PERSONALIZING INFORMATION RETRIEVAL USING INTERACTION BEHAVIORS IN SEARCH SESSIONS IN DIFFERENT TYPES OF TASKS

by

CHANG 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, Information and Library Studies

Written under the direction of

Nicholas J. Belkin, Ph.D.

and approved by

____

____

____

____

New Brunswick, New Jersey

October, 2012

ABSTRACT OF THE DISSERTATION

PERSONALIZING INFORMATION RETRIEVAL USING INTERACTION BEHAVIORS IN SEARCH SESSIONS IN DIFFERENT TYPES OF TASKS

By CHANG LIU

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

When using information retrieval (IR) systems, users often pose short and ambiguous query terms. It is critical for IR systems to obtain more accurate representation of users' information need, their document preferences, and the context they are working in, and then incorporate them into the design of the systems to tailor retrieval to individual users. The proposed study is to personalize IR systems by tailoring search result content to individual users through the inference of useful documents during their information seeking episode, in different types of tasks. Specifically, this dissertation has two research goals: (1) generate predictive models of document usefulness based on multiple user behaviors as in different types of tasks; (2) generate predictive models of task type through observing users' search behaviors. To address these research goals, this study analyzed data collected in a controlled lab experiment. Thirty-two students were invited to participate in the study, each worked on four search tasks, and these tasks were designed to be different types. During search sessions, all users' interactions were recorded by multiple loggers. Predictive models of document usefulness and task type were generated using various statistical analysis methods. Our results demonstrate that multiple behavioral measures on both content pages and search result pages can be indicators of document usefulness. More importantly, task type affected the relationship between the behavioral measures and document usefulness, and it may therefore be necessary to build task-specific predictive models of document usefulness, which can achieve better prediction accuracy than a non-task specific predictive model. In addition, behavioral measures on within-session level and whole-session levels could be able to generate predictive models of task type. The results improve our understanding on how to infer users' search context information and document usefulness from user behaviors, and then to use this knowledge to improve the information searcher's experience; that is, to make their information search more effective and pleasurable. The research findings have theoretical and practical implications for using behavioral measures and taking account of contextual factors in the development of personalized IR systems. Future studies are suggested for making use of these findings as well as research on related issues.

ACKNOWLEDGEMENT

During the course of my graduate career I have benefited from the help, support, advice and suggestions of a great many people. I would like to express my gratitude to these individuals.

I would like to express my deep and sincere gratitude to my mentor and advisor, Dr. Nicholas J. Belkin, for his guidance, encouragement, patience, and support during my graduate research and this dissertation study. I cannot forget the weekly meetings with him through the whole process of my dissertation research, and his painstaking and critical comments on my writings. More importantly, he always demonstrated his faith in my work and my ability. I am fortunate to have had him as my advisor and role model in research and in life.

I would also like to thank the members of my dissertation committee, Dr. Jacek Gwizdka, Dr. Smaranda Muresan and Dr. Diane Kelly, for their time and effort for reviewing and commenting this research. Special thanks to Dr. Jacek Gwizdka, who have greatly inspired me during my research. Also thanks to Dr. Diane Kelly, who always never tired of answering my questions and shared her expertise in methodology and evaluation in interactive information retrieval. Dr. Smaranda Muresan was always eager to engage my thoughts and has been very generous with her time and insights.

I am very grateful I had the opportunity to be a member of the Personalization of the Digital Library Experience (PoODLE) project. This three-year project was very important in my research journey, from which I learned methods in conducting user experiments and various methods in describing and evaluating user experience in information search. I thank the project members, Dr. Nick Belkin, Dr. Jacek Gwizdka, Dr. Xiangmin Zhang, Dr. Jingjing Liu, Michael Cole, Dr. Ralf Bierig, and Jun Zhang. It was really a great time collaborating with them and learning from them. They provided many invaluable suggestions in facilitating my journey to this dissertation. I also thank them for allowing me to use the data collected in the experiment in this project for this dissertation study.

There are researchers outside Rutgers that I would like to thank for their insightful comments on my dissertation at various conferences and workshops, including Marcia Bates, Dr. Pia Borlund, Dr. Katriina Byström, Dr. Luanne Freund, Dr. Gene Golovchinsky, Dr. Ian Ruthven, Dr. Susan Dumais, Dr. Cathy Blake, Dr. Michael Twidale, Dr. Chengxiang Zhai, I would also like to thank the anonymous reviewers from the SIGIR'12 conference. I would like to specially thank Dr. David Lewis and Dr. Tina Eliassi-Rad for the discussion and guidance in the statistical methods in my dissertation study.

The faculty of the School of Communication and Information (SC&I) at Rutgers University has provided me with a tremendous graduate education: they have taught me how to think about research in library and information; have provided resources for conducting research; and have provided opportunities for me to teach. I would like to thank Dr. Claire McInerney, Dr. Marija Dalbello, Dr. Nina Wacholder, Dr. Dan O'Conner, Dr. Paul Kantor, Dr. Marie Radford, Dr. Ya-Ling Lu, Dr. Chirag Shah and Dr. Xiaomu Zhou, for their invaluable suggestions and support during my graduate study at Rutgers. I would also like to thank Dr. Anselm Spoerri, Dr. Stew Mohr, and Dr. Emily Knox for their guidance, collaboration and suggestions for my teaching practices.

