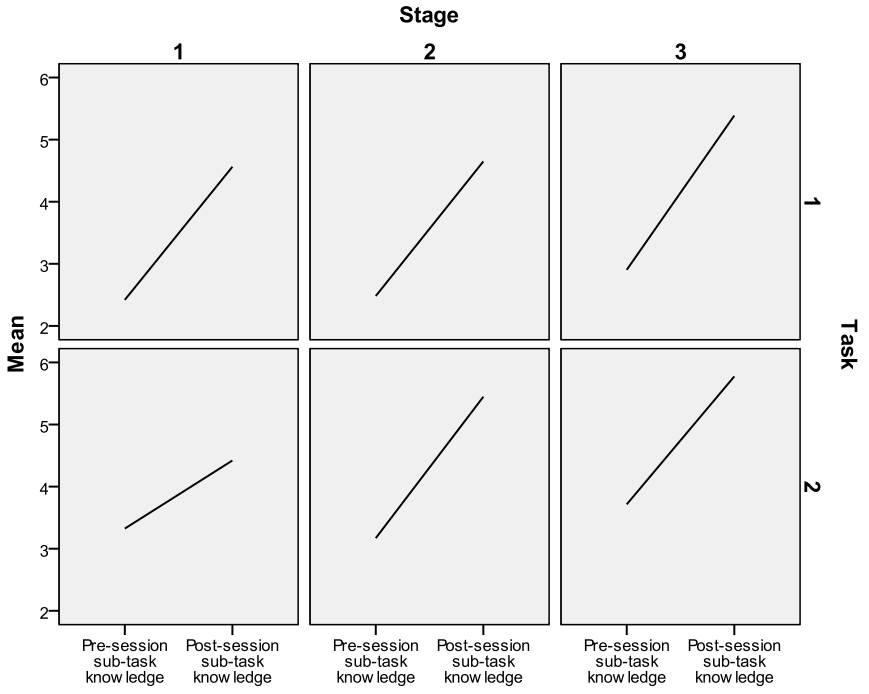


Figure 5.30 Pre- and post-session sub-task topic knowledge across 3 stages in both tasks

Figure 5.30 shows the tendency of the changes of pre- and post-session sub-task topic knowledge in two tasks across the 3 stages. As can been seen, in general, post-session general task topic knowledge is higher than pre-session. In addition, topic knowledge of the general task in the parallel task (task 2) was higher than that in the dependent task (task 1). However, the pre-session sub-task topic knowledge in later stages was not higher than that in the previous stages. The following sections provide statistical analyses results which examine which type(s) of topic knowledge showed significant differences across stages and between tasks.

5.5.1.2 Within-stage comparison for general task and sub-task topic knowledge

5.5.1.2.1 Pre- and post-session topic knowledge in all three stages in general

As mentioned before, all four types of topic knowledge were not normally distributed, so non-parametric Wilcoxon paired test was used to compare their differences.

Test Statistics[b]

	Pre-session task topic knowledge - Post-session task topic knowledge	Pre-session sub-task topic knowledge - Post-session sub-task topic knowledge
Z	-5.045[a]	-6.311[a]
Asymp. Sig. (2-tailed)	.000	.000

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

The above data show that for general task topic knowledge, the post-session rating score was higher than the pre-session rating score (Z=-5.05, p<.001). This is also true for the sub-task topic knowledge (Z=-6.31, p<.001).

5.5.1.2.2 Within-stage comparison of pre- and post-session topic knowledge in stage 1

The following data show that in stage 1, post-session general task topic knowledge score is higher than the pre-session score (Z=-3.32, p=.001). This is also the case for the sub-task topic knowledge (Z=-3.00, p<.005).

Test Statistics[b]

	Pre-session general task topic knowledge - Post-session general task topic knowledge	Pre-session sub-task topic knowledge - Post-session sub-task topic knowledge
Z	-3.323[a]	-2.996[a]
Asymp. Sig. (2-tailed)	.001	.003

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

5.5.1.2.3 Within-stage comparison of pre- and post-session topic knowledge in stage 2

Test Statistics[b]

	Pre-session task topic knowledge - Post-session task topic knowledge	Pre-session sub-task topic knowledge - Post-session sub-task topic knowledge
Z	-3.685[a]	-3.970[a]
Asymp. Sig. (2-tailed)	.000	.000

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

The above data show that in stage 2, the post-session general task topic knowledge score is higher than the pre-session score (Z=-3.69, p<.001). This is also the case for the sub-task topic knowledge (Z=-3.97, p<.001).

5.5.1.2.4 Within-stage comparison of pre- and post-session topic knowledge in stage 3

The following data show that in stage 3, the post-session rating score of general task topic knowledge is not higher than the pre-session rating score (Z=-2.384, p>.05). However, the post-

session rating score of sub-task topic knowledge is still higher than the pre-session rating score (Z=-4.04, p<.001).

Test Statistics[b]

	Pre-session task topic knowledge - Post-session task topic knowledge	Pre-session sub-task topic knowledge - Post-session sub-task topic knowledge
Z	-1.620[a]	-4.036[a]
Asymp. Sig. (2-tailed)	.105	.000

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

5.5.1.3 Between-stage comparison of topic knowledge

Test Statistics[a,b]

	Pre-session general task topic knowledge	Pre-session sub-task topic knowledge	Post-session sub-task topic knowledge	Post-session general task topic knowledge
Chi-Square	14.894	.165	8.177	13.900
df	2	2	2	2
Asymp. Sig.	.001	.921	.017	.001

a. Kruskal Wallis Test

b. Grouping Variable: Stage

Results show that the rating scores for general task topic knowledge increase along stage. Specifically, both pre-session general task topic knowledge, $\chi^2(2, N = 72) = 14.89$, $p < .005$ and post-session general task topic knowledge, $\chi^2(2, N = 72) = 13.90$, $p < .005$ increase along stages. These results mean that users did learn in the process of completing the tasks for the general task. As for sub-task topic knowledge, for pre-session subtask topic knowledge, the rating scores did not change significantly across stages, $\chi^2(2, N = 72) = .165$, $p > .05$. Since pre-session sub-task topic knowledge measured users' baseline knowledge on the different sub-tasks before working with them, it is reasonable to see that this type of user knowledge did not have differences, i.e., users showed equal knowledge on the different sub-tasks. As for the post-session sub-task topic

knowledge, user ratings also increased along stages, $\chi^2(2, N = 72) = 8.18, p < .05$. This means that although users' baseline knowledge of sub-tasks did not have differences, users perhaps learned more for the sub-tasks in the later stages.

5.5.2 Sub-question 2b

Sub-question 2b: In general, i.e., in both the parallel and the dependent tasks, does the user's topic knowledge help in interpreting time as an indicator of document usefulness?

As mentioned in Section 5.5.1, there were four types of topic knowledge. However, for the purpose of helping interpret time as an indicator of document usefulness, so that the system might help users in their search, it perhaps makes the most sense to use the pre-session, instead of the post-session, topic knowledge. General task topic knowledge elicited in the three stages measured users' knowledge increase of the same overall task, but sub-task topic knowledge in the three stages measured users' knowledge of the different sub-tasks. Since users in the study worked with a multi-session task, it is natural to consider their knowledge of the whole task for RQ2. Therefore, in this section, only the pre-session general task topic knowledge was used in investigating the relationship of general task topic knowledge with document usefulness and time. In the rest of section 5.5, topic knowledge refers to, unless specified, pre-session general task topic knowledge.

Just as was done for RQ1 in Section 5.4, in this section, all three types of time, i.e., total display time, total dwell time, and decision time use the transformed data using logarithm with a base of 10. Also as was done in Section 5.4, usefulness scores were collapsed into smaller scaled groups (3 groups) from the original 7-point rating scores elicited from the users. This was also applied to the topic knowledge scores that users originally rated based on a 7-point scale. They were combined into fewer groups since it is appropriate for the system to differentiate user

knowledge based on 3 levels: not familiar (little knowledge), somewhat familiar (some knowledge), and very familiar (much knowledge). In the rest of this section, unless specified, usefulness and topic knowledge refer to combined usefulness and combined topic knowledge.

5.5.2.1 Total display time

The following reports findings about whether topic knowledge could help in interpreting total display time as an indicator of document usefulness. To examine the relations between factors, GLM Univariate analysis was conducted in which the log(10) of total display time was used as the dependent variable while topic knowledge and usefulness were used as the independent variables (factors).

The usefulness data have been examined previously in Section 5.4 for their distribution as well as the means to collapse them. In this section's analyses, the same means groupings were used, i.e., scores 1- 2 into a low-useful group, 3-5 into a mid-useful group, and 6-7 into a high-useful group. To refresh, Figure 5.32 and Figure 5.32 depict the distributions of the original and the combined data. Again, in Section 5.5.2, unless specified, usefulness refers to collapsed usefulness.

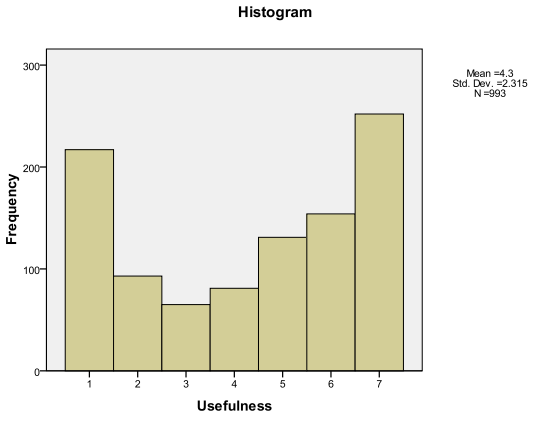


Figure 5.31 Distribution of the original usefulness data

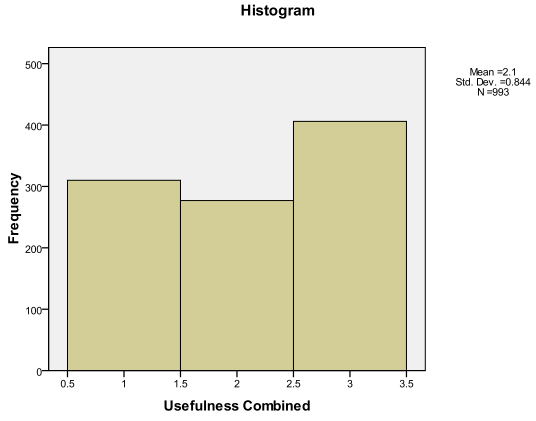


Figure 5.32. Distribution of the original usefulness data

Topic knowledge was also collapsed into groups. Figure 5.32 depicts the distribution of the original data. From the figure, it is reasonable to combine scores 1-2 into a little knowledge group, 3-4 into a medium knowledge group, and 5-7 into a much knowledge group. Figure 5.33 shows the distribution of pre-session task topic knowledge after it was collapsed into these three groups. In the rest of Section 5.5.2, unless specified, topic knowledge refers the collapsed pre-session task familiarity.

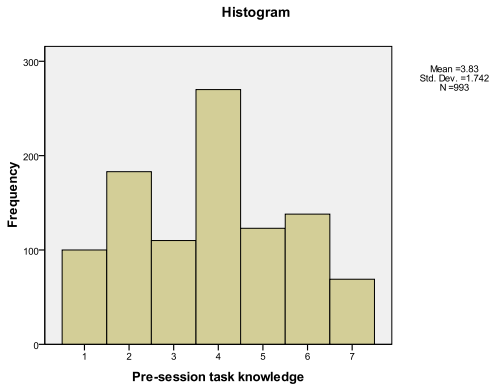


Figure 5.34 Distribution of the original topic knowledge data

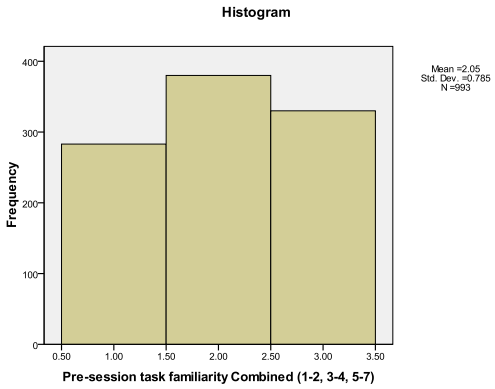


Figure 5.33 Distribution of the combined topic knowledge data

Table 5.15 GLM Univariate results of the effects of usefulness, topic knowledge, and their interaction on mean log(10) of total display time in both tasks combined

Effect source (factor or interaction)	F	p
Usefulness	123.112	.000
Topic knowledge	1.314	.269
Usefulness*Topic knowledge	3.050	.016

Table 5.15 and Figure 5.35 show the GLM Univariate results of the effects of usefulness, topic knowledge, and their interaction on mean log(10) of total display time. As the data show, usefulness had a significant main effect, $F(2, 990)=123.112$, $p<.001$, meaning that the relation between usefulness and log(10) of total display time was significant, i.e., 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, $F(4, 988)=3.050$, $p<.016$, on total display time, meaning that the relationship patterns of usefulness and total display time varied according to different levels of topic knowledge.

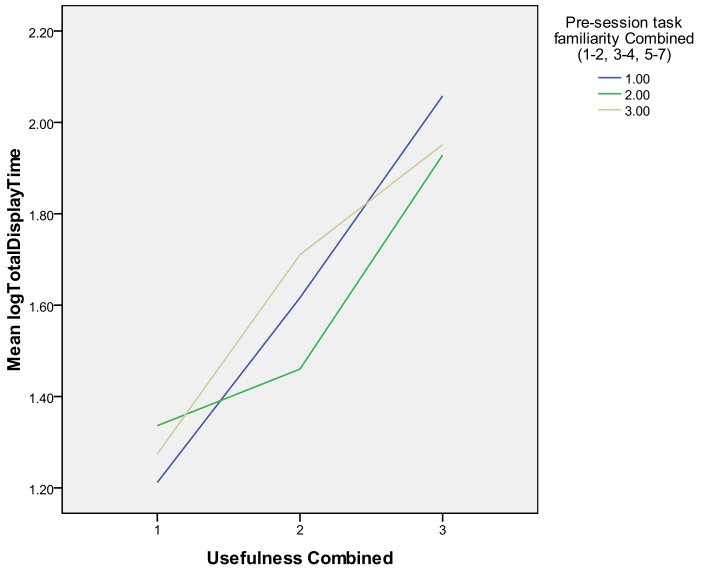


Figure 5.35 Relations between usefulness, topic knowledge, and log(10) of total display time in both tasks combined

In other words, the results indicate that in general, for all users (without differentiating their knowledge levels), the longer the documents were displayed, the more useful the documents were. For all documents (without considering their usefulness degrees), average display time of users with different topic knowledge did not show differences. However, for documents with different levels of usefulness, users with different levels of knowledge had different display time patterns. Users with different degrees of topic knowledge did not show differences in total display time for either little useful and very useful documents. However, there were differences in total display time of somewhat useful documents among users with different degrees of topic knowledge. Specifically, for documents which were somewhat useful,

those users who had much topic knowledge had them displayed longer than those who had little topic knowledge, and than those who had some topic knowledge.

5.5.2.2 Total dwell time

Table 5.16 and Figure 5.36 report findings of whether topic knowledge played roles in interpreting total dwell time as an indicator of document usefulness.

Table 5.16 GLM Univariate results of the effects of usefulness, topic knowledge, and their interaction on mean log(10) of total dwell time in both tasks combined

Effect source (factor or interaction)	F	p
Usefulness	76.114	.000
Topic knowledge	.102	.903
Usefulness*Topic knowledge	3.501	.008

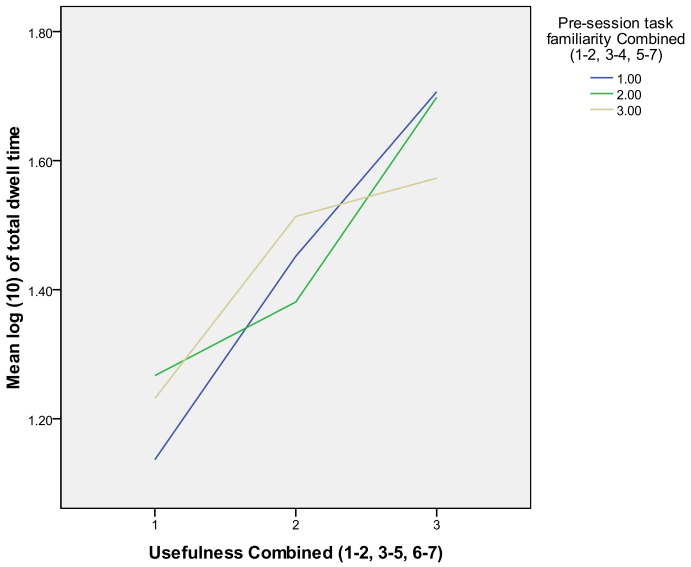


Figure 5.36 Relations between usefulness, topic knowledge, and log(10) of total dwell time in both tasks combined

Results show that usefulness had significant main effect, $F(2, 990)=76.114$, $p<.001$, on total dwell time, but topic knowledge did not. There was also a significant interaction effect between usefulness and topic knowledge, $F(4, 988)=3.501$, $p<.01$, on total dwell time.

The results indicate that in general, for all users (without differentiating their knowledge levels), the longer they dwelled on the documents, the more useful the documents were. For all documents (without considering their usefulness degrees), average dwell time of users with different topic knowledge did not have differences. However, for documents with different levels of usefulness, users with different levels of knowledge had different dwell time patterns. Users with little topic knowledge dwelled for a very short time on little useful documents, and they dwelled for a (relatively) very long time on very useful documents. Users with some topic knowledge dwelled for a longer time on little useful documents than those with little knowledge, and they had a similar dwell time as those with little knowledge. Those with much topic knowledge dwelled for a similar time on little useful documents as those having some knowledge, which was also longer compared with those with little knowledge, but for very useful documents, they dwelled for a shorter time than those who had only some or little knowledge.