I want to thank the good friends that I have made in Rutgers for their support throughout my graduate study. I would like to thank my cohort doctoral students at Rutgers: Dr. Marianne Martens, Hannah Kwon, Sarah Legins, who had helped me get through the difficulty times in the beginning of my graduate study in U.S. and continued encouragement and support through the dissertation. I would like to thank Dr. Xiaojun Yuan, Dr. Yuelin Li, Dr. Ying-Hsang Liu and Dr. Colleen Cool, who have always been very supportive in my research and career. Many thanks to my friends: Haiyan Zhang, Xinyi Cui, Zhe Li, Jing Ning, Yu Zhang, Jun Tan, Xing Luo, Xin Tong, Yuan Lin, Li Gu, Hang Liu, Vicky Chu, Miao Chen, Xiaozhong Liu, Chih-Hui Lai, Cindy Lu, Kyong Eun Oh, Si Sun, Chenfang Yang, Yongwei Yang. Their friendship and support helped me finish my Ph.D. program.

I also would like to thank IMLS grant LG#06-07-0105, which has provided financial support for my research in the PoODLE project; and thank SC&I for offering me assistantship and financial support which helped me finish my dissertation study.

I am deeply grateful for the support my parents provided during the last phase of this dissertation. The time they gave of their own personal lives to come here from across the world was a big help in my effort to complete the dissertation. I most want to thank my husband Tao for strongly supporting me and caring for me through this process. I also credit our daughter, Grace, for inspiring and amazing me every day.

DEDICATION

This thesis is dedicated to my husband, Tao Xu, my daughter, Grace Xu, and my parents, Xiwen Ge and Xiangbin Liu. Thank you for your love and support.

Table of Contents

ABSTRACT OF THE DISSERTATION .... **II**

ACKNOWLEDGEMENT .... **IV**

DEDICATION .... **VII**

TABLE OF CONTENTS .... **VIII**

LIST OF TABLES .... **XI**

LIST OF ILLUSTRATIONS .... **XVI**

CHAPTER 1. INTRODUCTION **1**

CHAPTER 2. LITERATURE REVIEW **5**

2.1 USERS' INTERESTS AND PREFERENCE LEARNING 5
2.1.1 Implicit measures of users' interests and preferences *5*
2.1.2 Discussion of research on implicit evidence for relevance feedback. *16*
2.2 CONTEXTUAL FACTORS IN INFORMATION RETRIEVAL RESEARCH 18
2.2.1 Task type effect on information search behaviors *19*
2.2.2 Users' knowledge effect on information search behaviors *32*
2.2.3 Other contextual factors *34*
2.3 THE EVALUATION OF PERSONALIZATION ALGORITHMS 35
2.3.1 Prediction accuracy *36*
2.3.2 Personalization performance *37*

CHAPTER 3. THEORETICAL FRAMEWORK **40**

3.1 RESEARCH MODEL 40
3.2 VARIABLES CONSIDERED IN THE CURRENT STUDY 42
3.3 RESEARCH QUESTIONS 43

CHAPTER 4. METHODOLOGY **45**

4.1 EXPERIMENTAL DESIGN 45
4.2 EXPERIMENTAL SYSTEM 45
4.3 PARTICIPANTS 46
4.4 PROCEDURE 47
4.5 SEARCH TASKS AND TASK TYPES 48
4.6 PREDICTIVE MODELING METHODS 53
4.7 MODELING ON BEHAVIORAL MEASURES 54

CHAPTER 5. RESULTS FOR RQ1 .... **58**

5.1 DESCRIPTIVE ANALYSIS 58
5.1.1 Descriptive analysis of participants *58*
5.1.2 Descriptive analysis of content pages visited *60*
5.1.3 Descriptive analysis of behavioral measures *61*
5.2 UNIVARIATE ANALYSES OF BEHAVIORAL MEASURES 62
5.3 GENERAL PREDICTIVE MODELING 64

5.3.1 Predictive modeling using recursive partitioning ... 64
5.3.2 Predictive modeling using logistic regression ... 66
5.4 SPECIFIC PREDICTIVE MODELING ... 68
5.4.1 Specific model for the BIC task ... 69
5.4.2 Specific model for the CPE task ... 71
5.4.3 Specific model for the INT task ... 74
5.4.4 Specific model for the OBI task ... 77
5.5 COMPARISON OF PREDICTIVE MODELS ... 79
5.5.1 Discussion of the two general predictive models ... 81
5.5.2 Comparison of four predictive models in the BIC task ... 82
5.5.3 Comparison of four predictive models in CPE ... 83
5.5.4 Comparison of four predictive models in the INT task ... 85
5.5.5 Comparison of four predictive models in the OBI task ... 86
5.5.6 Comparison of the prediction performance of the specific models in the four tasks ... 88
5.6 DISCUSSION FOR RQ1 ... 89
5.6.1 Behavioral measures as predictors of document usefulness ... 89
5.6.2 Task effect on the predictive models ... 95
5.7 SUMMARY FOR RQ1 ... 97

CHAPTER 6. RESULTS FOR RQ2 ... 99

6.1 DESCRIPTIVE ANALYSIS OF THE SELECTED BEHAVIORAL MEASURES ... 99
6.1.1 Descriptive analysis of behavioral measures on the whole-session level ... 99
6.1.2 Descriptive analysis of behavioral measures on the within-session level ... 101
6.2 COMPARISON OF SEARCH BEHAVIORS BY TASKS ... 103
6.2.1 Overall comparison by task ... 103
6.2.2 Comparison by task facets ... 105
6.3 PREDICTING INDIVIDUAL TASK TYPES ... 113
6.3.1 Prediction using whole-session level measures ... 115
6.3.2 Prediction using within-session level measures ... 120
6.3.3 Prediction using both within-session and whole-session level measures ... 126
6.4 DISCUSSION FOR RQ2 ... 128
6.4.1 Task and task facet effect on behavioral measures ... 128
6.4.2 Predicting task type from behavioral measures ... 133
6.4.3 Application of the predictive models of task type ... 136

CHAPTER 7. CONCLUSION ... 143

7.1 SUMMARY OF THIS DISSERTATION RESEARCH ... 143
7.2 IMPLICATIONS FOR PERSONALIZED IR SYSTEM DESIGN ... 146
7.3 LIMITATIONS AND FUTURE STUDIES ... 147

APPENDICES ... 151

A. DESCRIPTIVE ANALYSIS RESULTS ON BEHAVIORAL MEASURES IN RQ1 ... 151
B. PREDICTIVE MODELING ON RANDOM SAMPLES USING RECURSIVE PARTITIONING IN RQ1 ... 156
a. For the general model ... 156
b. For BIC task ... 158
c. For CPE task ... 160

d. For INT task *162*
e. For OBI task *164*
C. SELECTED PREDICTORS IN RANDOM SAMPLES IN RQ1 165
a. For the general model *165*
b. For BIC task *168*
c. For CPE task *171*
d. For INT task *173*
e. For OBI task *176*
D. DESCRIPTIVE ANALYSIS RESULTS ON WHOLE-SESSION LEVEL MEASURES IN RQ2 179
E. DESCRIPTIVE ANALYSIS RESULTS ON WITHIN-SESSION LEVEL MEASURES IN RQ2 182

REFERENCES 186

VITA 193

List of Tables

Table 1 Classification of behavioral signals as implicit relevance feedback and related literature 14

Table 2 Classification of behavioral measures in information search tasks (from the literature review) 32

Table 3. Facets of task which were varied in this study 49

Table 4 Variable facet values for the search tasks 52

Table 5. Behavioral measures examined in RQ1 55

Table 6. The behavioral measures examined in RQ2 56

Table 7. Frequency table of the saved and non-saved pages 60

Table 8. Frequency table of the saved and non-saved pages in each of the tasks 61

Table 9. Descriptive statistics of behavioral measures on the whole-session level 62

Table 10. Univariate comparisons of candidate predictor variables between saved and non-saved pages. 63

Table 11. The coefficients for selected predictors in ten samples (general model) 67

Table 12. The prediction performance using the general model 68

Table 13. The selected predictors from recursive partitioning in ten samples (BIC) 69

Table 14. The coefficients for selected predictors in the ten samples of BIC task 71

Table 15. The selected predictors from recursive partitioning in ten samples (CPE) 72

Table 16. The coefficients for selected predictors in the ten samples of CPE task 73

Table 17. The selected predictors from recursive partitioning in ten samples (INT) 74

Table 18. The coefficients for selected predictors in the ten samples of INT task 76

Table 19. The selected predictors from recursive partitioning in ten samples (OBI) 77

Table 20. The coefficients for selected predictors in the ten samples of OBI task 79

Table 21. Prediction performance comparison by the two general predictive models 81

Table 22. Prediction performance comparison by general and specific models in BIC 82

Table 23. Prediction performance comparison by general and specific models in CPE 84

Table 24. Relative importance of the predictors in CPE... 84

Table 25. Prediction performance comparison by general and specific models in INT 85

Table 26. Relative importance of the predictors in INT 86

Table 27. Prediction performance comparison by general and specific models in OBI................ 87

Table 28. Relative importance of the predictors in OBI 88

Table 29. Prediction performance comparison among four tasks.. 88

Table 30. The summary table of behavioral measures identified as predictors of document usefulness (+ = positive relationship, - = negative relationship) 90

Table 31. Descriptive statistics of behavioral measures on the whole-session level 100

Table 32. Descriptive statistics of behavioral measures on the within-session level.................. 102

Table 33. Comparison of behavioral variables on whole-session level among tasks 104

Table 34. Comparison of behavioral variables on within-session level among tasks................. 105

Table 35. Comparison of behavioral variables on whole-session level by product..................... 106

Table 36. Comparison of behavioral variables on within-session level by product 107

Table 37. Comparison of behavioral variables on whole-session level by complexity 108

Table 38. Comparison of behavioral variables on within-session level by complexity............... 109

Table 39. Comparison of behavioral variables on whole-session level by level 110

Table 40. Comparison of behavioral variables on within-session level by level......................... 111

Table 41. Comparison of behavioral variables on whole-session level by goal (quality) 112

Table 42. Comparison of behavioral variables on within-session level by goal (quality)........... 113

Table 43. The Step Summary table for whole-session level measures....................................... 115

Table 44. Model Fitting Information in MLR on whole-session level 116

Table 45. Likelihood Ratio Tests in MLR on whole-session level... 116

Table 46. Parameter Estimates in MLR on the whole-session level... 117

Table 47. Classification table in MLR on whole-session level... 120

Table 48. The Step Summary table for within-session level measures....................................... 121

Table 49. Model Fitting Information in MLR on within-session level....................................... 122

Table 50. Likelihood Ratio Tests in MLR on within-session level 122

Table 51. Parameter Estimates in MLR on the within-session level 124

Table 52. Classification table in MLR on within-session level 126

Table 53. Parameter Estimates in MLR on both whole-session and within-session levels 127

Table 54. Classification table in MLR on both within-session and whole-session levels 128

Table 55. Comparison of predictive models on the whole-session level.................................... 129

Table 56. Comparison of predictive models on the within-session level 131

Table 57. The important predictors of task type on the whole-session level.............................. 134

Table 58. The important predictors of task type on the within-session level.............................. 135