5.5.2.3 Decision time

Table 5.17 and Figure 5.37 report findings of whether topic knowledge played a role in interpreting decision time as an indicator of document usefulness.

Table 5.17 GLM Univariate results of the effects of usefulness, topic knowledge, and their interaction on mean log(10) of decision time in both tasks combined

Effect source (factor or interaction)	F	p
Usefulness	3.170	.042
Topic knowledge	4.498	.011
Usefulness*Topic knowledge	3.039	.017

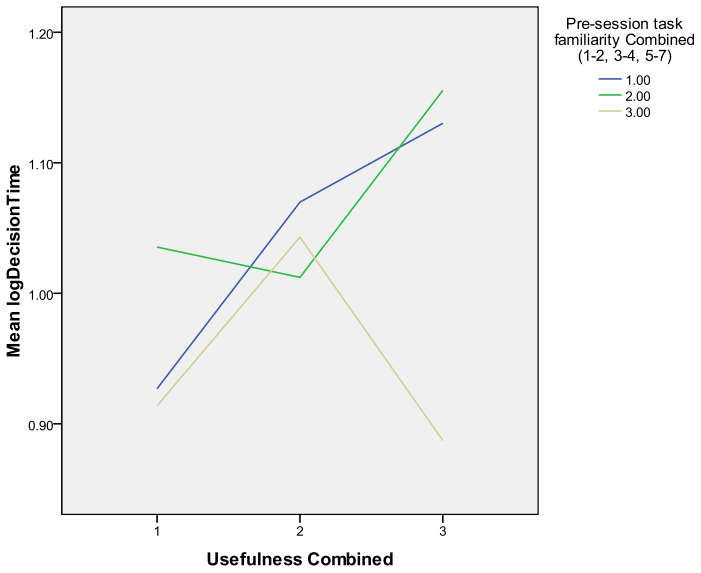


Figure 5.37 Relations between usefulness, topic knowledge, and log(10) of decision time in both tasks combined

Although in Table 5.17, usefulness had a significant p value, 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 logarithm (10) of decision time, $F(2, 990)=4.498$, $p<.05$. In addition, there was a significant interaction effect between usefulness and the topic knowledge on log(10) of decision time, $F(4, 988)=3.039$, $p< .05$.

These findings indicate that in general, all users (without considering their topic knowledge levels) seemed to have equally quickly decided the usefulness of retrieved documents which had different levels of usefulness. For all documents that they viewed, users with different levels of knowledge spent different times judging the usefulness of the documents. Specifically, those who had much knowledge made decisions more quickly than those with only some or little knowledge. For documents with different levels of usefulness, users with different levels of knowledge also had different decision time patterns. As Figure 5.37 shows, for little useful documents, those users who had little knowledge and much knowledge had lower decision

time compared with those who had medium knowledge. For the somewhat useful documents, there were no differences in decision time between/among users with all levels of knowledge. For the very useful documents, those who had much knowledge had much lower decision time than those had little or some knowledge on the task topics.

Looking at the users with different levels of knowledge, users with little topic knowledge spent a very short time deciding that a document was little useful, and they spent a very long time deciding that a document was very useful. Users with some topic knowledge spent a longer time than those had little knowledge to determine that a document was little useful, but they spent a similar length of time as those had little knowledge to judge whether a document was very useful. Meanwhile, those with much topic knowledge appeared to have judged as quickly as those with little knowledge if a document was little useful, but judged more quickly on very useful documents than those with only some or little knowledge.

5.5.3 Sub-question 2c

Sub-question 2c: In the dependent task, does the user's topic knowledge help in interpreting time as an indicator of document usefulness?

Unlike the previous sub-question RQ2a which considered both types of task, the current sub-question RQ2b looks particularly at the dependent task. Again, all three types of time were examined. Usefulness and topic knowledge were collapsed into groups.

5.5.3.1 Total display time

For usefulness data, the grouping was the same as what was used for combining usefulness scores for RQ1b, as described in Section 5.2.2. To help refresh, scores 1- 2 were combined into one group, 3-5 into the 2nd group, and 6-7 into the 3rd group. Figure 5.38 and

Figure 5.39 show the distributions before and after grouping. In Section 5.5.3, unless specified, usefulness refers to the combined usefulness.

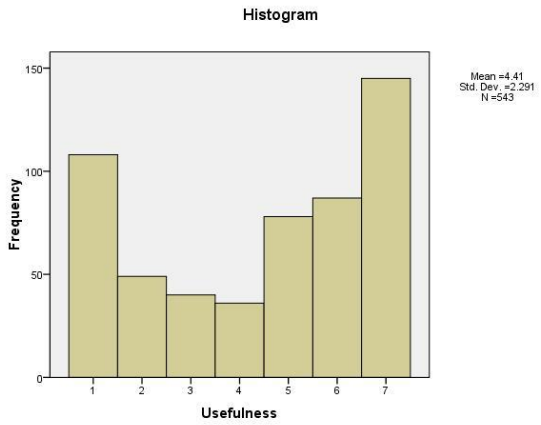


Figure 5.38 Distribution of the original usefulness data in the dependent task

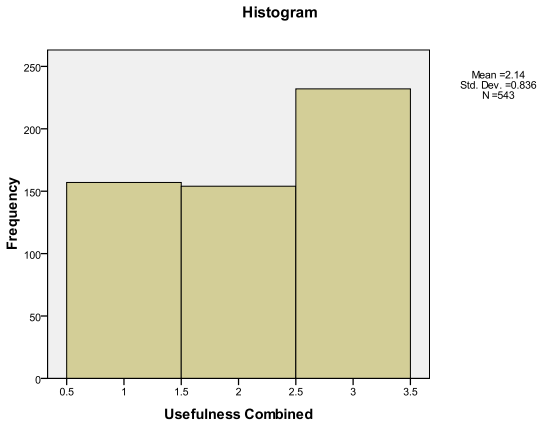


Figure 5.39 Distribution of the combined usefulness data in the dependent task

For topic knowledge, in order to collapse it into fewer groups, the distribution of the original data in the dependent task was first examined, as shown in Figure 5.41. It seems reasonable to combine 1-2 into a little knowledge group, 3-4 into a some knowledge group, and 5-7 into a much knowledge group. The distribution after collapsing is depicted by Figure 5.40. In the following part of Section 5.5.3, unless specified, topic knowledge refers to the combined pre-session task topic knowledge.

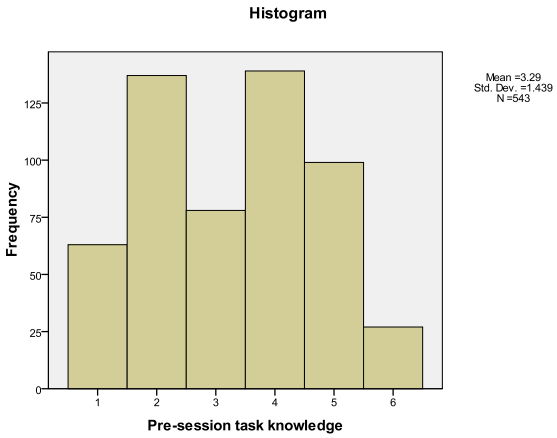


Figure 5.41 Distribution of the original topic knowledge data in the dependent task

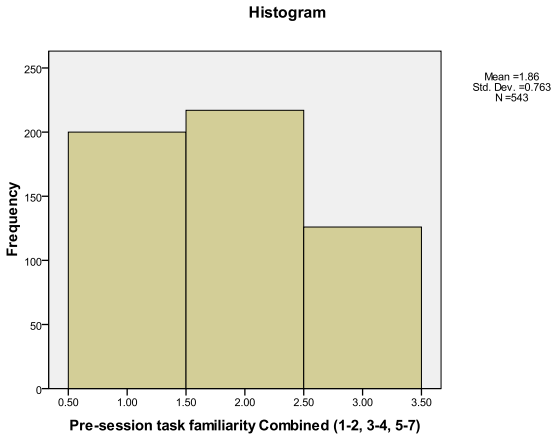


Figure 5.40 Distribution of the combined topic knowledge data in the dependent task

Table 5.18 and Figure 5.42 describe the relations between topic knowledge, usefulness, their interaction, as well as logarithm (10) of total display time.

Table 5.18 GLM Univariate results of the effects of usefulness, topic knowledge, and their interaction on mean log(10) of total display time in the dependent task

Effect source (factor or interaction)	F	p
Usefulness	60.574	.000
Topic knowledge	2.252	.106
Usefulness*Topic knowledge	1.345	.252

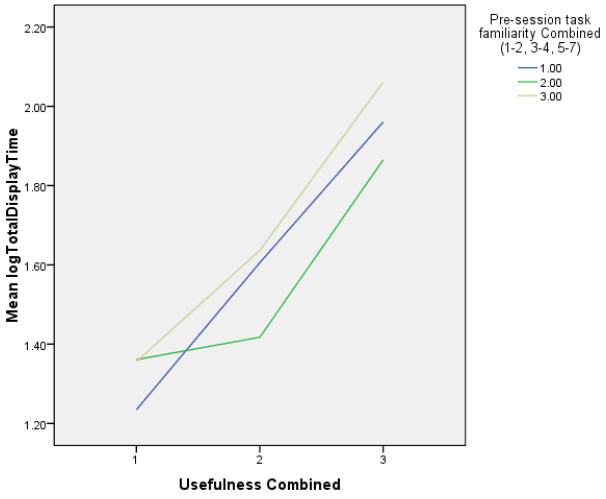


Figure 5.42 Relations between usefulness, topic knowledge, and log(10) of total display time in the dependent task

Results show that usefulness had a significant main effect, $F(2, 540)=60.574$, $p<.001$, on log(10) of total display time, meaning that the more useful the documents, the longer the total display time. This suggests that total display time could be a reliable indicator of document usefulness. Topic knowledge did not show a significant main effect. There was no interaction effect, either.

5.5.3.2 Total dwell time

The following reports findings about whether topic knowledge played a role in interpreting total dwell time as an indicator of document usefulness in the dependent task. Table 5.19 shows the GLM Univariate analysis results using log(10) of total dwell time as the dependent variable, and usefulness and topic knowledge as the factors. Figure 5.43 shows the relations between these variables.

Table 5.19 GLM Univariate results of the effects of usefulness, topic knowledge, and their interaction on mean log(10) of total dwell time in the dependent task

Effect source (factor or interaction)	F	p
Usefulness	29.806	.000
Topic knowledge	.751	.472
Usefulness*Topic knowledge	1.865	.115

Usefulness was found to have a significant effect, $F(2, 540)=29.806$, $p<.001$, on total dwell time, but topic knowledge did not. The interaction between usefulness and knowledge did not show a significant effect, either. This indicates that in general, for all users, the longer they had the documents displayed, the more useful the documents were. Mean dwell time of the documents viewed by the users with different levels of knowledge was similar. In other words, no matter how much knowledge the users had, they dwelled on their viewed documents for equal lengths of time in a session.

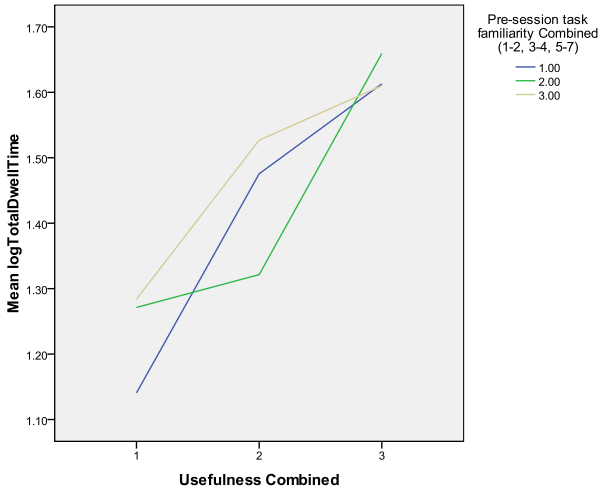


Figure 5.43 Relations between usefulness, topic knowledge, and log(10) of total dwell time in the dependent task

5.5.3.3 Decision time

The following reports findings on whether topic knowledge could help in interpreting decision time as an indicator of document usefulness. Table 5.20 shows the GLM Univariate analysis results using log(10) of decision time as the dependent variable, and usefulness and topic knowledge as the factors. Figure 5.44 shows the relations between these variables.

Table 5.20 GLM Univariate results of the effects of usefulness, topic knowledge, and their interaction on mean log(10) of decision time in the dependent task

Effect source (factor or interaction)	F	p
Usefulness	3.312	.037
Topic knowledge	1.097	.335
Usefulness*Topic knowledge	.940	.440

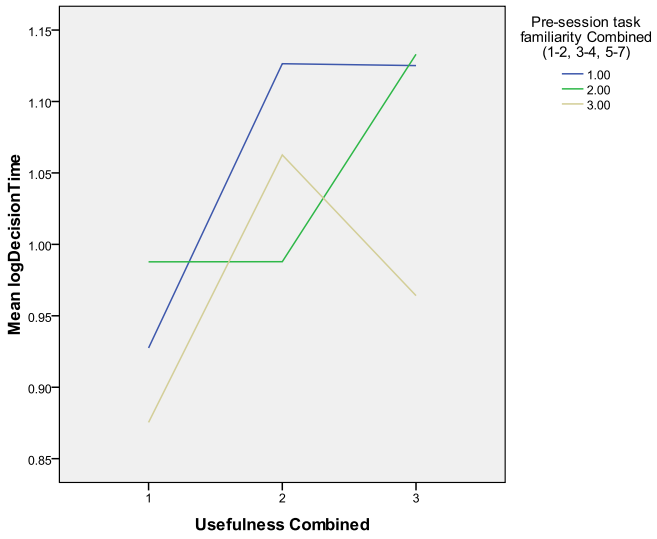


Figure 5.44 Relations between usefulness, topic knowledge, and log(10) of decision time in the dependent task

As what was found when total display time or total dwell time was used, usefulness had a significant main effect on decision time, $F(2, 540)=3.312$, $p<.05$. 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 other words, there was not a 3-way relationship found between decision time, usefulness, and topic knowledge.

These results indicate that decision time could be interpreted as a reliable indicator of document usefulness, i.e., the longer users took to make a decision on a document, the more likely the documents were to be useful. Mean decision time of the documents viewed by the users with different levels of knowledge was similar. In other words, no matter how much knowledge the users had, they took an equally long time to make a decision about the documents' usefulness.

5.5.4 Sub-question 2d

Sub-question 2d: In the parallel task, does the user's topic knowledge help in interpreting time as an indicator of document usefulness?

5.5.4.1 Total display time

The following reports findings on whether topic knowledge helps in interpreting total display time as an indicator of document usefulness in the parallel task. Again, usefulness and pre-session task topic knowledge were both collapsed into three groups. For usefulness, the same way was used to combine original scores as what was done in Section 5.4 for the parallel task, i.e., 1-2 into a little useful group, 3-5 into a medium useful group, and 6-7 into a very useful group. To refresh, Figure 5.46 and Figure 5.45 shows the distributions before and after combining. In the following of Section 5.5.4, unless specified, usefulness refers to the combined usefulness.

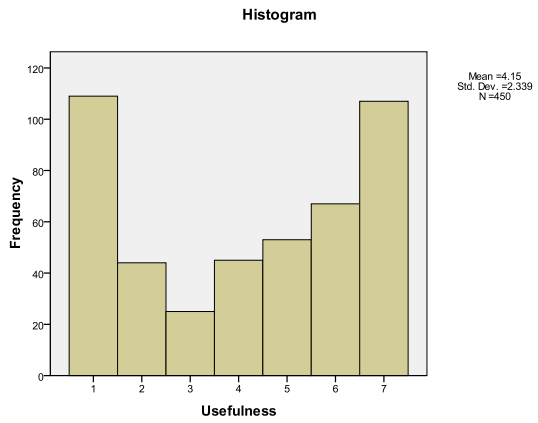


Figure 5.46 Distribution of the original usefulness data in the parallel task

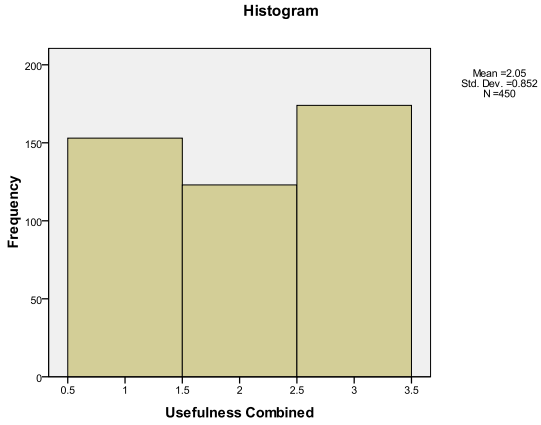


Figure 5.45 Distribution of the combined usefulness data in the parallel task

For pre-session task topic knowledge, the distribution of the original data in the parallel task is depicted by Figure 5.48. The combination was again in the following way: 1-2 into a low-knowledge group, 3-5 into a mid-knowledge group, and 6-7 into a high-knowledge group. Figure 5.47 shows the distribution after combination. In Section 5.5.4, unless specified, topic knowledge refers to the combined pre-session task topic knowledge.