Table 59. The default Cost-Benefit Matrix.. 139

Table 60. The “actual” Cost-Benefit Matrix.. 139

Table 61. Selected predictors in sample 1 (general model) 165

Table 62. Selected predictors in sample 2 (general model) 166

Table 63. Selected predictors in sample 3 (general model) 166

Table 64. Selected predictors in sample 4 (general model) 166

Table 65. Selected predictors in sample 5 (general model) 166

Table 66. Selected predictors in sample 6 (general model) 167

Table 67. Selected predictors in sample 7 (general model) 167

Table 68. Selected predictors in sample 8 (general model) 167

Table 69. Selected predictors in sample 9 (general model) 167

Table 70. Selected predictors in sample 10 (general model) 168

Table 71. Selected predictors for sample 1 (BIC).. 168

Table 72. Selected predictors for sample 2 (BIC).. 168

Table 73. Selected predictors for sample 3 (BIC)........168

Table 74. Selected predictors for sample 4 (BIC)........169

Table 75. Selected predictors for sample 5 (BIC)........169

Table 76. Selected predictors for sample 6 (BIC)........169

Table 77. Selected predictors for sample 7 (BIC)........169

Table 78. Selected predictors for sample 8 (BIC)........170

Table 79. Selected predictors for sample 9 (BIC)........170

Table 80. Selected predictors for sample 10 (BIC)........170

Table 81. Selected predictors for sample 1 (CPE)........171

Table 82. Selected predictors for sample 2 (CPE)........171

Table 83. Selected predictors for sample 3 (CPE)........171

Table 84. Selected predictors for sample 4 (CPE)........171

Table 85. Selected predictors for sample 5 (CPE)........172

Table 86. Selected predictors for sample 6 (CPE)........172

Table 87. Selected predictors for sample 7 (CPE)........172

Table 88. Selected predictors for sample 8 (CPE)........172

Table 89. Selected predictors for sample 9 (CPE)........172

Table 90. Selected predictors for sample 10 (CPE)........173

Table 91. Selected predictors for sample 1 (INT)........173

Table 92. Selected predictors for sample 2 (INT)........173

Table 93. Selected predictors for sample 3 (INT)........174

Table 94. Selected predictors for sample 4 (INT)........174

Table 95. Selected predictors for sample 5 (INT)........174

Table 96. Selected predictors for sample 6 (INT)........175

Table 97. Selected predictors for sample 7 (INT)........175

Table 98. Selected predictors for sample 8 (INT) 175

Table 99. Selected predictors for sample 9 (INT) 176

Table 100. Selected predictors for sample 10 (INT) 176

Table 101. Selected predictors for sample 1 (OBI) 176

Table 102. Selected predictors for sample 2 (OBI) 176

Table 103. Selected predictors for sample 3 (OBI) 177

Table 104. Selected predictors for sample 4 (OBI) 177

Table 105. Selected predictors for sample 5 (OBI) 177

Table 106. Selected predictors for sample 6 (OBI) 177

Table 107. Selected predictors for sample 7 (OBI) 178

Table 108. Selected predictors for sample 8 (OBI) 178

Table 109. Selected predictors for sample 9 (OBI) 178

Table 110. Selected predictors for sample 10 (OBI) 178

List of Illustrations

Figure 1. The theoretical model of the study 41

Figure 2. The interface of the experiment system........ 47

Figure 3 The general predictive model 66

Figure 4. Specific decision-tree predictive model for BIC 70

Figure 5. Specific decision-tree predictive model for CPE........ 73

Figure 6. Specific decision-tree predictive model for INT 76

Figure 7. Specific decision-tree predictive model for OBI........ 78

Figure 8. The process of implementing predictive models of document usefulness 97

Figure 9. Application graph of the personalization models 142

Figure 10. The distribution and the normal Q-Q Plot of dwell time........ 151

Figure 11. The distribution and the normal Q-Q Plot of number of mouse clicks........ 151

Figure 12. The distribution and the normal Q-Q Plot of number of keystrokes 152

Figure 13. The distribution and the normal Q-Q Plot of time to first click 152

Figure 14. The distribution and the normal Q-Q Plot of content _mean 152

Figure 15. The distribution and the normal Q-Q Plot of content_sum 153

Figure 16. The distribution and the normal Q-Q Plot of content_count 153

Figure 17. The distribution and the normal Q-Q Plot of serp_mean 153

Figure 18. The distribution and the normal Q-Q Plot of serp_sum 154

Figure 19. The distribution and the normal Q-Q Plot of serp_count 154

Figure 20. The distribution and the normal Q-Q Plot of prop_content........ 154

Figure 21. The distribution and the normal Q-Q Plot of query interval 155

Figure 22. Predictive modeling on ten samples using recursive partitioning for the general model 157

Figure 23. Predictive modeling of BIC on ten samples using recursive partitioning 159

Figure 24. Predictive modeling of CPE on ten samples using recursive partitioning.................161

Figure 25. Predictive modeling of INT on ten samples using recursive partitioning 163

Figure 26. Predictive modeling of OBI on ten samples using recursive partitioning 165

Figure 27. The histogram and the normal Q-Q Plot of task completion time.............................179

Figure 28. The histogram and the normal Q-Q Plot of number of all documents 179

Figure 29. The histogram and the normal Q-Q Plot of Numbers of unique documents 179

Figure 30. The histogram and the normal Q-Q Plot of Number of SERPs.................................180

Figure 31. The histogram and the normal Q-Q Plot of Number of unique SERPs......................180

Figure 32. The histogram and the normal Q-Q Plot of Number of queries 180

Figure 33. The histogram and the normal Q-Q Plot of Total time spent on documents (seconds) 181

Figure 34. The histogram and the normal Q-Q Plot of Total time spent on SERPs (seconds)....181

Figure 35. The histogram and the normal Q-Q Plot of Ratio of document time to all 181

Figure 36. The histogram and the normal Q-Q Plot of Ratio of SERP time to all.......................182

Figure 37. The histogram and the normal Q-Q Plot of Mean dwell time of all documents.........182

Figure 38. The histogram and the normal Q-Q Plot of Mean dwell time of unique documents..183

Figure 39. The histogram and the normal Q-Q Plot of Mean dwell time of all SERPs...............183

Figure 40. The histogram and the normal Q-Q Plot of Mean dwell time of unique SERPs........183

Figure 41. The histogram and the normal Q-Q Plot of Number of documents per query 184

Figure 42. The histogram and the normal Q-Q Plot of Number of unique documents per query 184

Figure 43. The histogram and the normal Q-Q Plot of Number of SERPs per query.................184

Figure 44. The histogram and the normal Q-Q Plot of Number of unique SERPs per query......185

Figure 45. The histogram and the normal Q-Q Plot of Average query interval 185

Figure 46. The histogram and the normal Q-Q Plot of Average Time To First Click.................185

Chapter 1. Introduction

With the development of information access, there is an overabundance of information for users to choose on the Internet. When using information retrieval (IR) systems, users often pose short and ambiguous query terms. Sometimes even though different searchers have posted the same queries, their information needs are different given their search contexts. Thus, it is critical for IR systems to obtain more accurate representations of users' information needs, their document preferences, and the context in which they are working in, and then incorporate this into the design of the systems to tailor retrieval to individual users. Personalized IR systems are proposed to address this issue.

Personalized IR systems attempt to take into account contextual factors and tailor various aspects of the search experience to individual users. There are many different ways to personalize IR systems, with respect to the particular aspects of search experience and different sources that are used to do personalization. Search experience encompasses a wide variety of aspects of the search, such as the interaction mode by users' preferences, adapting interfaces according to a person's cognitive style, tailoring search result content with respect to the documents' relevance or usefulness, information presentation with respect to the display format, and so on. This dissertation focuses on tailoring search result content through inference of useful documents to users during their search episodes.

Relevance feedback has been used as a method to increase performance effectiveness in information retrieval research. A typical relevance feedback system requires users to assess the relevance of some retrieved documents after they get the result list with their initial queries, and then the system generates new terms to expand or modify their initial queries and provides a revised result list (Rocchio, 1971). The traditional approach of relevance feedback requires users to explicitly provide feedback, which is an additional activity for users beyond their own information seeking behaviors and work task. Due to such reasons, the effectiveness of explicit techniques to obtain document relevance can be limited, and researchers started to investigate user behaviors as implicit evidence of document relevance or usefulness as a substitute for explicit feedback. Implicit evidence is obtained by unobtrusive observation of the interactions between users and systems, and has the advantage of not interrupting users from their search.

There have been a number of studies that have investigated which user behaviors can serve as implicit indicators of document preferences (Claypool, et. al, 2001; Nichols, 1997; Oard and Kim, 2001; Kelly and Teevan, 2003). The behavior sources that have been investigated as sources of implicit feedback include clickthrough on search result pages, dwell time on content pages, saving, printing and so on. The results of these studies have been encouraging, but most of them have focused on single behaviors, either on search result pages or content pages, and none of them have combined multiple behaviors on both search result pages and individual content pages to infer document usefulness.

Besides predicting document usefulness for personalization of retrieval system, it is also important for the systems to identify the context of the current search; for example, the task type, users' knowledge of the task, their search stage, etc. In the research field of human information behavior, many studies have shown that these contextual factors influence the way users search for information (Li, 2008; Toms et al., 2008; Kim, 2009), the type of information users expect (Murdock et al., 2006; Freund, 2008); and more importantly, that these contextual factors could affect the reliability of user behavioral measures for implicit relevance feedback (Kelly, 2004; Kelly and White, 2006; Liu and Belkin, 2010). However, there has not been much research on automatically identifying such contextual factors through user behaviors.

This dissertation hopes to contribute to search result content personalization in four aspects: (1) investigate which behavioral measures on search result pages and content pages can serve as implicit predictors of document preferences; (2) generate predictive models of document usefulness based on multiple user behaviors in different types of tasks; (3) compare whether task-specific predictive models can lead to better prediction performance of document usefulness than a non-task specific predictive model; (4) examine which search behaviors can be used to predict task type.

To address these research goals, this study analyzed data collected in a controlled lab experiment. Thirty-two students were invited to participate in the study, each worked on four search tasks, designed to be of different types. Users' search interactions were logged by multiple loggers and explicit judgments of document usefulness were collected by allowing users to save useful pages. We first generated a general predictive model that was not task specific by considering the interactions in all tasks in this experiment. We then generated task-specific predictive models of document usefulness in each type of task, and then compared them with the general model, in both the selected behavioral measures as predictors and the prediction performance. Our results demonstrate that multiple behavioral measures on both content pages and search result pages could be indicators of document usefulness. More importantly, task type affected the relationship between the behavioral measures and document usefulness, and it is necessary to build task-specific predictive models of document usefulness, which could achieve better prediction accuracy than the non-task specific predictive model. Secondly, we explored behavioral measures to generate predictive models of task type. Our results demonstrate that behavioral measures collected during search (on the within-session level) could be used to predict task type, and if combined with behavioral measures after the task is done (on the whole-session level), the system may achieve better prediction performance.

Chapter 2. Literature Review

This chapter contains a review of literature about learning users' interests and the effect of contextual factors on users' search behaviors. It begins with implicit relevance feedback techniques, including what behaviors have been examined in previous studies, followed by a discussion of the limitations of current research. Then the effects of contextual factors on users' information search behaviors are presented. In the third part, the methods that have been used to evaluate the performance of personalization algorithms are reviewed.