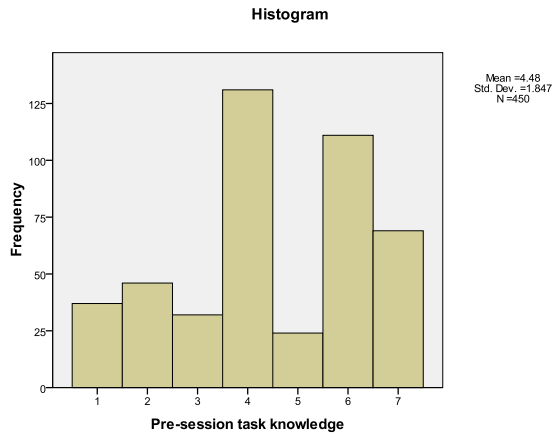


Figure 5.48 Distribution of the original topic knowledge data in the parallel task

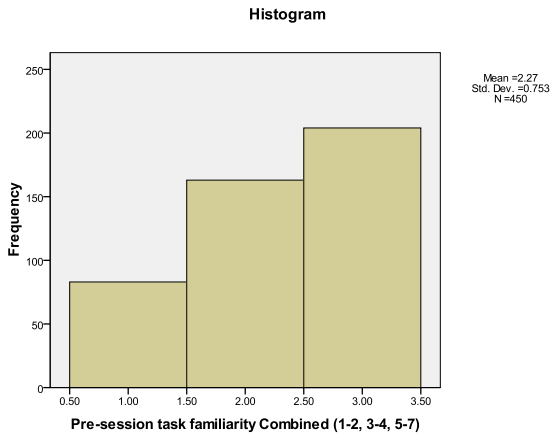


Figure 5.47 Distribution of the combined topic knowledge data in the parallel task

Table 5.21 GLM Univariate results of the effects of usefulness, topic knowledge, and their interaction on mean log(10) of total display time in the parallel task

Effect source (factor or interaction)	F	p
Usefulness	65.704	.000
Topic knowledge	.895	.410
Usefulness*Topic knowledge	3.938	.004

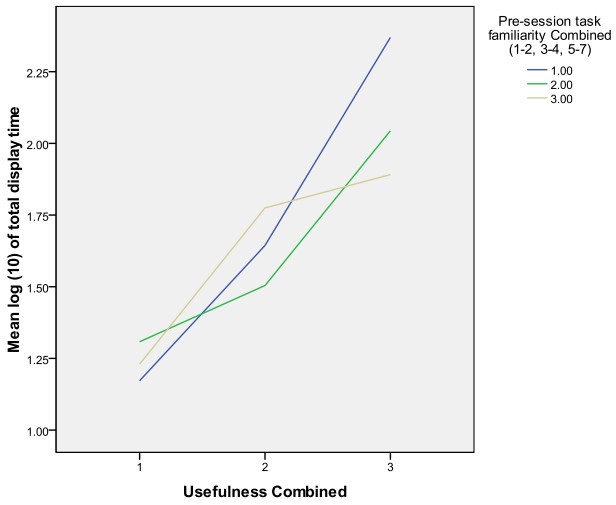


Figure 5.49 Relations between usefulness, topic knowledge, and log(10) of total display time in the parallel task

As Table 5.21 shows, usefulness was found to have a significant main effect on total display time, $F(2, 447)=65.704$, $p<.001$, meaning that there was a significant relationship

between usefulness and total display time. Topic knowledge did not have a significant main effect on total display time. The interaction between usefulness and knowledge had a significant effect on total display time, $F(4, 445)=3.938$, $p<.005$, meaning that the relation between usefulness and total display time varied across different levels of topic knowledge.

The results indicate that in general, for all users (without differentiating their knowledge levels), the longer the documents were displayed, the more useful the documents were. For all documents (without considering their usefulness degrees), average display time of users with different topic knowledge was not different. However, for documents with different levels of usefulness, users with different levels of knowledge had different patterns in the length of how long their viewed documents were displayed. Although users with different degrees of topic knowledge did not show differences in total display time for little useful documents, there were differences in total display time of somewhat and very useful documents between users with different degrees of topic knowledge. Specifically, for documents which were somewhat useful, those users who had much topic knowledge had them displayed longer than those who had little topic knowledge, and than those who had some topic knowledge. For documents which were very useful, those with much knowledge had them displayed more briefly than those with some knowledge, and than those with little knowledge.

5.5.4.2 Total dwell time

This subsection shows the findings on whether topic knowledge helps in interpreting total dwell time as an indicator of document usefulness in the parallel task. Table 5.22 and Figure 5.50 show the GLM Univariate results when usefulness and topic knowledge were considered as factors and log(10) of total dwell time were used as the DV.

Table 5.22 GLM Univariate results of the effects of usefulness, topic knowledge, and their interaction on mean log(10) of total dwell time in the parallel task

Effect source (factor or interaction)	F	p
Usefulness	51.787	.000
Topic knowledge	1.478	.229
Usefulness*Topic knowledge	4.688	.001

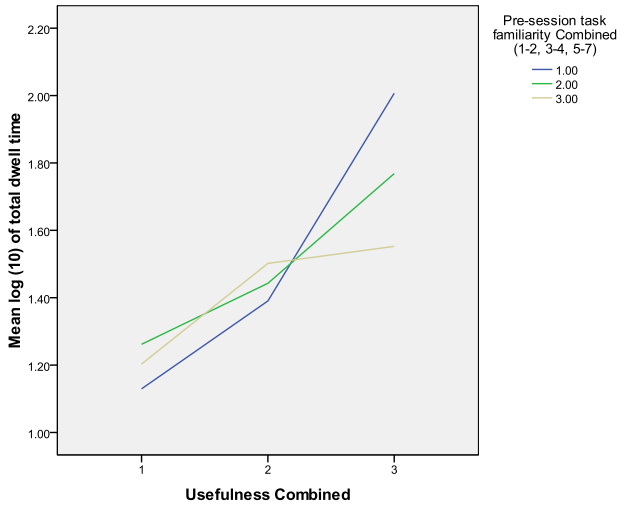


Figure 5.50 Relations between usefulness, topic knowledge, and log(10) of total dwell time in the parallel task

Usefulness was found to have a significant main effect on total dwell time, $F(2, 447)=51.787$, $p<.001$, meaning that the relation between usefulness and total dwell time was significant. Topic knowledge did not show a significant main effect on total dwell time. The interaction between usefulness and knowledge was found to have significant effect on time, $F(4, 445)=4.688$, $p<.005$, meaning that the relation between usefulness and total dwell time varied across different levels of topic knowledge.

Results indicate that in general, for all users (without differentiating their knowledge levels), the longer they dwelled on the documents, the more useful the documents were. For all documents (without considering their usefulness degrees), average total dwell time of users with different topic knowledge did not have differences. However, for documents with different levels

of usefulness, users with different levels of knowledge had different patterns in the length of how long they dwelled on the documents. Although users with different degrees of topic knowledge did not seem to have differences in terms of how long they dwelled on the little and somewhat useful documents, they did seem to differ on very useful documents. Specifically, for documents which were very useful, those with much knowledge had shorter dwell time than those with some knowledge, and than those with little knowledge, who spent a longer time dwelling on the documents.

5.5.4.3 Decision time

This subsection describes the findings on whether topic knowledge helps in interpreting decision time as an indicator of document usefulness in the parallel task. Table 5.23 and Figure 5.51 show the GLM Univariate results when usefulness and topic knowledge were considered as factors and log(10) of decision time were used as the DV.

Table 5.23 GLM Univariate results of the effects of usefulness, topic knowledge, and their interaction on mean log(10) of decision time in the parallel task

Effect source (factor or interaction)	F	p
Usefulness	.880	.415
Topic knowledge	4.666	.010
Usefulness*Topic knowledge	2.302	.058

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, $F(2, 447)=4.666$, p=.01. The interaction effect between usefulness and knowledge on decision time was marginally significant, $F(4, 445)=2.302$, p=.058.

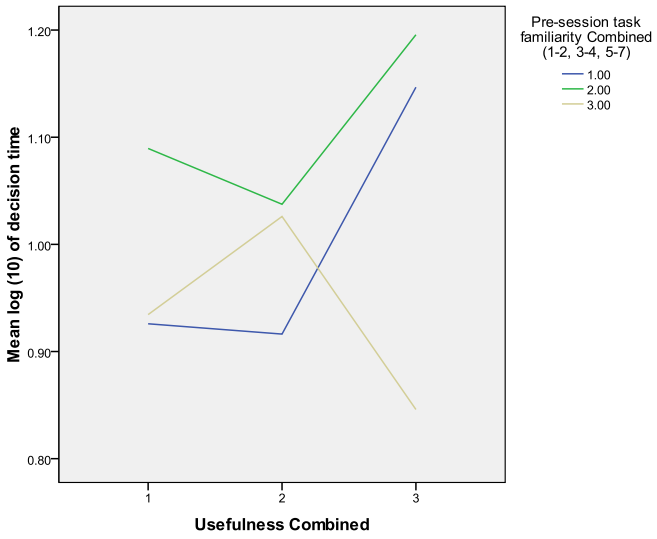


Figure 5.51 Relations between usefulness, topic knowledge, and log(10) of decision time in the parallel task

The results indicate that in general, all users (without considering their topic knowledge levels) seemed to have equally quickly decided the usefulness of retrieved documents which had different levels of usefulness. However, users with different levels of knowledge spent different lengths of time on judging the usefulness of documents. Specifically, those who had much knowledge made the decisions more quickly than those with only little knowledge. For documents with different levels of usefulness, users with different levels of knowledge also had different decision time patterns. As Figure 5.51 shows, for the little useful documents, those users who had little knowledge and much knowledge had less decision time compared with those who had mediocre knowledge. For the somewhat useful document, those with little knowledge made the decision more quickly than those with at least some knowledge. For the very useful documents, those who had much knowledge made the usefulness judgment more quickly than those had less than some knowledge on the task topics.

When looking at the users with different levels of knowledge, it was found that users with little topic knowledge quickly left the documents if they were of little use, and they spent a (relatively) very long time on the page before leaving, if it was very useful to them. Users with some topic knowledge spent longer time than those had little knowledge to determine a document was little useful, and they spent a bit less time to determine if a document was somewhat useful, but much longer time when a document was very useful. Meanwhile, those with much topic knowledge appeared to have judged as quickly as those with little knowledge on little useful document, and they judged a bit slower if on somewhat useful documents, but much more quickly on very useful documents.

5.5.5 Summary of the above results

Table 5.24 shows the summary of the main effect of usefulness, topic knowledge, and the interaction effect of topic knowledge and usefulness. Data were F and p (in parenthesis) values obtained in GLM analyses. Again, Usefulness and pre-session task topic knowledge both used combined data.

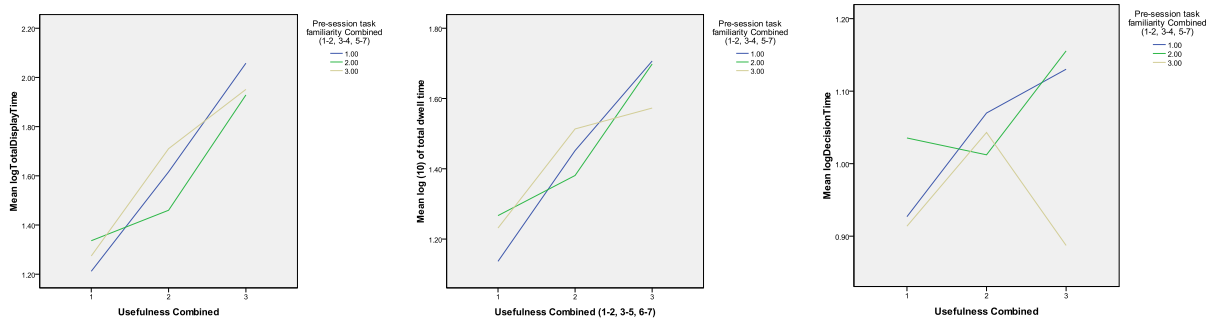
Table 5.24 Summary of the *F*(p) values of factors (results of GLM analyses)

Task	Type of time	Topic knowledge	Usefulness	Topic knowledge*Usefulness
In both tasks	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)

Figure 5.52 summarizes the figures again showing the relationships among time, topic knowledge, and usefulness:

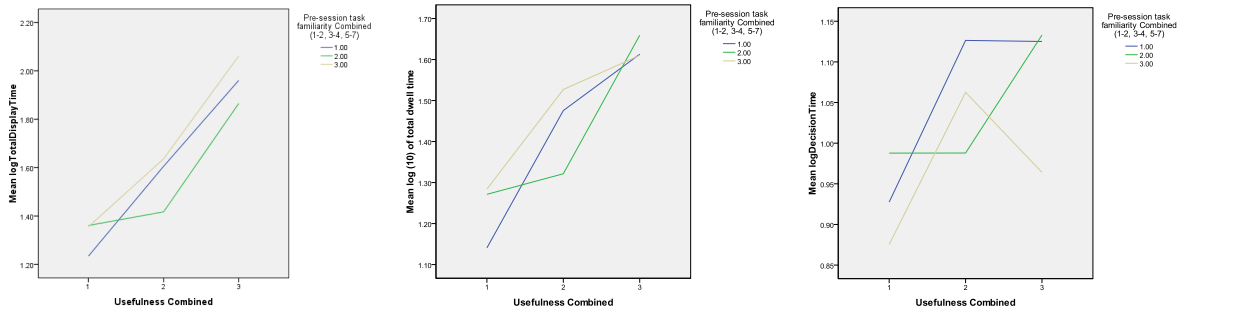
In both tasks

Log(10) total display time Log(10) total dwell time Log(10) decision time



In the dependent task:

Log(10) total display time Log(10) total dwell time Log(10) decision time



In the parallel task:

Log(10) total display time Log(10) total dwell time Log(10) decision time

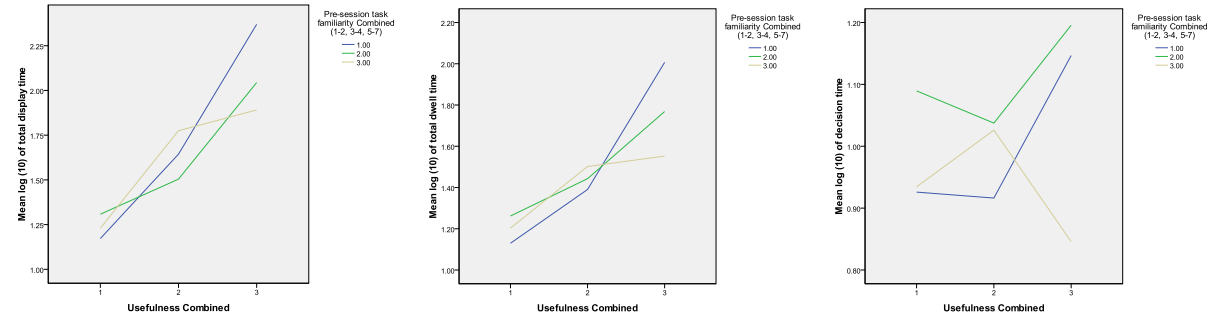


Figure 5.52 Relations between time, topic knowledge, and usefulness

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, 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 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 will help interpreting these times as indicators of usefulness.

5.5.6 Sub-question 2e

Sub-question 2e: How does topic knowledge compare with stage in terms of its role in helping interpret decision time as an indicator of document usefulness?

This section compares the two factors: task stage and topic knowledge, in terms of their roles in help interpreting time as an indicator of document usefulness. The differences will be examined, and more GLM analyses including both factors will be conducted to confirm the results. Since both topic knowledge and stage showed effects only when decision time was considered but not the other two types of time, analysis for this sub-RQ2e only looks at decision time.

5.5.6.1 Difference examination

Figure 5.53 and Figure 5.54 show the relations between task stage, usefulness, and decision time in both tasks and in the parallel task. Figure 5.55 and Figure 5.56 show the relations between topic knowledge, usefulness, and decision time in both tasks considered together and in the parallel task.