2.1 Users' interests and preference learning

2.1.1 Implicit measures of users' interests and preferences

When users search on IR systems, they may be neither familiar with their current information problems, nor able to specify these problems with accurate or complete query terms. Search results for users' original queries, especially ambiguous or broad ones, might not be good. Relevance feedback is an important technique in IR to obtain relevant documents automatically by modification of users' initial queries (Rocchio, 1971). The assumption underlying this idea is that an ideal query exists for each information problem within each document store, which can differentiate a particular set of documents from all the others in the store. When the user gets a result list after he/she issues an initial query, relevance feedback requires users to assess the relevance of some retrieved documents. By using the content of the relevant documents, the system can either adjust the weights of terms in the initial query, or extract additional terms to expand the query, and then provide a revised search result list. The method that explicitly requires users to provide their judgments of documents is called explicit relevance feedback (ERF). The disadvantage of ERF is that requiring users to judge document usefulness/relevance is an additional task besides searching and their current work tasks. Therefore, a method that could infer users' preferences based on the observation of search behaviors is proposed, which has attracted much attention in the research. Compared to ERF, the implicit method is called implicit relevance feedback (IRF), using information about searchers' preferences and interests to identify documents to be used for relevance feedback. The advantage of IRF is that the system can unobtrusively observe users' search behaviors, and automatically infer users' preferences and better the search results. This section of review examines what user behaviors and measures of behaviors have been found to be significant in learning users' interests and preferences.

1) Behaviors on documents

Nichols (1997) provided the first classification of behaviors as implicit feedback, in which he suggested 13 types of implicit information. Subsequent work by Oard and Kim (2001), Claypool, Le, Waseda, and Brown (2001), and Kelly and Teevan (2003) further developed the classification of behavioral sources for implicit feedback to infer users' preferences and relevance judgments. Some behaviors among these implicit sources have been extensively investigated, like reading time, saving, printing, selecting and referencing to be able to infer users' preferences.

Dwell time, usually understood as the length of time a document is displayed, has been examined for implicit relevance feedback in many studies. Some previous studies have shown contradictory results for the prediction of usefulness from dwell time. For example, Morita and Shinoda (1994) found the reading time for articles rated as interesting was longer than for articles rated as uninteresting. However, Kelly and Belkin (2004) did not find a significant relationship between display time and usefulness judgments in their naturalistic study. Recent studies have been focusing on the interaction of contextual factors (e.g. tasks, stage of searching, etc.) with user' behaviors on documents' usefulness prediction. Based on the same dataset in Kelly and Belkin (2004), White and Kelly (2006) further explored the interactions between dwell time and the factors of user and task respectively. They examined whether establishing a threshold for dwell time with information about the user and/or the task to predict document usefulness could improve the personalization performance without considering information of the user or the tasks. They found that the performance of implicit relevance feedback using dwell time as an indicator of relevance was improved when the task information was considered.

Kellar, Watters, Duffy and Shepherd (2004) examined the relationship between task type, reading time and documents' relevance, and their results showed that the performance of reading time as an indicator of relevance varied in different types of tasks. In particular, they found that reading time is a reliable indicator of relevance on complex tasks when users only judge general relevance, and is not good on simple tasks when a user wants to find a specific answer. Liu and Belkin (2010) examined the effects of search stage and usefulness on decision time. Their results showed that users spent the longest time on very useful pages in early search stages, but spent the shortest time in later stages. These studies demonstrate that understanding how behaviors change with respect to contextual factors (task type, stage of search, topical knowledge, etc.) can improve the effectiveness of using dwell time for IRF.

Goecks and Shavlick (1999) examined whether **the amount of page activity** was related to a user's interest. They measured users' mouse and scroll activities as well as browsing behaviors. While they found correlations between a high degree of page activities and users' interest, they did not test this relationship against explicit judgments by the real users. Claypool, et al. (2001) analyzed some user behaviors as implicit indicators and compared them with the explicit ratings. They developed an experimental browser, the Curious Browser, to collect users' behaviors on the Web pages they visited. They found that the time spent on a page and the amount of scrolling on a page had a strong positive relationship with explicit interest, while individual scrolling methods and mouse-clicks were not correlated with explicit interest.

2) Query logs

Query logs have also been used to disambiguate users' current queries. The query logs include not only previous queries of the current searcher, but also the click-through behaviors following the queries. Shen, Tan and Zhai (2005) explored methods to construct users' query language models using users' immediate search context and feedback, i.e. the query history and click-through history in the same search session, and found the performance of later queries could be improved using such methods without any cost to the user. They further (Shen, Tan and Zhai, 2006) conducted a user study to examine personalization on two types of queries: recurring queries and fresh queries, using different lengths of search history. Their results showed that recent history tended to be much more useful than remote history, especially for fresh queries; while the entire history was helpful for improving the search accuracy of recurring queries. In addition, such results also match the ostensive model (proposed by Campbell and van Rijsbergen, 1996), which adds a temporal dimension to relevance, suggesting a recently viewed step in the relevance path is more indicative of the current information need than a previously viewed one.

Some other studies that used query logs to do personalization often involve mapping queries to categories using users' local profile and/or external classification files like ODP (Open Directory Project) or WordNet, and then extract related terms for query expansion. Liu, Yu and

Meng (2002) proposed to personalize users' queries by mapping their queries into categories to disambiguate the query terms. They examined three methods of mapping queries to categories: using user profile only, using general profile only, and using both user and general profiles. Their results indicate that the accuracy of using both profiles is consistently better than those using any one alone. Bai et al. (2007) used a method of a combination of two types of contexts to select appropriate terms for query expansion. The contexts they used included the topical domain of interest of the query using ODP as context around query, and the semantic term relationship within each query as context within query. Their results demonstrated that such combined method brought significant improvements in retrieval effectiveness.