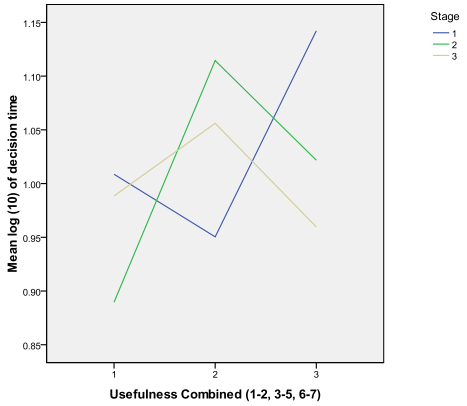


Figure 5.53 Relations between usefulness, stage, and decision time in both tasks

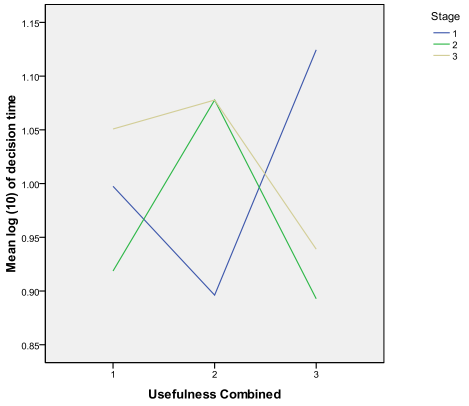


Figure 5.54 Relations between usefulness, stage, and decision time in the parallel task

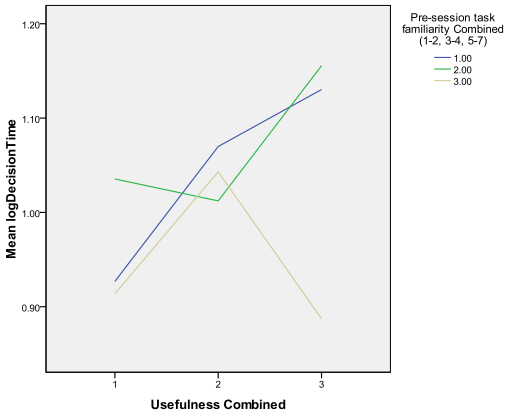


Figure 5.55 Relations between usefulness, topic knowledge and decision time in both tasks

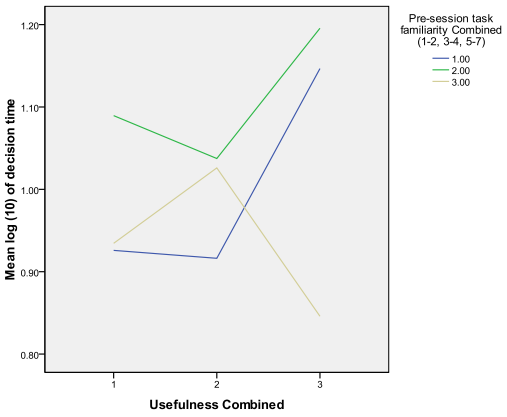


Figure 5.56 Relations between usefulness, topic knowledge and decision time in the parallel task

When considering both tasks together, for little useful documents, users in stage 2 had shorter decision times but those in stage 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 spent a long time to make the decision, just as people in stage 1 or stage 3.

For somewhat 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 very 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. Meanwhile, those with some or little knowledge spent a long time to make a usefulness decision, but those with much knowledge spent very short time to make a usefulness decision. This seems to indicate that in making a usefulness judgment when it was actually very 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 spent 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 little 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 & 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 spent a long time to make the decision, as people in stage 3.

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

For very useful documents, in stage 1, users had very long decision time, but in stages 2 and 3, they had short decision time. Users with only some or even little knowledge had long decision time, but those with much knowledge had short decision time. This seems to indicate a correspondence between stage 1 and knowledge levels 1 & 2 (long decision time), and stages 2 & 3 and knowledge levels 3 (short decision time). This seems to indicate that in making a usefulness judgment when it was actually very 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 spent a long time to make the decision, as people in stage 1.

5.5.6.2 More GLM analyses on various factors

Further analysis using the GLM model was conducted to confirm these findings and to compare the factors considered in RQ2, i.e., knowledge, with that in RQ1, i.e., 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.

Table 5.25 GLM results when both task stage and topic knowledge was considered

Tests of Between-Subjects Effects

Dependent Variable:decision time in seconds

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	54591.722[a]	26	2099.682	3.322	.000	.082	86.379	1.000
Intercept	204797.010	1	204797.010	324.044	.000	.251	324.044	1.000
usefulnessCombined	3488.049	2	1744.025	2.760	.064	.006	5.519	.545
Stage	246.330	2	123.165	.195	.823	.000	.390	.080
preTaskFamCom	2472.378	2	1236.189	1.956	.142	.004	3.912	.406
usefulnessCombined * Stage	4525.569	4	1131.392	1.790	.129	.007	7.161	.548
usefulnessCombined * preTaskFamCom	6130.774	4	1532.693	2.425	.047	.010	9.701	.698
Stage * preTaskFamCom	2378.318	4	594.580	.941	.440	.004	3.763	.301
usefulnessCombined * Stage * preTaskFamCom	9279.643	8	1159.955	1.835	.067	.015	14.683	.785
Error	610515.821	966	632.004					
Total	1046104.250	993						
Corrected Total	665107.543	992						

a. R Squared = .082 (Adjusted R Squared = .057)

b. Computed using alpha = .05

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 pre-session task familiarity played a more significant role than task stage in interpreting decision time as an indicator of usefulness.

5.5.7 Answers to RQ2 and sub-RQs

Based on the analysis, RQ 2 is answered as follows:

RQ2: Does the user's topic knowledge help in interpreting time as an indicator of document usefulness?

Based on the findings in the current study, the brief answer to RQ2 is: yes, when pre-session general task topic knowledge is considered, and when decision time is used.

Sub-question 2a: What are the patterns of the users' pre- and post-session general task topic knowledge and pre- and post-session sub-task topic knowledge?

The answer to this sub-question is: post-session topic knowledge was higher than pre-session knowledge for both general tasks and sub-tasks. For general tasks, pre-session knowledge in later stages was higher than that in previous stages, so was post-session knowledge. However, for sub-tasks, pre-session knowledge in later stages was not higher than that in previous stages, although post-session knowledge in later stages was still higher than that in previous stages, meaning that users perhaps learned more on sub-tasks in later stages.

Sub-question 2b: In general, i.e., in both the parallel and the dependent tasks, does the user's topic knowledge help in interpreting time as an indicator of document usefulness?

The answer to this sub-question is: yes, when pre-session task topic knowledge was used to help interpret decision time as an indicator of document usefulness (a significant 3-way relationship was found among pre-session task topic knowledge, usefulness, and decision time, i.e., there was a significant interaction effect between pre-session task topic knowledge and usefulness on decision time); but no, pre-session task topic knowledge was not found to help in interpreting total display time or total dwell time as an indicator of document usefulness (no significant 3-way relationship was found among pre-session task topic knowledge, usefulness, and total display time; or among pre-session task topic knowledge, usefulness, and total dwell time).

Sub-question 2c: In the dependent task, does the user's topic knowledge help in interpreting time as an indicator of document usefulness?

The answer to this sub-question is: no, pre-session task topic knowledge was not found to help in interpreting any type of time (total display time, total dwell time, or decision time) as an

indicator of document usefulness (no significant 3-way relationship was found among pre-session task topic knowledge, usefulness, and total display time; or among pre-session task topic knowledge, usefulness, and total dwell time; or among pre-session task topic knowledge, usefulness, and decision time).

Sub-question 2d: In the parallel task, does the user's topic knowledge help in interpreting time as an indicator of document usefulness?

The answer to this sub-question is: yes, when pre-session task topic knowledge was used to help interpret decision time as an indicator of document usefulness (a significant 3-way relationship was found among pre-session task topic knowledge, usefulness, and decision time, i.e., there was a significant interaction effect between pre-session task topic knowledge and usefulness on decision time); but no, pre-session task topic knowledge was not found to help in interpreting total display time or total dwell time as an indicator of document usefulness (no significant 3-way relationship was found among pre-session task topic knowledge, usefulness, and total display time; or among pre-session task topic knowledge, usefulness, and total dwell time).

Sub-question 2e: How does topic knowledge compare with stage in terms of their roles in helping interpret time as an indicator of document usefulness?

The answer to this sub-question is: pre-session general task topic knowledge was found to play a more significant role than stage in helping to interpret total display time as an indicator of document usefulness.

5.6 Results of RQ3

RQ3 was aimed at comparing the performance of a system which displayed useful terms extracted from users' past work helps their search in later stages, with one which did not. One

version of the system was the blank IE window and the other had a query-suggestion interface with a side bar listing suggested terms extracted from previous session(s). In the process of data analysis, a number of measurements, such as time spent to find the first useful document, the mean ranking of useful documents, etc., were used to compare the differences between the two groups of people who used the two versions of systems. No significant results were found. It was then detected that the two groups of people had different initial topic knowledge of the tasks. Specifically, there were significant differences in their knowledge with the general task (combined pre-session task topic knowledge, the same variable as is used in the previous two RQs) in the beginning of stage 1 (Mann–Whitney $U = 39.0$, $n1 = n2 = 12$, $p < 0.05$ two-tailed). The participants were unfortunately not from the same sample. Therefore, RQ3 cannot be addressed.

However, it should be noted that this does not affect the analysis of RQs 1 and 2 because those two RQs were not comparing two groups of people on a task level, but rather comparing time and usefulness on a document level.

Chapter 6. Discussion

This chapter discusses the main findings of the study. It begins with a discussion of the three types of time used in the study with respect to how they were measured and how they can be used in modeling users and personalizing search. Following it are discussions of each research question about the major findings, potential explanations, and possible implications. This chapter ends with a discussion of the relationships between the examined factors, as well as the implication of the findings on the theoretical model proposed in Chapter 3.

6.1 Three Types of Time and Their Usefulness in Modeling Users and Personalizing Search

It was shown in this study that there were different types of time that can be used to measure the time that users spend on documents. The study identified and used three types of time: total display time, total dwell time, and decision time, while previous studies tended to have a single type of time, being dwell time or total dwell time. One exception in previous studies was Kelly's (2004) naturalistic and longitudinal study, in which some of the participants did perform a certain task other than search for information, for example, writing an email which included information that were searched in parallel with writing the email. Nevertheless, these cases happened only occasionally in Kelly (2004) study, and these cases were not differentiated with others that did not involve performing a task with a task product. Therefore, even though the time in Kelly (2004) was defined similarly as the total dwell time in the current study, the results of these two studies cannot be compared in a simple manner.

Among the three types of time, total display time and total dwell time were both measured at a whole session level. This means that they will not be captured until a session is

finished. While in a multi-session 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 it can be used for adapting search for the current session, in addition to adapting search in the following sessions. The findings of the current study with the three types of time are discussed in the following.

6.2 Time, task stage, and document usefulness

RQ1 of this study was concerned with exploring whether task stage helps in interpreting time as an indicator of document usefulness. This was addressed by examining the relationships between task stage, usefulness, and time in GLM analyses. All three types of time were considered. The relationships were examined in both task types in general as well as in the dependent and the parallel tasks individually. The following discusses each factor of time, stage, and task type regarding their roles in helping interpret time as an indicator of document usefulness. Possible explanations and implications are also discussed.

6.2.1 Time as an indicator of Usefulness in the Stage Model

This part discusses the use of time as an indicator of usefulness when stage is considered in the GLM model, simplified as the stage model. Results show that when both tasks combined were considered, in GLM analysis, the relations between usefulness and total display time, as well as that between usefulness and total dwell time, were significant. This means that those documents which 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 total dwell time and total display time of those documents which were more useful were therefore increased. These

findings indicate that in both tasks combined, i.e., when task type was not specified, total display time and total dwell time were rather reliable indicators of document usefulness.

However, decision time was found not to have a significant main relationship with usefulness. Users did not necessarily spend more time on making decisions of documents of different levels of usefulness, when no distinction between tasks was made. This indicates that when task type was not specified, decision time alone cannot be a reliable indicator of document usefulness.

In the dependent task, when any type of time (total display time, total dwell time, or decision time) was used in the GLM model, usefulness was found to have a significant relationship with time. This means that all three types of time appeared to be a reliable indicator 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 it 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 the users 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 in the dependent task 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 point 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, as described in the preceding paragraph, 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 point 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.

The differences in findings in the two types of tasks, as well as those in both tasks combined, also have some implications for the usefulness of task type information[9]. First, when decision time is used to personalize search, it is important to know the task type in order to obtain better personalization performance because the findings for decision time in the two individual task types were different, and the finding for decision time when both tasks were combined was different from that in the dependent task. If no task type information was specified, the use of decision time would follow the findings in both tasks combined, i.e., decision time only cannot be a reliable indicator of usefulness. If the task type is parallel, the same pattern follows. However, if it is known that the task type is dependent, then decision time could be used

[9] More discussion about task type is available in Section 6.2.3 when stage information is also included. This section here about the differences in task type focuses on time as a single predictor of usefulness.

as a reliable indicator of usefulness. Second, when total dwell time or total display time is used, task type information is not important since findings for these two types of time, with respect to their relation with usefulness, in the two individual task types were the same, and that is also consistent with the findings when both tasks were combined. In short, task type information is important when decision time is used alone as an indicator of usefulness, but not when total dwell time or total display time is used as a single indicator of usefulness.

To sum up this sub-section on the use of time as an indicator of usefulness, 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 only as a single indicator of usefulness only worked in the dependent task. Given the above mentioned 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 6.1 shows a summary of the indicators of usefulness.

Table 6.1 Summary of indicators of usefulness

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

6.2.2 Stage as a Helpful Contextual Factor in Inferring Usefulness

As can be seen from **Error! Reference source not found.**, 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, stage showed a significant interaction effect with usefulness on decision time, where usefulness only did not have a significant relation with decision time. This means that in the parallel task, decision time alone cannot be a reliable indicator of document usefulness, however, when task stage information was also considered, it 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, sub-task topics changed across stages, but sub-task patterns did not, and users were dealing with roughly the same things (e.g., exterior and interior feature, performance, safety, prices, and colors) on different car models. Users were very likely to have gained some knowledge on sub-tasks, 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 indicative value of any of the three types of time. Explanations of why stage did not play a role in the dependent task could possibly be that in the dependent task, not only were sub-task topics different, but sub-task 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 different things at the three sessions, and they most likely looked at different web pages on different web sites. Users would not have gathered knowledge over stages that may change the time that they spent 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).

The different findings in terms of the role of task stage in helping indicate document usefulness in two individual tasks as well as in both tasks combined finds that it is important to

know task type information[10]. First, when decision time is used for personalization, knowledge of task type is important because task stage played a role in the parallel task but not in the dependent task. If task type is not specified, stage is shown to play roles, which does not hold true in the dependent task. Second, when total display time is used for personalization, task type information is also important. Although task stage did not seem to play a role in individual tasks, it did in both tasks combined[11]. If no task information is specified, interpretation of task stage's role will be different. (Note: When total dwell time is used, stage did not play any roles, so it is not discussed here.)

In sum, task stage was found 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 in interpreting total display time as an indicator of usefulness. This finding can be used for the subsequent search/work sessions, 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 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 be helpful for personalizing search for specific users in that decision time can be a reliable indicator of document usefulness given the task stage (and task type) information. In addition, unlike the total dwell time which would not be available until a session is over, decision time can be easily obtained since it is the time duration between opening the document till the user's first action, which can be easily captured by the system.

[10] This is a re-statement from the stage perspective of the differences among task types, as described in Section 6.2.1.
[11] The reason of this finding could be due to the sample size. When individual tasks were considered separately, the sample size is not as big as when they were combined. The tendency that when sample size gets bigger, the role of stage became more salient also reminds us to pay attention to this contextual factor.

6.2.3 Task Type as a Helpful Contextual Factor in Inferring Usefulness

The importance of task type has been addressed in Sections 6.2.1 and 6.2.2 from the time and the stage perspectives. It is discussed here as a single subsection to summarize the findings on task type.

Generally speaking, the most important finding is that when interpreting decision time as an indicator of document usefulness so that it may help personalize search for the current session, 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, sub-tasks that the users worked with in the three sessions were different not only in their topics, but also in the sub-task patterns. However, in the parallel task, sub-tasks that the users worked with in the three sessions were different only in their topics; they had the same sub-task 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. This made it possible that users gained knowledge across stages in the parallel task on the usefulness of some 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, i.e., 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, i.e., the overall correctness of usefulness prediction.

What also needs to be mentioned is that when total display time is 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 (the possible reason that sample size may have influenced this result has been mentioned previously in a footnote). 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.

6.3 Time, topic knowledge, and usefulness

RQ2 of this study was concerned with exploring whether users' knowledge of task topics helps interpret time as an indicator of document usefulness. This was addressed by examining the relationships among topic knowledge, usefulness, and time in GLM analyses. All three types of time were considered. The relationships were examined in general in both task types combined as well as in the dependent and the parallel tasks individually. Since there were different types of knowledge elicited, their patterns are first discussed. This is followed by discussions on each factor of time, topic knowledge, and task type with respect to their roles in inferring document usefulness. Possible explanations and implications are discussed.

6.3.1 Topic knowledge patterns across 3 stages

There were four types of topic knowledge elicited in the study: pre-session general task topic knowledge, pre-session sub-task topic knowledge, post-session general task topic knowledge, and post-session sub-task topic knowledge. General task topic knowledge and sub-task topic knowledge are discussed separately.

6.3.1.1 General task topic knowledge patterns

General task topic knowledge in both the parallel and the dependent task measured the same thing across 3 stages, which is the users' familiarity with the general task. Recall that this knowledge of users was found to have increased in the experiment from an average rating score of less than 3 (out of 7) in the beginning of the first session to an average rating score of above 5

(out of 7) by the end of the third session. Not only did it increase in each session/stage, i.e., the self-rated post-session knowledge level was higher than the pre-session one, but it also increased across stages, i.e., knowledge in later stage(s) was higher than that in the earlier stage(s). It was also found that pre-session knowledge in stage 2 was a bit lower than post-session knowledge in stage 1, although pre-session knowledge in sage 3 was roughly the same as post-session knowledge in stage 2.

These findings demonstrate that users did learn through the experiment. It is reasonable to see that they learned through searching and writing in each session, and that they forgot a bit after an interval after session 1, when their knowledge was only mediocre (a rating of about 4 out of 7). After stage 2, their average knowledge was pretty high (rating of 5 out of 7), and their knowledge did not reduce as much as it did after stage 1.

When looking at user knowledge of the different task types (the dependent and the parallel), it was found that these patterns roughly apply to both types of tasks, although in the 3rd session, users still gained knowledge with the dependent task but not with the parallel task. Instead, their descriptive average self-rated knowledge level lowered a bit. In addition, user knowledge of the parallel task was higher than that of the dependent task, especially in the first 2 sessions. In the beginning of the 3rd session, users showed higher knowledge in the parallel task, but by the end of the 3rd session, their knowledge levels in both tasks combined were roughly the same.