In addition, previous queries that were used for searching the same or similar as the current information needs can also be used a source for query expansion or query suggestions. Huang, Chien and Oyang (2003) proposed to extract relevant terms for query expansion in a similar search session on a search engine. By conducting a user study, Kelly, Gyllstrom and Bailey (2009) provided query suggestions based on the query logs generated by a separate group of users who were searching for the same search task. Their results revealed that searchers used queries generated by other users for the same task more frequently than queries generated by systems. Methods of suggesting other searchers' queries on similar topics are also used by current major commercial search engines, like Google, Yahoo, Bing, etc., or retrieval systems like the ACM digital library.

Teevan, Dumais & Horvitz (2010) examined users' clicks on search results and explicit judgments for the same queries and found people's explicit judgments for the same queries differed greatly from one another. They proposed that the *potential for personalization* could be defined by the gap between how well search engines could perform if they were to tailor results to the individual, and how well they currently perform by returning results designed to satisfy everyone. Their results demonstrated that the variability in clicks on search results for the same query might indicate the potential benefits for personalization of the search results.

3) Behaviors on search result page

Besides interaction behaviors on content pages (e.g. dwell time), users' behaviors on search result pages have also been found to be effective evidence for implicit relevance feedback. For example, click-through, the selection behavior on search result page, has received a lot of attention because this behavior is easy to obtain from proxy servers and is logged by most search engines. Joachims and his colleagues have used click-through data to re-rank search results (Joachims, 2002). However, click-through for implicit feedback has been critiqued by researchers for its problematic assumption that all the pages requested by the user are equally useful to the user. Recent studies have tried to combine other behaviors together with click-through behaviors to improve its performance. Radlinski and Joachims (2005) proposed to infer users' relative relevance judgments of documents based on the click order in one query and within multiple queries in one search session, and got better performance than no relevance feedback. Fox, et al. (2005) analyzed the association between explicit ratings of user satisfaction and implicit measures of user interest using a Bayesian modeling method. Their results revealed that a combination of implicit measures could better predict user satisfaction than using click-through behaviors alone. In addition, they found time, including time from when the user left the result list to the time he/she returned, and time during which a page was in focus, and exit type (e.g. kill browser window, new query, URL entry and etc.) were the two best predictors of satisfaction.

Liu, Gwizdka and Liu (2010) examined users' search behaviors during query reformulation intervals (QRIs) to identify whether users found useful documents during that interval. They examined this issue by comparing behavioral variables that characterized QRIs during which useful pages were found with those during which no useful pages were found. Their results demonstrated that the QRI duration and the total time spent on content pages during QRIs with useful pages was significantly longer than during QRIs with no useful pages. In addition, users were also found to have viewed more content pages and spent more time on content pages than on search result pages during QRIs with useful pages. This study did not only focus on user behaviors on single search result lists, but on a series of search result lists after each query that users have issued in a search task episode.

When investigating users' query reformulation behaviors (QRBs) in a controlled user study, Liu, Gwizdka, Liu, Xu and Belkin (2010) found that after visiting and saving useful page(s), Generalization as a QRB was less likely to be used while New query as a QRB was more likely to be used. It was also found that the query reformulation types after finding useful page(s) might vary in different types of tasks. For example, in Simple and Hierarchical tasks, Specialization was mostly used after saving useful pages; while in Parallel tasks, Word Substitution was used most frequently after saving useful pages. Such results indicate that how users reformulate their queries when searching for a search task may help infer whether they have found useful pages after their previous queries.

Agichtein, et al. (2006) proposed a user behavioral model to predict web search result preferences by observations of searchers' natural interactions with a commercial search engine. Different from Fox et al. (2005), which was in general query-independent, this study examined query features, which are query-specific. The behaviors their model considered included: query features, browsing features, and click-through features. Query features included query length, fraction of shared words between query and title, summary, URL, and domain, and the overlap between two adjacent queries. Browsing features were used to characterize interactions with pages beyond the results page, such as dwell time and number of clicks on each document. Click-through features include result ranking, click frequency, and whether there is a click on the next or the previous result. Their results indicated that considering the number of clicks in the result list and the position of clicks in the result list together with click-through behaviors could improve the prediction of users' preferences. These studies indicated that a combination of several behaviors as evidence for implicit feedback might perform better than methods using a single type of behavioral evidence.

4) Combination of multiple behaviors

Using single user behaviors to predict user's preference seems insufficient, and some studies have extended such research by combining multiple sources of evidence for implicit feedback, or forming relative indicators by combining several behaviors. Shapira, Taieb-Maimon and Moskowitz (2006) suggested six new implicit indicators, and among them, four were relative measures: mouse movement relative to reading time, scrolling time relative to reading time, reading time normalized by page size, number of links visited on a page relative to the number of existing links on the page; and the other two were single indicators: number of links visited on a page, and level of interaction on a page. All these indicators of users' behaviors on individual content pages were captured during their everyday searching by using a specially developed browser, and participants were also asked to explicitly rate the relevance of each page using 5-scale points. The results of this study showed that the best single indicator is "the relative mouse movement", which was more accurate than reading time or mouse movement, in distinguishing different levels of interests on documents. With respect to the combination of multiple implicit indicators, they used stepwise linear regression to obtain a regression model containing the "best" subset of indicators. However, this paper did not mention what behaviors were included in this "best" subset.