The differences between knowledge on the two tasks could possibly be explained by the reason that the three car models in the parallel tasks may be more familiar to the users than the three activities in the dependent tasks, for which they did not know as much until after they searched on them. It should be noted that the differences here, we think, reflect only the

differences of users' knowledge on the task topics but should not be interpreted as to affect the relations among topic knowledge, time, and usefulness (in other words, the roles that topic knowledge play in inferring document usefulness) in the two types of tasks.

6.3.1.2 Sub-task topic knowledge patterns

About users' knowledge on sub-task topics, it was found that this knowledge of users increased in each session/stage, and post-session sub-task topic knowledge increased across stages, but pre-session sub-task topic knowledge did not increase across stages. In other words, the self-rated post-session knowledge level was higher than the pre-session one, and post-session knowledge in later stage(s) was significantly higher than that in the earlier stage(s), but pre-session knowledge in later stage(s) was not significantly higher than that in the earlier stage(s). These results are reasonable. Users learned on each sub-task through searching and writing within the session. They did not have much higher pre-session knowledge on sub-task(s) in the later session(s) because this sub-task topic knowledge measured different things in the three stages, which was the users' familiarity with the different sub-tasks. Nevertheless, they had higher post-session knowledge on sub-task(s) in the later session(s), meaning that users perhaps learned more about the sub-tasks in later sessions.

When looking at users' sub-task topic knowledge in the different task types (the dependent and the parallel), it was found that the above described patterns in both tasks combined apply to both types of individual tasks. In other words, in either the dependent or the parallel task, users' knowledge on each of the three sub-tasks was roughly the same. This indicates that in the parallel task, users' knowledge on the car models that they have finished did not help them with other car models in terms of how familiar they were with the sub-task before searching and working on those car models; in the dependent task, the fact that the completion of

some sub-task(s) depended upon the completion of other sub-task(s) means that knowledge of the subsequent task was not enhanced.

In general, users' sub-task topic knowledge in the parallel task was higher than that in the dependent task. This difference could possibly be explained by the reason that each of the three car models in the parallel tasks may be more familiar to the users than each of the three activities in the dependent tasks, for which they did not know as much until after they searched on them.

Recall that although there were four types of topic knowledge, only pre-session task topic knowledge was used in data analysis since only this knowledge among the four reflects users' knowledge change across stages in multi-session tasks, and it may be obtained a-priori (possibly through monitoring users' past search, etc.) in order to be used for modeling user behaviors and adapting search. In the following discussion, topic knowledge refers specifically to the pre-session general task topic knowledge.

6.3.2 Time as an Indicator of Usefulness in the Topic Knowledge Model

This section discusses the use of time as an indicator of usefulness when topic knowledge is considered in the GLM model, simplified as topic knowledge model. In general, the results found all three types of time as indicators of usefulness in the topic knowledge model were consistent with those in the stage model.

Specifically, results in the topic knowledge model showed that in the dependent task, each of the three types of time: total display time, total dwell time, and decision time, appeared to be reliable indicators of document usefulness without consideration of other factors such as topic knowledge. The longer a page was displayed and/or viewed, the more useful the page was. It is reasonable to see this since the user was working in parallel with writing while searching for and reading Web pages, and those pages which were displayed or viewed while writing were

most likely to be useful ones. In addition, the longer it took the user to judge the usefulness of a page in the process of working with their tasks, the more useful it was. This meant that for more useful pages, users needed to read them longer (than little useful pages) before going to other documents or starting to write in MS WORD, etc.

Given the limitation of total dwell time and total display time that they cannot be captured until the end of a session, their use as a reliable indicator of usefulness in the dependent task can only be applied to subsequent sessions in a multi-session task, but not in the current session. On the other hand, decision time can be used as a reliable indicator of usefulness in the dependent task for personalizing search in the current session, as well as in the following sessions.

By contrast, in the parallel task or when task type was not specified (i.e., in both tasks combined), total display time and total dwell time still were shown to be reliable indicators of document usefulness, but decision time was not[12]. This means that in order for the interpretation based on decision time to be accurate, other factors such as topic knowledge must be taken into account. Possible explanations of topic knowledge are discussed in the next section (Section 6.3.3).

In both tasks combined, i.e., when the task type information was not specified, findings were similar to those in the parallel task. This means that the parallel task had so great an impact on the combined tasks, that the findings still hold true even when the dependent task was also included in analysis.

[12] The findings on relationships between time and usefulness here in RQ2 when topic knowledge was considered are slightly different from relationships between time and usefulness in RQ1 when task stage was considered. It is reasonable to see such a difference because the total factors considered in GLM analyses were different. These findings should not be understood as inconsistency with each other, but they were caused when different other factors were considered in a model. Nevertheless, when only time and usefulness were considered, they should still have significant correlation. More discussion on the effect when both task stage and topic knowledge were considered in the model is available in the later section of this chapter (Section 6.3.5.3).

Concerning the differences of task type, it can be seen from the above discussion that task type information is very important. For one thing, results in the parallel and in the dependent tasks were quite different. For another, results in both tasks combined were different from those in the dependent task, too. If no task information was specified, and if a task is indeed dependent but the system does not know, interpretation of such a task would follow the patterns in both tasks combined, that decision time is not a reliable indicator of usefulness. This would be inaccurate, because in fact, all times should be reliable indicators of usefulness.

Table 6.2 is a summary of the indicators of document usefulness when topic knowledge was considered in the model.

Table 6.2 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

6.3.3 Topic Knowledge as a Helpful Contextual Factor in Inferring Usefulness

Although in the dependent task, topic knowledge did not seem to play any role in interpreting any of the three types of time as an indicator of usefulness, in the parallel task and in both tasks combined, topic knowledge was found to have played a significant role since significant interaction effects between topic knowledge and usefulness were found on time in GLM analysis. Specifically, topic knowledge was found to have a significant interaction effect with usefulness on total display time and total dwell time (where usefulness was also found to have a main effect). Topic knowledge was also found to have significant interaction effect with usefulness on decision time, where usefulness did not show any main effect.

In the parallel task, for very 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 very 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 point 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 indicated 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 the ongoing session. Nevertheless, 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. The significant role of topic knowledge when decision time is used for inferring usefulness in both tasks combined and in the parallel task was extremely important because in these cases, decision time only cannot be used as an indicator of usefulness (usefulness did not have main effect on decision time). In sum, these findings can be used for personalizing search for specific users with different topic knowledge.

6.3.4 Task Type as a Helpful Contextual Factor in Inferring Usefulness

As described above, our findings are that the roles that topic knowledge played in the dependent task and in the parallel task were different. In the dependent task, topic knowledge was not found to play any role in inferring document usefulness, but in the parallel task or when task type was not specified, it did. This was true for all three types of time.

Possible explanation of this could be based on the differences in the nature of the two types of tasks. The sub-tasks in the dependent task varied not only on the topics but also on the patterns, and users were asked to focus on different things in the different sub-tasks. On the other hand, the sub-tasks in the parallel task varied on the topics but not on the patterns, and users were asked to work on the same or similar things (such as write on exterior and interior features, performance, price, color, etc.) of different car models. Therefore, in the dependent task, for users who had much topic knowledge, it is very likely that they had this knowledge before they worked, not through working, with this task at hand. Therefore, they may have known little about the layout, format, or possible outline of a document when they first opened a page in a session, since this could be the first time the user sees such a page or a similar page in a domain/website for this whole task. In other words, their knowledge of the task topic would not have helped them much on being familiar with the web pages. That the users (even those had

much topic knowledge) did not have much knowledge on the document itself would then cause them to spend a longer time in getting familiar with very useful documents and the useful information in them before leaving these documents and starting to use them in writing (prolonged decision time). This reason would also possibly increase the chances that users kept referring back to the documents in parallel with writing the reports (prolonged total dwell time), and the chances that users kept these useful documents open in writing (prolonged total display time).

On the other hand, in a parallel task, for users who had much topic knowledge, it is possible that they had increased their knowledge on the topic through working with the task at hand. In addition, they may also have gained some knowledge of the layout or possible outline of a document in a domain/website that they had seen in the earlier phases/sessions of this task, which should be helpful for them to make a decision more quickly on useful documents (decision time), as well as shorten the reading time when referring back to these documents in parallel with writing the reports (total dwell time), and the chances and time that users kept these useful documents open in writing (total display time).

When there is no task information specified, i.e., in both tasks combined, topic knowledge also played a role. This is due to, we think, the strong role of topic knowledge in the parallel task. For a task which was a dependent task, absence of task type information will have generated inaccurate interpretations since interpretations according to findings in both tasks combined would be different from findings in the dependent task. In sum, it is important to know task type information in order to better infer document usefulness from time.

6.3.5 The relation between time and topic knowledge

Although it was not the main focus of the current study, the above results also included the two-way relationship between time and topic knowledge. It is discussed here in comparison with the previous findings of Kelly & Cool (2002) who examined the same relationship. Kelly & Cool (2002) found that with the increase of users' topic knowledge, users' reading time tended to decrease, although no statistically significant differences were found between topic familiarity (based on a 5-point scale) and reading time. The current study found that there were statistically significant differences in the time spent on documents between people with different levels of topic knowledge, and that this relation between time and topic knowledge was true only in certain circumstances. Specifically, significant relations between topic knowledge and decision time were found in both tasks combined, and in the parallel task. The patterns of these two relationships were the same: when users' knowledge was only little or medium, their decision time did not differ. However, when users' knowledge was much, the decision time decreased. These findings in the current study suggest certain conditions in which the findings in Kelly & Cool (2002) are applicable: when users' knowledge was medium and above, when decision time was considered, and when task type was the parallel or both parallel and dependent tasks combined. Findings of the current study extend the literature by discovering the conditions of interpreting the relationship between topic knowledge and reading time.

6.3.6 Comparison of the roles of task stage and topic knowledge in interpreting time as an indicator of usefulness

6.3.6.1 Both could help, in general

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 shown to be related to 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, it is reasonable to see that they both could help in general.

6.3.6.2 Examining their potential in detail

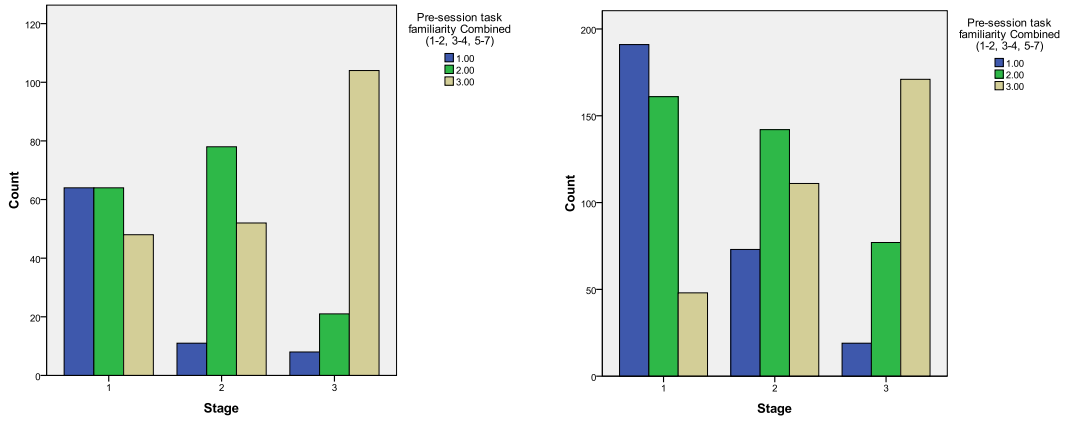
Although task stage and topic knowledge were both found to be potentially important in interpreting time as an indicator of document usefulness, the ways that they played roles were not always the same when considering the specific values of these two variables. In other words, it is not the case that stage 1 corresponded to topic knowledge level 1, stage 2 to topic knowledge level 2, and stage 3 to knowledge level 3. The following discusses their similarities and differences in more detail.

Table 6.3 Frequency of knowledge levels by task stages in the parallel task

Stage	Pre-session topic knowledge level		
	1	2	3
1	64	64	48
2	11	78	52
3	8	21	104

Table 6.4 Frequency of knowledge levels by task stages in both tasks combined

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



a. In the parallel task b. In both tasks combined

Figure 6.1 Frequency of knowledge levels by task stages

As results in Table 6.3 and

Table **6.4** 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 somewhat useful documents, and it dropped down for very useful documents to a similar or lower decision time for not useful documents. This could be explained by the point 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 6.3 and 6.4, and Figure 6.1.

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 corresponded so 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 was descriptively but not statistically significant higher than level 3 (N=48) (Table 6.3 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 was not exactly the same match by values.

When task type was not specified (i.e., in both tasks combined), 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.

6.3.6.3 More comparison of their roles

The result when both task stage and topic knowledge were considered in GLM analyses (Section 5.5.6.2) indicates that task stage did not play any significant role any more, but topic knowledge did. This could be interpreted as that topic knowledge could play a more significant and maybe more accurate role than task stage in helping infer usefulness. If both users' topic knowledge and task stage information are 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 of the single factor.

6.3.7 Implications of these findings

6.3.7.1 Implications on designing systems that can predict document usefulness

6.3.7.1.1 Application strategies

One direction of implications of the findings in RQs 1 and 2 is that they can be used for personalization system design. In this type of application, time as a user behavior can be easily detected by the system, and stage or knowledge is supposed to be detectable, too[13]. It is hoped that task type can also be learned through some means[14]. Using these types of information, the system could make predictions on the usefulness of the documents, based on the findings in this study.

Results show that if the system does not consider the roles that task stage and topic knowledge play and predict usefulness based solely on time, the prediction would not always be accurate. Given the task stage and/or topic knowledge information, the system can enhance its performance of usefulness prediction.

If the system does not consider stage information, it would have one single threshold criterion across all stages for not useful, somewhat useful, and very useful documents. For example it may simply classify those with decision time of less than, say, 10 seconds as little useful documents, those with decision time of longer than, say, 11.2 seconds as very useful, and those in between as medium 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 one, the system would classify documents with decision time of

[13] Detection of stage or knowledge is explained later in this section.
[14] Possible ways of detecting task type is discussed later in this section.

less than, say,1.4 seconds as low useful documents, and those with decision time of longer than, say, 1.5 seconds as very useful documents. At stage 2, the thresholds are different. Maybe those with decision time of less than 12.6 seconds would be low useful, longer than 12.6 seconds would be medium useful, and those with decision time of 12.6 seconds would be very useful. At stage 3, the thresholds would again be different than the previous 2 stages. In general, based on the findings of the roles of stage, this approach to setting different thresholds at different stages is hoped to lead to better performance in predicting document usefulness. Further studies will attempt to discover the thresholds for different stages, making predictions, and generating the ROC curve that describe the prediction performance (both correctness and error rate).

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 decision time of less than, say, 10.5 seconds as low useful documents, and those with decision time of longer than 10.5 seconds as very useful documents. For those with medium level of knowledge, the thresholds will be different from those for people with little knowledge. Documents with decision time of less than, say, 1.6 seconds would be low useful, and longer than 1.6 seconds would be very useful. For those with high level of knowledge, the threshold will be different again. In general, this approach of setting different thresholds based on different levels if knowledge should enhance performance of usefulness prediction. Further 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 perfect to infer usefulness, for example, at stage 3, it is difficult to differentiate the very useful and not useful documents since the means of these two groups were roughly the same. However, the purpose of this dissertation was to explore the role of stage 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 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)[15]. 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 all promising findings.

6.3.7.1.2 Ways to detect stage, knowledge, and task type

As mentioned in the beginning of this section, in order for the findings to be applicable in personalization system design, the system will need to be able to know the information of task stage and/or 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 of working on the same task/project. For example, people usually build a folder in a certain directory for a specific task, and the time that a person builds this folder is probably the starting point of a task. If the user opens the same MS WORD file at other times, or the files under the

[15] To use the heuristic of writing in WORD files or not as the single criterion to infer document usefulness is also possible but is not as accurate. Although a web page being followed by MS WORD writing possibly means it is very useful, web page not being followed by MS WORD does not necessarily means it is not useful. Combining decision time and this heuristic will have better prediction performance.

same directory, then the system can track how many times the person has worked with the files before to estimate task stage.

For topic knowledge, similar approaches can also be used to estimate user's knowledge to be 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, Dumais, & Teevan (2009), that users frequently going to the medical database PubMed, etc., are likely to be domain experts, while those 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, using 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, & Belkin (2010) found that in parallel tasks, users tended to employ more frequently a query reformulation strategy called Word Substitution, i.e., to substitute part 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 of all these contextual factors is detectable by the system.

6.3.7.2 Predicting stage, topic knowledge, or task type

Different from the implication introduced in Section 6.3.6.1, another way that the findings in this study can be applied 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, i.e., 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 with 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 in 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.

6.4 How Findings Relate to the Theoretical Model

In Chapter 3, a theoretical model was proposed which addresses the relationships among various factors in personalization of IR. Findings of the current study have established some relationships in the model, as well as generated methodological support.

From the data perspective, based on the results of the current study, the model proposed in Chapter 3 can be established in the following way which validates some relationships to be discovered. As Figure 6.2 shows, there was a significant relationship between document usefulness and time when task stage and task type (based on structure) information was also considered. There was also a significant relationship between document usefulness and time when topic knowledge and task type (based on structure) information was also considered. The significant relationship between topic knowledge and time happened only in some task types. Relationships that were supposed to be examined by RQ3 was not able to be answered, and will be examined in future studies.

From a methodology perspective, the findings of this study demonstrated that it is an effective way to follow the proposed model to generate research questions that can investigate salient factors that play significant roles in personalization, and accordingly to design studies to

answer the research questions. There are many other factors that can be addressed in this way, which opens the door for a sequence of future studies.

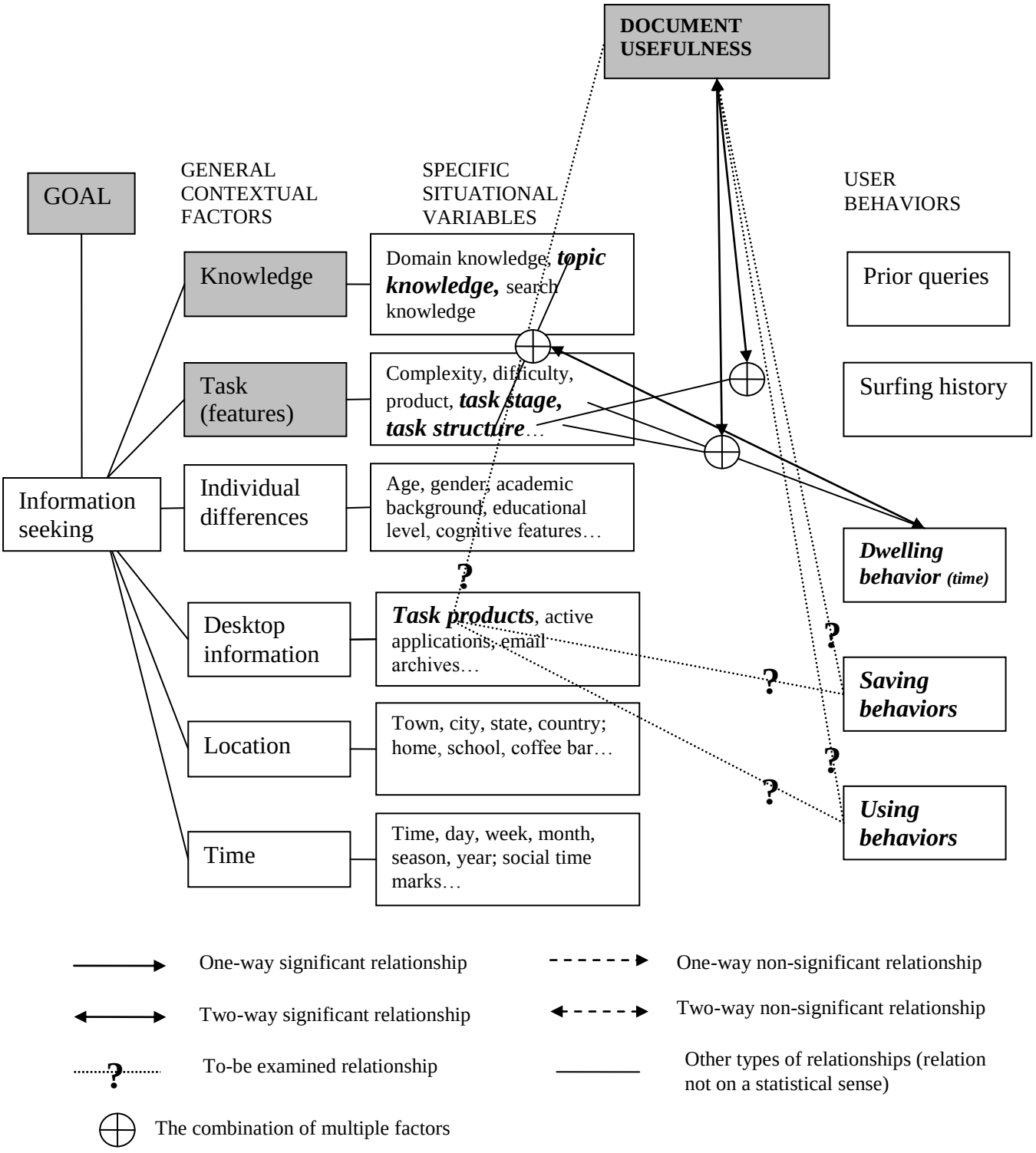


Figure 6.2 The revised IR model showing the relations between factors

Chapter 7. Conclusions

7.1 Summary of this Dissertation Research

This dissertation research was aimed at exploring how information systems can personalize search towards individual users and user situations in multi-session tasks. Three research questions were developed to address the general research goal to look particularly at how task stage and user's topic knowledge may help the search system to infer document usefulness, given how long a person looked at a document, as well as how a user's past work product may help the system generate significant terms that are useful for query reformulation.

This research is important in several ways. Three types of time that users spend on a retrieved document were identified. 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 following paragraphs describe these points in details.

First, it was observed that there were different types of time that users spend on a retrieved document, and that they played different roles in indicating document usefulness. Previous research that studied time as a user behavior 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 did not differentiate 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). One reason is, perhaps, due to the fact that most previous studies have not been

conducted in the context of doing a task with an output other than finding documents. The current dissertation research took a different approach with consideration of how long users spent on specific documents. Observations in the experiments as well as examinations of the data revealed that there could be three types of time for an individual document: total dwell time, total display time, and decision time (i.e., first dwell time). They were found to have different roles in indicating usefulness. In general, total dwell time or total display time alone were shown to be reliable indicators of document usefulness in both types of tasks (although in some cases other factors can help enhance the interpretation accuracy, i.e., the overall correctness of usefulness prediction in future studies based on a certain time threshold). However, due to the limitation that these two types of time cannot be captured until a session is over, their use for personalization can only be applied to subsequent sessions in multi-session tasks but not the current session. On the other hand, using decision time alone to infer document usefulness was found to be reliable only in the dependent task but not in the parallel task, or when task type information was not specified. The use of decision time can be applied to the current session since it can be captured in the early phase of a session.

Second, task stage was found to be helpful in inferring document usefulness from time. While there have been many studies in the literature on task stage looking at users' changes in behaviors and relevance judgment (e.g., Vakkari, 2000, 2001; Taylor et al., 2007), they did not consider how users' behaviors can be helpful in predicting document usefulness. The current study found that task stage was helpful in inferring document usefulness especially from decision time in the parallel task or when no task type information was available, but not in the dependent task, where time (any type) alone is a reliable indicator of usefulness (sine usefulness was shown to have significant main effect on time, i.e., there was significant correlation between usefulness

and time). In the parallel task, decision time alone was not a reliable indicator of document usefulness, but if taking task stage information into account, it could be. Task stage plays an important role in inferring usefulness from decision time for personalizing search in the current session. When task type information was not specified, the finding on decision time was similar to that in the parallel task. In addition, although total display time seemed to be able to infer document usefulness, task stage can potentially enhance the interpretation accuracy (due to the significant interaction effect found between topic knowledge and usefulness on decision time). This role of task stage can be used to personalizing search in subsequent sessions.

Third, user's knowledge of task topic was found to be helpful in interpreting time as an indicator of document usefulness. This role was found to be in effect in the parallel task or when no task type information was available, but not in the dependent task, where time (any type) alone is a reliable indicator of usefulness. In the parallel task, although total display time and total dwell time were found to be reliable indicators of document usefulness, the interpretation accuracy would potentially be enhanced if topic knowledge was also considered (due to the significant interaction effect found between topic knowledge and usefulness on decision time). Decision time alone could not reliably indicate usefulness, but if taking topic knowledge into account, it could. Topic knowledge plays an important role in inferring usefulness from decision time for personalizing search in the current session. When task type information was not specified, the findings were similar to those in the parallel task.

Fourth, when either task stage or topic knowledge information is available, either of them was found to have the potential of helping the system improve usefulness prediction performance by setting personalized thresholds of decision time across different stages or for people with different levels of topic knowledge. When both task stage and topic knowledge information is

available, topic knowledge was found to be a more accurate factor to help infer document usefulness.

Fifth, task type was also an important factor to take into account in inferring document usefulness. In general, it was found that in the dependent task, any of the three types of time: total display time, total dwell time, and decision time, was found to be a reliable indicator of document usefulness, and that factors of task stage and user's topic knowledge did not help. Nevertheless, in the parallel task, task stage and topic knowledge was found to help especially when decision time was used as an indicator of document usefulness because this decision time alone either cannot infer usefulness or stage and topic knowledge can enhance the interpretation accuracy.

Sixth, the study extends the literature by discovering the conditions of some relationships between certain factors that have been found in previous studies. For example, it was found in this study that the significant relationship between time and topic knowledge discovered by Kelly & Cool (2002) happened only in certain circumstances in terms of the knowledge of the users (medium and above), the time used (decision time) and the task type (parallel task, or both tasks combined). This provides conditions for interpreting and making use of these relationships in appropriate and hence more effective ways in future studies and/or in system design.

Seventh, the findings of this study on the effect of task type, as well as findings of other studies on task type classified along other task features (e.g., Freud 2008), provide evidence that task type is a significant factor influencing user behaviors, and that task classification should probably be the right and first thing to do in exploring user behaviors in IIR studies (Kelly, 2004). Future studies should look into task types classified along other task features, such as those suggested by Li (2008): task complexity, task product, and so on.

Last, but not the least, the study has generated a systematic way to examine the roles of contextual factors in helping to infer document usefulness. This can be employed for examination of the roles of other contextual factors, such as other task types based on other features, users' cognitive abilities, as well.

Due to the method being a controlled lab experiment with college students working on certain assigned tasks, care should be taken when generalizing the findings universally. 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 suggested by Li (2008) and vary only in one feature while keep others constant, which makes it possible to generalize the findings relatively safely to other tasks of the same type without concerns of the topicality issue.

7.2 Implications for System Design

This study has implications 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. This study also demonstrates that contextual factors should also 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.).

Practical implications of findings of this study are two-fold: to predict usefulness based on contextual factors of stage, knowledge, and task type; or to predict stage, knowledge and task

type based on usefulness. In both cases, user behavior, i.e., time spent on the documents, and behavior changes, are easily detected. On the one hand, if the system can obtain the stage, knowledge, and/or task type information by tracking users' other behaviors, such as how many times the users have accessed the folder created for the task, or how the users reformulate their queries, the system may be able to predict the usefulness of a document. On the other hand, if the system learns the document usefulness through other user behaviors, such as if the user makes use of the information in the document in his/her writing, then the system may be able to predict the task stage, topic familiarity, or task type.

Finally, this study seems to indicate that many aspects of user behaviors are not isolated, but instead, are related to each other. Effectively integrating findings from multiple user studies in user modeling in system design has the potential to obtain optimized prediction performance, and further studies can be conducted to look into it. The important issue in doing so is to determine which factors should be used in what circumstances, for what purposes, and how to make use of findings on multiple behaviors in operational system design.

7.3 Limitations and Future Studies

This study has some limitations in answering the research questions as well as in interpreting results. First, the sample size is limited, with 24 participants in total, and 12 using each version of system. This did not seem to affect answering RQs 1 and 2 which consider all viewed documents in terms of their usefulness. However, this sampling did affect answering of RQ3. As described above, RQ3 was not able to be addressed due to the reason that the two groups of participants using each version of system were unfortunately not from the same sample. If the sample size were bigger, the imbalance of the two groups of people (only 12 in each group

in the study) could have perhaps been washed out, and this sampling issue may not have been a problem.

The study employed a 3-sub-task design to operationalize task stages. In reality, task stage may have various other types besides this simple design with clear stage boundaries. The findings on the roles that task stage plays are valid, nevertheless, it is an issue to accurately obtain task stage information. Possible suggestions have been discussed in Chapter 6, but future studies will need to actually test the applicability in real life.

The tasks were designed mimicking real life work tasks. Nevertheless, they are still assigned tasks to the users. A longitudinal and naturalistic study in the future is needed to see if there are differences in findings on task stage and topic knowledge when users work with their own tasks at hand. In all these future approaches in real life situations, it should be borne in mind that some tasks do not have clear stage boundaries, and future studies should be able to take care of such cases, as well.

The study found significant three-way relationships between contextual factors (task stage, topic knowledge, task type), usefulness, and time. However, it should be noted that the effect size was not big. Partial eta squared varied from 0.01-0.03 (e.g., see results in Table 5.25). This indicated that time only is not enough to predict usefulness. Other types of behaviors, for example, saving, using, and reading behaviors, etc., will also need to be considered in order to have a better prediction performance of document usefulness based on user behavior. Incorporating different types of user behaviors including time spent on documents to predict document usefulness will be done in future studies.

This study did not consider document length as an influential factor on the time users spent on documents. The reason is that the given task and topic are in general familiar to people,

and the retrieved documents for these tasks are in general easy to be read, so it is not very likely that users will have to spend significantly longer time on longer documents. Given time, future studies can examine the relation between usefulness, time, and document length, to confirm the findings.

This research focused on user behaviors, specifically, the time that they spent on documents, only, therefore, users' performance in terms of how good or how detailed their writings were, was not addressed. Future studies could examine user performance and conduct analysis on the relation between user performance and their topic knowledge, task stage, time spent on documents, time spent on the whole sessions, and so on.

This study focused on the two contextual factors of task stage and topic knowledge. There are other factors that may possibly play roles in inferring usefulness from time, for example, some cognitive characteristics such as need for cognition, task difficulty level, task goal as being specific or general, task product as being factual or intellectual (Li, 2008), to name a few. These will be considered in future studies.

As previously mentioned, this study discovered the relationships among factors and indicated the roles played by task stage and topic knowledge. Future effort is needed to find the thresholds of decision time for documents with different levels of usefulness, in different stages, or for people with different levels of topic knowledge, and compare the prediction correctness and error rates. Based on the threshold of decision time, and maybe some other behavioral evidence, such as saving and using behaviors, personalization systems will be designed and evaluated.

Finally, as mentioned in Chapters 5 and 6, RQ3 was not able to be addressed due to the sampling issue. This RQ remains unanswered until further studies are conducted. As previously

discussed in the method chapter, future studies will also consider using implicit ways to automatically expand user queries using significant terms extracted from users' used and saved Web pages, as well as generated reports.

In conclusion, this research has contributed to a better understanding of how information-seeking behaviors, specifically, 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 research findings have theoretical and practical implications for using behaviors and contextual factors in the development of personalization systems. Future studies are suggested on making use of these findings as well as research on related issues.

APPENDICES

A. Recruitment Notice

A doctoral student at the School of Communication, Information, and Library Studies (SCILS), Jingjing Liu, is seeking upper-division Journalism/Media Studies students to participant in her dissertation research experiment. Participants will be invited to come and work on a journalism assignment three times within a two-week period, at their convenient time slots. Each of the three experiment sessions will last about one hour, and will be held in the Communication and Interaction Laboratory in the SCILS building. The experiment will involve conducting online searching for information that is useful for accomplishing the assignment, and submitting the assignment report at the end of each session. Participants will be asked to complete some questionnaires about their demographic information, search experience, knowledge on the assignment topic, etc., and to make judgments on the usefulness of the information that they have viewed and used during their searches. Various aspects of their searching behaviors will be recorded for subsequent analysis.

All volunteers for this study will receive $30.00 for their participation, and the six volunteers who submit the most detailed assignments will receive an additional $20.00. Taking part in this study has no risks and offers the advantage for students of giving them the opportunity to perform real searching for information to support tasks that journalists encounter every day in their professional lives.

For more information about this study, and to volunteer to be a participant, please send email to Jingjing Liu at jingjing@eden.rutgers.edu. You may also contact Dr. Nicholas J. Belkin, who serves as the Chair to this dissertation research, at nick@belkin.rutgers.edu.

This study is part of the research project, "Personalization of the Digital Library Experience", funded by the U.S. Institute of Museum and Library Services.

B. Consent Form

a. General consent form

You are invited to participate in a research study that is being conducted by Jingjing Liu, who is a doctoral student in Communication, Information, and Library Studies at Rutgers University, along with Professor Nicholas Belkin, Co-Principal Investigator on the study. The purpose of this research is to determine significant factors that can influence search system users on their usefulness judgment of a search result.

Approximately 30 subjects above 18 years old will participate in the study, and each individual's participation will include three sessions, each lasting approximately an hour.

Participation in this study will be conducted in three sessions. In the 1st session, you will complete a background questionnaire first. Then you will select a sub-task to work on, search for useful online information, and write a report for your sub-task. You will complete a before- and a post-session sub-task questionnaire eliciting your knowledge on the task topic. In the 2nd and 3rd session, you will repeat the process of selecting sub-tasks, searching, writing, and completing pre- and post-session sub-task questionnaires. At the end of the 3rd session, you will complete a post-session task questionnaire and an exit interview.

This research is confidential. The research records will include some information about you, such as age, gender, major, online search experience, etc. I will keep this information confidential by limiting individual's access to the research data and keeping it in a secure location.

The research team and the Institutional Review Board at Rutgers University are the only parties that will be allowed to see the data, except as may be required by law. If a report of this study is published, or the results are presented at a professional conference, only group results will be stated. If you do not give permission for further use of the data, your data will be kept until September 2013; if you give permission, the data will be kept in the PI's office and released to others only on explicit permission.

There are no foreseeable risks to participation in this study.

You have been told that the benefits of taking part in this study may be to help develop improved search systems to access library and other document collections. However, you may receive no direct benefit from taking part in this study. You will receive $30.00 for completing the entire study. If you are among the top six participants who have the submitted the most detailed reports for the assignment, you will receive an additional $20.00.

Participation in this study is voluntary. You may choose not to participate, and you may withdraw at any time during the study procedures without any penalty to you. In addition, you may choose not to answer any questions with which you are not comfortable.

If you have any questions about the study or study procedures, you may contact myself at

Jingjing Liu, Ph.D. Candidate
Communication, Information, and Library Studies
Room 303, SCILS building, College Avenue Campus
Rutgers University
4 Huntington Street
New Brunswick, NJ 08901-1071, USA
Office phone: 732-932-7500 ext. 8045
Mobile: (917) 519-1659
Email: jingjing@rutgers.edu

Or you can contact my advisor, Dr. Nicholas J. Belkin, at

Nicholas J. Belkin
Professor (II) of Information Science
Department of Library and Information Science
School of Communication, Information & Library Studies
Room 202, Huntington House, College Avenue Campus
Rutgers University
4 Huntington Street
New Brunswick, NJ 08901-1071, USA
Phone +1 732 932 7500 x8271
Fax +1 732 6916
Email belkin@rutgers.edu

If you have any questions about your rights as a research subject, you may contact the IRB Administrator at Rutgers University at: Rutgers University, the State University of New Jersey, Institutional Review Board for the Protection of Human Subjects, Office of Research and Sponsored Programs, 3 Rutgers Plaza, New Brunswick, NJ 08901-8559
Tel: 732-932-0150 ext. 2104, Email: humansubjects@orsp.rutgers.edu

You will be given a copy of this consent form for your records.

Sign below if you agree to participate in this research study:

Subject (Print) _____

Subject Signature _____ Date _____

Principal Investigator Signature _____ Date _____

b. Consent for audio and video recording

As part of the Study, we will record various aspects of your interaction with the web search system. Your search requests, mouse clicks and other keyboard interactions will be logged. We will also make audio and video recordings of the experimental sessions and record your eye movements.

To participate in the Study you must consent to allow the data recording described above. These logs and recordings will not have your name or any personal information attached to ensure your anonymity and will be treated as previously described.

If you agree to allow the data recording describe above, please sign this form below. If you do not wish to allow this data recording, do not sign this form. If you do not sign this form, you will not be able to participate in the Study.

I, _____, do hereby agree to permit recording of my interaction with the search system during my participation in the Personalizing Information Retrieval Using Task Stage, Topic Knowledge, and Task Product Study, including audio and video recording of search activity and keyboard and mouse use.

_____ _____

Signature Date

_____ _____

Investigator's Signature Date

c. Consent for Further Use of Recorded Data

We would like to ask your permission to use the data collected in this investigation for further research, for demonstration in teaching, and for presentation during conferences. As described above, the data and recordings do not include your name or other identifying information. If you do not want to give your permission for us to use your data, you may still participate in the Study and receive compensation if you complete the experiment.

Use of your data could entail any of the following:

1. Researchers, both at Rutgers and at other institutions, re-analyzing the logs of your tasks, and associated eye tracking, audio and/or video recordings. Such use would be only on approval of this investigator.
2. Playing excerpts of the log audio and/or video recordings of the tasks during presentations of the research results of this project at scholarly conferences or other research or educational meetings. In such use, your face will be made blurred, so you cannot be identified.

If you agree to our making use of the data recorded during your tasks as specified above, please sign this form below. If you do not wish to permit such use, do not sign this form. If you do not sign this form, the logs and recordings will be treated as previously described.

I, _____, do hereby agree that the logs and audio, video, and eye tracking recordings made during my participation in the Study may be used for research, teaching, and demonstration purposes as described above.

Subject Signature _____ Date _____

Principal Investigator Signature _____ Date _____

C. General Instructions

Thank you for volunteering as a participant. The following describes the general instructions of this experiment. Please read it carefully.

1. About SESSIONS
   • You will come in three times, each time finishing one sub-assignment of your general assignment.
   • You can freely choose which sub-assignment you want to work on in each session.
   • Each session will last about an hour. You will have 40 minutes to search for information and write a report. You will also work on some questionnaires and evaluate the web pages that you view. (There will be more details about these in the instructions later on.)
2. About your DESKTOP

   On your desktop, you can find the following items that you will need or may want to use in the experiment.
   • There are three folders named Session 1, Session 2 and Session 3 respectively. These folders will be where you save your stuff (web pages and documents that you save, and the assignment report that you write) in all three sessions.
   • Internet Explorer (IE)
   • Microsoft Word, PowerPoint, and Excel
3. About SEARCHING
   • Please feel free to use any online systems to search for useful information.
   • Please feel free to save any online resources (e.g., Web pages, documents, etc.) that you want to.
4. About WRITING
   • Please feel free to use any online resources that you have located to help write your session reports.
   • Please feel free to use the note sheet to make any notes that you may have.
   • Please remember to submit the sub-assignment report at the end of each session.
5. About PAYMENT
   • If you complete all the requirements for the three sessions in this study, you will receive $30. If you are among the top 6 participants who have the most detailed reports, you will receive an additional $20.

D. Session Instructions

Before you start working with your sub-assignment, please read the following instructions carefully.

• You will have 40 minutes to search and write your report.

• Please use the main monitor to search and write your report, and use the 2nd monitor for the questionnaire ONLY.

• Please keep only ONE IE window open. Do NOT open more than one IE window and Do NOT close the window you use. Use the browser back button to go back to the previous page.

• Remember that you can use the note sheet.

• Remember that you can save the online resources that you like.

• You may freely choose in which ways you want to save the web page content, including:

--- bookmarking,

--- saving web pages,

--- copying and pasting the content into text documents, or

--- other ways that you like.

• When you save your web pages, the copy-and-pasted text documents, or the file you take your notes, please save them under the right folder named according to your session number on the desktop, i.e., folder "Session 1" in your session 1, folder "Session 2" in your session 2, and folder "Session 3" in your session 3.

• Remember to submit your report at the end of each session by saving it in the folder named according to your session number on the desktop, i.e., folder "Session 1" in your session 1, folder "Session 2" in your session 2, and folder "Session 3" in your session 3.

E. General Task

The Parallel task

Suppose you are a beat reporter for automobiles in a newspaper. You have an assignment now, which is 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 hybrid,

• Toyota Camry sedan hybrid, and

• Nissan Altima sedan hybrid.

You want to write about the features of each of the three models, including aspects such as: standard features and specifications, safety, pricing, reviews, and so on.

You have three sessions to finish this assignment. In each session, you will need to finish writing on ONE and ONLY ONE car and submit a report for this car.

The order of the three cars is up to you. In the end of the final session, you will need to integrate all three reports into one FINAL report, which would be your FINAL feature story for this assignment.

The Dependent task

Suppose you are a beat reporter for automobiles in a newspaper. You have an assignment now, which is 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.

You will have three sessions to finish this assignment. The last paragraph of this assignment description lists three activities that you will work on for this assignment. In each session, you will need to finish ONE and ONLY ONE activity and submit a report for this activity.

The order of the three activities is up to you. In the end of the final session, you will need to integrate all three reports into one FINAL report, which would be your FINAL feature story for this assignment.

Here are the three activities:
• Collect information on what manufacturers have hybrid cars. You want to collect the different models that are good for mid-level income families.

• Select three hybrid models that you will focus on in this feature story. You want to introduce their specific features that make you choose them out of other models.

• Compare the advantages and disadvantages of three models of hybrid cars that you choose to focus on in this feature story.

F. Note sheet

Please use this work sheet to take notes of anything that you might want to during the process of searching for information and working on the session report.

G. Background questionnaire

Please tell us about your background information:

1. What is your gender?*
   - Female
   - Male

2. What is your age (in years)?*

3. What is your major?*

4. What is your academic level?*
   - Freshman
   - Sophmore
   - Junior
   - Senior
   - Graduate student
   - Others, please specify

5. What undergraduate and/or graduate degree(s) have you earned? (Please list majors)
   1
   2
   3
   4

Please tell us about your search Experience:

Please indicate the number that most closely describes your searching experience.

6. Please indicate your level of expertise with computers:

Novice						Expert
1	2	3	4	5	6	7

7. How many years have you been doing online searching? _____.

8. What browser(s) do you use in your life?
Enter at least 1 response.

Most frequently used:

2nd most frequently used:

3rd most frequently used:

Others:

9. If 1 stands for "Novice", and 7 stands for "Expert", what is your level of expertise with COMPUTERS?

	1	2	3	4	5	6
Level of expertise with computers	◉	○	○	○	○	○

10. How many years have you been doing online searching?

11. What browser(s) do you use in your life?
Enter at least 1 response.

Most frequently used:

2nd most frequently used:	
3rd most frequently used:	
Others:	

12. If 1 stands for "Novice", and 7 stands for "Expert", what is your level of expertise with SEARCHING?

	1	2	3	4	5	6	7
Level of expertise with searching							

13. Please indicate your level of expertise with searching:

Novice Expert

1 2 3 4 5 6 7

H. Pre-session task questionnaire

1. How familiar are you with the topic of this assignment?
   - 1 Not at all
   - 2
   - 3
   - 4 Somewhat
   - 5
   - 6
   - 7 Extremely

2. How much experience do you have with this kind of assignment?
   - 1 None
   - 2
   - 3
   - 4 Some
   - 5
   - 6
   - 7 A great deal

3. How difficult do you think it will be to find the information for this assignment?
   - 1 Not at all
   - 2
   - 3
   - 4 Somewhat
   - 5
   - 6
   - 7 Extremely

I. Pre-session sub-task questionnaire

1. How familiar are you with the topic of this SUB-assignment?
   - 1 Not at all
   - 2
   - 3
   - 4 Somewhat
   - 5
   - 6
   - 7 Extremely

2. How much experience do you have with this kind of SUB-assignment?
   - 1 None
   - 2
   - 3
   - 4 Some
   - 5
   - 6
   - 7 A great deal

3. How difficult do you think it will be to find the information for this SUB-assignment?
   - 1 Not at all
   - 2
   - 3
   - 4 Somewhat
   - 5
   - 6
   - 7 Extremely

J. Usefulness Evaluation Questionnaire

Here are several questions that you will need to answer for each page the experimenter shows you:

--- How useful was this page in helping you complete and/or understand your sub-assignment?
 1 Not at all
 2
 3
 4 Somewhat useful
 5
 6
 7 Extremely useful

--- How confident are you in your usefulness rating?
 1 Not at all
 2
 3
 4 Somewhat confident
 5
 6
 7 Extremely confident

--- Did you visit this page before?
 0 No
 1 Yes

--- If "yes", how familiar were you with this page?
 1 Not at all
 2
 3
 4 Somewhat familiar
 5
 6
 7 Extremely familiar

K. Post-session sub-task questionnaire

1. At this point, how familiar are you with the topic of this SUB-assignment?
   - 1 Not at all
   - 2
   - 3
   - 4 Somewhat
   - 5
   - 6
   - 7 Extremely

2. How difficult was it to find the information you needed for this SUB-assignment?
   - 1 Not at all
   - 2
   - 3
   - 4 Somewhat
   - 5
   - 6
   - 7 Extremely

3. How successful do you think you were in gathering the information to complete this SUB-assignment?
   - 1 Not at all
   - 2
   - 3
   - 4 Somewhat
   - 5
   - 6
   - 7 Extremely

4. How satisfied are you with the report that you submitted for this SUB-assignment?

- 1 Not at all
- 2
- 3
- 4 Somewhat
- 5
- 6
- 7 Extremely

L. Post-session task questionnaire

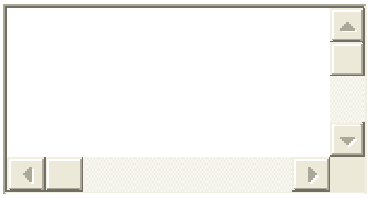
1. At this point, how familiar are you with the topic of this GENERAL assignment?*

- 1 Not at all
- 2
- 3
- 4 Somewhat
- 5
- 6
- 7 Extremely

M. Exit Interview

1. How much did you learn on the topic of this assignment during the whole process?
   - 1 None
   - 2
   - 3
   - 4 Some
   - 5
   - 6
   - 7 A great deal

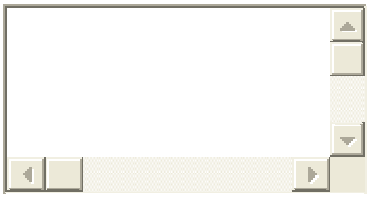
2. Why did you choose this order of the three sub-assignments?



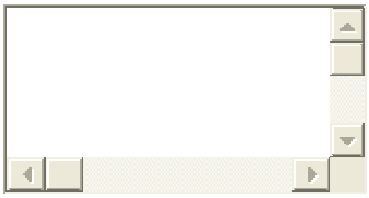
3. How do you think of the system suggested keyword feature in general through the two sessions?[16]
   - 1 Very harmful
   - 2
   - 3
   - 4 Neutral
   - 5
   - 6
   - 7 Very useful

4. Why do you give this rating for question 22?

[16] This question was for those who had used QE version only.



5. Do you have any other comments on the experiment in general?



N. Receipt

ALL INFORMATION WILL BE KEPT PRIVATE AND CONFIDENTIAL

Receipt for payment of $30.00 for Study Participation:

Date: _____

Time slot: _____

Name: _____

Signature: _____

If you are among the top 6 who submit the most detailed report, you will receive an additional $20.00. In order for us to mail the additional $20.00 to you, please leave your mailing address:

_____

_____

Thank you for your participation!

O. Codes for left-side panel in QE interface

```
<html>
<body>

<!--
// the following javascript code is for moving, resizing, and opening the second window.
// -->

<SCRIPT type="text/javascript">
window.moveTo(0,0);
window.resizeTo(280,1024);
var sidebarWidth=280;
var sidebarBorder=30;
var bottomOffset=50;
var evalSidebarWin = window.open("about:blank","yourName", "menubar=1, scrollbars=1,
resizable=1, location=1, toolbar=1, copyhistory=0, width=1000,height=1024, left=280, top=0");
</SCRIPT>


<SCRIPT >
<!--
// by Nannette Thacker
// http://www.shiningstar.net
// This script checks and unchecks boxes on a form
// Checks and unchecks unlimited number in the group...
// Pass the Checkbox group name...
// call buttons as so:
// <input type=button name="CheckAll"   value="Check All"
        //onClick="checkAll(document.myform.list)">
// <input type=button name="UnCheckAll" value="Uncheck All"
        //onClick="uncheckAll(document.myform.list)">
// -->

<!-- Begin
function checkAll(field)
{
for (i = 0; i < field.length; i++)
        field[i].checked = true ;
}
function uncheckAll(field)
{
for (i = 0; i < field.length; i++)
        field[i].checked = false ;
}
//  End -->
```

```
</script>

<!--
// the following javascript code is for copying to clipboard.
// -->

<script>
function copyTerms(field)
{
var textvalue = "";
for (i = 0; i < field.length; i++)
   {
      if (field[i].checked == true)
            textvalue +=  field[i].value+" ";
   }

window.clipboardData.setData('text', textvalue);
}
</script>

<br><br>
<DIV align="center"><TABLE width=50%>
<FONT FACE="Arial" SIZE="4">Recommended terms:<br><br></font>
<FONT FACE="Arial" SIZE="3.5">
<form name="myform">
<input type="checkbox" name="list" value="Term1"> 1. Term 1<br>
<input type="checkbox" name="list" value="Term2"> 2. Term 2<br>
<input type="checkbox" name="list" value="Term3"> 3. Term 3<br>
<input type="checkbox" name="list" value="Term4"> 4. Term 4<br>
<input type="checkbox" name="list" value="Term5"> 5. Term 5<br>
<input type="checkbox" name="list" value="Term6"> 6. Term 6<br>
<input type="checkbox" name="list" value="Term7"> 7. Term 7<br>
<input type="checkbox" name="list" value="Term8"> 8. Term 8<br>
<input type="checkbox" name="list" value="Term9"> 9. Term 9<br>
<input type="checkbox" name="list" value="Term10"> 10. Term 10
</font></tr></table></div>
<DIV align="center"><table width=50%>
<td><input type="button" name="CheckAll" value="Select All"
onClick="checkAll(document.myform.list)" style="width:100; background-color:lightgrey;
font:bold">
<input type="button" name="UnCheckAll" value="Select None"
onClick="uncheckAll(document.myform.list)" style="width:100; background-color:lightgrey;
font:bold">
<br><br>
```

```
<input type="button" value="COPY" onclick="copyTerms(document.myform.list);"
style="width:100; background-color:lightblue; font:bold">
</td>
</TABLE></div>

</body>
</html>
```

Bibliography

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.

Allen, B.L., & Kim, K.-S. (2000). Person and context in information seeking: Interactions between cognitive and task variables. Paper presented at *ISIC 2000: Information Seeking in Context: The 3rd International Conference on Information Needs, Seeking, and Use in Different Contexts*; 2000 August 16-18; Gothenburg, Sweden. Available from: the authors, School of Information Science and Learning Technologies, University of Missouri-Columbia.

Belkin, N.J. (1980) Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, vol. 5: 133-143.

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. (2006). Getting personal: Personalization of support for interaction with information. Talk in *The 1st International Workshop on Adaptive Information Retrieval*, October 2006, Glasgow, UK.

Belkin, N.J., Chaleva, I., Cole, M., Li, Y.-L., Liu, Y.-H., Muresan, G., Smith, C.L., Sun, Y., Yuan, X.-J., & Zhang, X.-M. (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. InProceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation, 7-8.

Belkin, N.J., Muresan, G., & Zhang, X.-M. (2004). Investigating the Effect of the Use of User's Context on IR Performance. *SIGIR 2004 IRiX workshop*, July 2004, Sheffield, UK.

Belkin, N. J., Oddy, R., & Brooks, H. (1982). ASK for information retrieval: Part I. *Journal of Documentation*, 38 (2), 61-71.

Beitzel, S.M., Jensen, E.C., Chowdhury, A., Frieder, O., & Grossman, D. (2007). Temporal analysis of a very large topically categorized Web query log. *Journal of the American Society for Information Science and Technology*, 58(2), 166-178.

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

Budzik, J., & Hammond, J.K. (1999). Watson: Anticipating and contextualizing information needs. In *Proceedings of the 62nd Annual Meeting of the American Society for Information Science*.

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.

Chirita, P.-A., Firan, C.S., & Nejdl, W. (2006). Summarizing local context to personalize global web search. In *Proceedings of CIKM 2006*, 287-296.

Chirita, P.-A., Firan, C.S., & Nejdl, W. (2007). Personalized query expansion for the Web. In *Proceedings of the 30th Annual ACM International Conference on Research and Development in Information Retrieval* (*SIGIR '07)*. Ámsterdam, The Netherlands, 7-14.

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

Croft, B., & Das, R. (1990). Experiments with query acquisition and use in document retrieval systems. *Proceedings of the 13th Annual ACM International Conference on Research and Development in Information Retrieval* (*SIGIR '90).* 349-368.

Dervin, B. (1980). Communication gaps and inequities: Moving toward a reconceptualization. In B. Dervin & M. Voigt (Eds.), *Progress in communication sciences* (Vol. 2, pp. 73-112). Norwood, NJ: Ablex.

Dervin, B. (1992). From the mind's eye of the user: The sense-making qualitative-quantitative methodology. In Jack D. Glazier & Ronald R. Powell (Eds.), *Qualitative Research in Information Management* (pp. 61-84). Englewood Cliffs, CO: Libraries Unlimited.

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.

Dou, Z., Song, R., & Wen, J.-R. (2007). A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of WWW 2007*, 581-590.

Dumais, S., Cutrell, E., Cadiz, J., Jancke, G., Sarin, R., & Robbins, D.C. (2003). Stuff I've seen: A system for personal information retrieval and re-use. *Proceedings of the 26th Annual ACM International Conference on Research and Development in Information Retrieval* (*SIGIR '03)*. Toronto, Canada: 72-79.

Dumais, S. (2007). Information Retrieval in Context. In *Proceedings of IUI '07*, 2.

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

Gwizdka, J. (2006). Findings to keep and organize: Personal information collections as context. Position paper presented at the SIGIR'2006 Workshop on Personal Information Management. Seattle, WA. August 2006.

Gwizdka, J. (2008). Revisiting search task difficulty: Behavioral and individual difference measures. *Proceedings of the American Society for Information Science & Technology*, Columbus, OH.

Gwizdka, J., & Spence, I. (2006). What can searching behavior tell us about the difficulty of information tasks? A study of web navigation. *Proceedings of Annual Meeting of ASIST 2006*, Nov. 3-8, Austin, TX.

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

Ingwersen, P. & Järvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context*. Heidelberg: Springer.

Järvelin, K. and Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual ACM International Conference on Research and Development in Information Retrieval* (*SIGIR '00)*, Athens, Greece, 41–48.

Jones, W. (2007). How people keep and organize personal information. In J. T. W. Jones (Ed.), *Personal Information Management*. Seattle, University of Washingtn Press, 35-56.

Jones, W., Teevan, J. (2007). Introduction. In J. T. W. Jones (Ed.), *Personal Information Managem*ent. Seattle, University of Washington Press: 3-21.

Keenoy, K., & Levene, M. (2005). Personalization of Web search. In S. Anand & B. Mobasher (eds.) *Intelligent Techniques for Web Personalisation*, Springer LNCS 3169, pp. 201-228.

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, D. (2004). *Understanding implicit feedback and document preference: A naturalistic user study*. Ph.D. Dissertation, Rutgers University.

Kelly, D., & Belkin, N.J. (2004). Display time as implicit feedback: Understanding task effects. In *Proceedings of the 27th Annual ACM International Conference on Research and Development in Information Retrieval* (*SIGIR '04)*, Sheffield, UK, 377-384.

Kelly, D., & Cool, C. (2002). The Effects of Topic Familiarity on Information search behavior. In *Proceedings of JCDL '02*, 74-75.

Kelly, D., Murdock, V., Yuan, X.-J., Croft, W.B., & Belkin, N.J. (2002). Features of documents relevant to task- and fact-oriented questions. In *Proceedings of CIKM '02*, 645-647.

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

Kim, J.H. (2006). *Task as a Predictable Indicator for Information Seeking Behavior on the Web*. Unpublished Dissertation. Rutgers University.

Koenemann, J., & Belkin, N.J.: A Case for Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness. *In Proceedings of CHI 1996*: 205-212.

Komlodi, A., Soergel, G., & Marchionini, G. (2006). Search histories for user support in user interfaces. *Journal of the American Society for Information Science and Technology*, 53(6), 803-807.

Komlodi, A., Marchionini, G., & Soergel, G. (2007). Search history support for finding and using information: User interface design recommendations from a user study. *Information Processing & Management*, 43, 10-29.

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 *Proceedings of CIKM '05*, 271-271.

Li, Y. (2004). Task type and a faceted classification of task. *Proceedings of American Society for Information Science and Technology 2004*, November 13-17, RI: Providence, USA.

Li., Y. (2008). *Relationships among work tasks, search tasks, and interactive information searching behavior*. Unpublished dissertation. Rutgers University.

Liu, C., Gwizdka, J., Belkin, N.J. (2010). Analysis of query reformulation types on different search tasks. *Proceedings of i-conference 2010*, Urbana-Champaign, IL, February 3-6, 2010.

Liu, J., Cole, M., Liu, C., Belkin, N.J., Zhang, J., Bierig, R., Gwizdka, J., & Zhang, X. (in press). Search behaviors in different task types. *To appear in Proceedings of JCDL 2010*.

Lin, S.-J. (2001). *Modeling and Supporting Multiple Information Seeking Episodes over the Web.* Unpublished dissertation. Rutgers University.

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.

Marchionini, G., Dwiggins, S., Katz, A., & Lin, X. (1993). Information seeking in full-text end-user-oriented search systems: The roles of domain and search expertise. *Library and Information Science Research*, 15(1), 35-59.

Morita, M., & Shinoda, Y. (1994). Information filtering based on user behavior analysis and best match text retrieval. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 272-281.

Murdock, V., Kelly, D., Croft, W.B., Belkin, N.J., & Yuan, X. (2007). Identifying and improving retrieval for procedural questions. *Information Processing and Management*, 43, 181-203.

Pitkow, J., Schütze, H., Cass, T., Cooley, R., Turnbull, D., Edmonds, A., Adar, E., & Breuel, T. (2002). Personalized search: A contextual computing approach may prove a breakthrough in personalized search efficiency. *Communications of the ACM*, 45(9), 50-55.

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". *LISR*. 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.

Schutz, A., & T. Luckmann (1973). *The Structures of the life-worlds*. (M. Zaner and H. T. Engelhardt. Jr., Trans.). Evanston, IL: Northwestern University Press.

Sihvonen, A., & Vakkari, P. (2004). Subject knowledge improves interactive query expansion assisted by a thesaurus. *Journal of Documentation,* 60(6), 673-690.

Sonnenwald, D. (1999). Evolving perspectives of human behavior: Contexts, situations, social networks and information horizons. In: Wilson, Thomas D.; Allen, David K., eds. *Exploring the Contexts of Information Behaviour: Proceedings of the 2nd International Conference on Research in Information Needs, Seeking and Use in Different Contexts*; 1998 August 13-15; Sheffild, UK. London, UK: Taylor Graham; 1999. 176-190.

Sparck-Jones, K. (1995). Reflections on TREC. *Information Processing and Management*, 31(3), 291-314.

Sparck-Jones, K. (2000). Further reflections on TREC. *Information Processing and Management*, 36(1), 37-85.

Spink, A., Wolfram, D., Jansen, B. J., & Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science*, 52(3), 226-234.

Tague, J., & Schutz, R. (1989). Evaluation of the user interface in an information retrieval system: A model. *Information Processing & Management*, 25(4), 377-389.

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.

Teevan, J., Dumais, S.T., & Horvitz, E. (2005). Personalizing search via automated analysis of interests and activities. In *Proceedings of 28th Annual ACM International Conference on Research and Development in Information Retrieval* (*SIGIR '05)*, 449-456.

Teevan, J., Dumais, S.T., & Liebling, D.J. (2008). To personalize or not to personalize: Modeling queries with variation in user intent. In *Proceedings of SIGIR '08*, 163-170.

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., Dawe, E., & MacNutt, A. (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.

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

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

White, R., Dumais, S., & Teevan, J. (2009). Characterizing the influence of domain expertise on Web search behavior. In *Proceedings of WSDM 2009*.

White, R., & Kelly, D. (2006). A study of the effects of personalization and task information on implicit feedback performance. In *Proceedings of CIKM '06*, 297-306.

White, R.W., Ruthven, I., & Jose, J.M. (2005). A study of factors affecting the utility of implicit relevance feedback. In *Proceedings of 28th Annual ACM International Conference on Research and Development in Information Retrieval* (*SIGIR 2005)*, 35-42.

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.

Curriculum Vitae

Jingjing Liu

(jingjing@eden.rutgers.edu)

EDUCATION

2005-2010	Ph.D., Information Science School of Communication and Information, Rutgers University, New Brunswick, NJ
2006-2010	Graduate Certificate in Cognitive Science Rutgers University Center for Cognitive Science, Rutgers University, New Brunswick, NJ
2002-2003	Master in Library and Information Science The University of Southern Mississippi, Hattiesburg, MS
1992-1996	Bachelor of Arts School of Information Resources Management, Renmin University of China, Beijing, China

EMPLOYMENT

09/2010-present	Assistant Professor Department of Information and Library Science, Southern Connecticut State University
01/2008-08/2010	Graduate Research Assistant School of Communication and Information, Rutgers University
2002-2003	Graduate Assistant School of Library and Information Science, The University of Southern Mississippi, Hattiesburg, MS
2001	Librarian Deloitte, Touche & Thomatsu Beijing Office, Beijing, China
1999-2001	Information Specialist Shun Hing Power Co. Ltd, Beijing, China
1996-1999	Editor *Chinese Archives News*, Beijing, China

PUBLICATIONS

2010:

Liu, J., Gwizdka, J., Liu, C., Belkin, N.J. (2010). Predicting task difficulty in different task types. To appear in *Proceedings of the Annual meeting of the American Society for Information Science & Technology (ASIS&T) 2010* (10p.). Pittsburgh, PA, October 22-27.

Liu, C., Gwizdka, J., Liu, J., Xu, T., Belkin, N.J. (2010): Analysis and evaluation of query reformulations in different task types. To appear in *Proceedings of the Annual meeting of the American Society for Information Science & Technology (ASIS&T) 2010* (10p.). Pittsburgh, PA, October 22-27.

Cole, M., Zhang, X., Liu, J., Liu, C., Belkin, N.J., Bierig, R., & Gwizdka, J. (2010). Are self-assessments reliable indicators of topic knowledge? To appear in *Proceedings of the Annual meeting of the American Society for Information Science & Technology (ASIS&T)* 2010 (10p.). Pittsburgh, PA, October 22-27.

Liu, C., Gwizdka, J., & Liu, J. (2010). Helping identify when users find useful documents: Examination of query reformulation intervals. In *Proceedings of Information Interaction in Context (IIiX) 2010*. New Brunswick, NJ, USA. August 18-22, 2010.

Cole, M., Gwizdka, J., Bierig, R., Belkin, N.J., Liu, J., Liu, C., Zhang, J., & Zhang, X. (2010). Linking search tasks with low-level eye movement patterns. In *Proceedings of European Conference on Cognitive Ergonomics*. Delft, Netherlands. August 25-27, 2010.

Hu, X. & Liu, J. (2010). Evaluation of Music Information Retrieval: Towards a User-Centered Approach. In *Proceedings of the 4th Workshop on Human-Computer Interaction and Information Retrieval (HCIR).* August 22, 2010, New Brunswick, NJ.

Liu, J., Liu, C., Zhang, J., Bierig, R., & Cole, M. (2010). Identifying queries in the wild, wild Web. In *Proceedings of Information Interaction in Context (IIiX) 2010*. New Brunswick, NJ, USA. August 18-22, 2010.

Liu, J. & Belkin, N.J. (2010). Personalizing information retrieval for multi-session tasks: The roles of task stage and task type. To appear in *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR '10)*. Geneva, Switzland, July 19-23, 2010.

Bierig, R., Cole, M., Gwizdka, J., Belkin, N.J., Liu, J., Liu, C., Zhang, J., & Zhang, X. (2010). An experiment and analysis system framework for the evaluation of contextual relationships. *Proceedings of the 2nd International Workshop on Contextual Information Access, Seeking and Retrieval Evaluation (CIRSE '10),* pp. 5-8.

Liu, J., Cole, M., Liu, C., Belkin, N.J., Zhang, J., Bierig, R. et al. (2010). Search behaviors in different task types. In *Proceedings of ACM-IEEE Computer Society Joint Conference on Digital Libraries (JCDL) 2010*.

Liu, J., Liu, C., Gwizdka, J., & Belkin, N.J. (2010). Can search systems detect users' task difficulty? Some behavioral signals. In *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR '10)*. Geneva, Switzland, July 19-23, 2010.

Liu, J. & Belkin, N.J. (2010). Personalizing information retrieval for people with different levels of topic knowledge. In *Proceedings of Joint Conference of Digital Libraries (JCDL '10)*. Goldcoast, Australia, June 21-25, 2010.

Liu, J. & Belkin, N.J. (2010). Personalizing information retrieval using task stage and task type. Poster presented at i-conference 2010, Urbana-Champaign, IL, February 3-6, 2010.

Liu, J. (2010). Personalizing search using task stage: Enhancing information retrieval system performance for multi-session tasks. Poster presented at Association of Library and Information Science Education Conference (ALISE) 2010, Boston, MA, January 12-15, 2010.

2009:

Liu, J. (2009). Personalizing information retrieval using task features, topic knowledge, and task products. *Bulletin of IEEE Technical Committee on Digital Libraries*, 5(3). Available online at http://www.ieee-tcdl.org/Bulletin/current/index.html.

Liu, J., & Zhang, X. (2009). The impact of presentation vs. interaction design on user satisfaction with digital libraries. In *Proceedings of the Annual meeting of the American Society for Information Science & Technology (ASIS&T) 2009* (10p.). Vancouver, Canada, November 2009.

Liu, J. (2009). Personalizing information retrieval using task features, topic knowledge, and task products. In *Proceedings of the 32nd Annual International ACM SIGIR Conference, pp.* 855.

Cole, M., Liu, J., Belkin, N.J., Bierig, R., Gwizdka, J., Liu, C., Zhang, J., & Zhang, X. (2009). Usefulness as the criterion for evaluation of interactive information retrieval. In *Proceedings of*

the 3rd Workshop on Human-Computer Interaction and Information Retrieval (HCIR) (pp. 1-4). October 23, 2009, Washington, DC: Catholic University of America.

Belkin, N., Cole, M., & Liu, J. (2009). A model for evaluation of interactive information retrieval. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pp.7-8.

Zhang, X., Liu, J., Li, Y., & Zhang, Y. (2009). How usable are operational digital libraries – A usability evaluation of system interactions. In *Proceedings of the 1st ACM SIGCHI symposium on Engineering Interactive Computing Systems (EICS) 2009*, pp. 177-186.

Chapman, G., & Liu, J. (2009). Numeracy, frequency, and Bayesian reasoning. *Judgment and Decision Making*, *4*(1), 34-40.

2008:

Zhang, X., Li, Y., Liu, J. & Zhang, Y. (2008). Effects of interaction design in digital Libraries on user interactions. *Journal of Documentation, 64*(3), 438-463.

Liu, J., & Zhang, X. (2008). The effect of need for cognition on search performance. In *Proceedings of the Annual Meeting of the American Society for Information Science & Technology (ASIS&T) 2008* (10p.). Columbus, Ohio, October 2008.

2007:

Liu, J., & Zhang, X. (2007). Document recommender systems: Approaches to increasing information retrieval effectiveness. *Library and Information Services, 51*(12), 11-18.

Zhang, X., Li, Y., Liu, J., & Zhang, Y. (2007). Effects of browse design in digital libraries on users' browsing experience. In *Proceedings of the Library in the Digital Age Conference (LIDA)* (10p.). Dubrovnik, Croatia, May 28-June 2, 2007.

2006:

Li, Y., Zhang, X., Liu, J., & Zhang, Y. (2006). Trained vs. untrained searchers' interaction with search features in digital libraries: a case study. In *Proceedings of the Annual Meeting of the American Society for Information Science & Technology (ASIS&T 2006)* (8p.). Austin, Texas, November 2006.

Li, Y., Zhang, X., Zhang, Y., & Liu, J. (2006). Impact of interaction design for search features in digital libraries on user searching experience. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR '06)* (pp. 669-670). Seattle, WA, August 6-11, 2006.

2005:

Belkin, N.J., Cole, M., Gwizdka, J., Li, Y.-L., Liu, J.-J., Muresan, G., Roussinov, D., Smith, C.A., Taylor, A., & Yuan, X.-J. (2005). Rutgers information interaction lab at TREC 2005: Trying HARD. In *Proceedings of Text Retrieval Conference (TREC) 2005.*

2003:

Liu, J. (2003). A bibliometric study: Author productivity and co-authorship features of JASIST 2001-2002. *Mississippi Libraries*, 67(4), 110-112.