Another study that examined a combination of multiple user behaviors for implicit relevance feedback was conducted by Yang, Xiang and Shi (2009). With a focus on user behaviors on small screen devices, they examined the relationship between user behaviors from four aspects (display time, viewing information items, scrolling and link selection) and user's interest on blocks. They conducted a statistical analysis first to identify the most significant types of implicit evidence, and then used machine learning methods to identify user's interest blocks by combining this evidence. They also compared three classical machine learning techniques: Support Vector Machine (SVM), C4.4 and Naïve Bayesian Method. Their results demonstrated that the aspect of viewing information items was less indicative than other aspects of behavior; and with respect to the effectiveness of machine learning techniques, they found Support Vector Machine (SVM) was the best, because Naïve Bayesian Method always performed worst and it always assumed that indicators were independent of each other, but this was not the case; and C4.4 had the drawback of over-fitting. They also found the usefulness of significant implicit evidence was influenced by users to a great extent, while the influence of the type of Websites on the usefulness of significant implicit evidence was not as great as that of users.

Melucci and White (2007) also investigated multiple aspects of user behaviors to represent the features of documents. The method they used was the Vector-Space Model (VSM), and each document was a vector, the relevance was a linear transformation, relevance statuses were the eigenvalues of the linear transformation, and the computation of the probability of relevance of a document as the projection of the document vector onto an eigenvector of the linear operator. Their results demonstrated that implicit relevance feedback was more effective when it was tailored to the task and personalized to the user. However, this study did not explore which behaviors and what combinations of behaviors were significant for the prediction of document usefulness.

5) Summary of behavioral evidence as implicit measures

The above reviewed studies on behavioral evidence as implicit measures of users' interest and preference can be summarized in the following table. In this table, users' observable behaviors are modeled using two axes, *Search stage* and *Signal Type*.

Table 1 Classification of behavioral signals as implicit relevance feedback and related literature

		Signal Type		
		Attention	Action	Content-based
Search stage	Before they search			Previous queries: Beeferman and Berger (2000); Liu, Yu and Meng (2002); Shen and Zhai (2003); Huang, Chien and Oyang (2003)
	On search result pages	Time on result list before first click: Fox, et al. (2005); Agichtein, et al. (2006); Total time on result lists: Fox, et al. (2005) Query Reformulation Interval time: Liu, Gwizdka and Liu (2010)	Issue query: Agichtein, et al. (2006); Clickthrough: Radlinski and Joachims (2005); Joachims et al. (2007); Fox, et al. (2005); Agichtein, et al. (2006)	Query content features: Agichtein, et al. (2006);

			Click order: Radlinski and Joachims (2005); Agichtein, et al. (2006) Click position: Agichtein, et al. (2006) Number of clicks: Agichtein, et al. (2006) The way in which the user exited the page: Fox, et al. (2005)	
	On content pages	Display (dwell) time: Morita and Shinoda (1994); Konstan et al. (1997); Oard and Kim (1998); Kelly and Belkin (2004); Fox, et al. (2005); Agichtein, et al. (2006); White and Kelly (2006) First dwell time: Liu and Belkin (2010) Eye movement: Joachims et al. (2007)	Scroll: Fox, et al. (2005) Mouse movement & clicks: Fox, et al. (2005) Number of visits: Fox, et al. (2005)	

	Further use of content pages		Print; Bookmark; Save; Delete; Add to favorites; Email; Cite; Rate; Edit; etc.
			Oard and Kim (1998);
			Fox, et al. (2005);	

Search stage (before they search, on search result pages, on content pages, further use of content pages, exit type of content pages) refers to the search behaviors in different search stages during the interactive search process. Before they conduct the current search, users' previous query log could represent their long-term and short-term interests. On search interfaces and search result pages, users could issue their queries to search; when they get the search result, their main decision is which search results to click, so related measures as click-through, time before click, click order, etc. can serve as implicit evidence of users' preferences. On each individual content page, their behaviors are mainly investigating the content of the page. After reading the content page, users decide whether to further use the content pages if they are useful or relevant.

Signal type (Attention, Action, and Content based) refers to the type of behavioral signals as implicit evidence for relevance feedback. Attention describes users' attention on the page, e.g. the dwell time they spend on the page, or the eye-movements on the page. Action refers to user' movements that can be observed during search, mainly including mouse and keyboard activities. Content-based signal mainly refers to the content of queries users issue during search, and analysis of the content match between queries and content pages.

This classification highlights users' behaviors on search result pages. Although the focus of implicit relevance feedback is to predict the usefulness or relevance of individual documents, users start examining these documents on search result pages, and by deciding to click on the document, they can browse the content of certain documents. Therefore, users' behaviors through the whole search session can serve as evidence for implicit relevance feedback.

2.1.2 Discussion of research on implicit evidence for relevance feedback.

This section reviewed implicit evidence that has been investigated in previous studies to personalize the content of search results for individual users. It also demonstrates some challenges for future research using implicit evidence.

First of all, with respect to behavioral evidence, we need further investigation on how to interpret behavioral evidence in different contexts. Behaviors, such as dwell time on content pages, were found not to be consistently reliable as indicators of preferences in all situations. For instance, Kellar et al. (2004) found that dwell time was a good indicator of preferences only for complex tasks, and was not good for simple tasks. One reason for this is that the behaviors that are considered for IRF are also influenced by contextual factors (task topic, type, topic knowledge and so on), and their ability to distinguish relevant and non-relevant documents varies in different contexts. Research on information-seeking behaviors could help find out how users' behaviors change with respect to contextual factors and may also help identify the best behavioral evidence in particular situations. Related work addressing this issue is reviewed in detail in the next section.

Secondly, there are differences between client-logged versus server-logged studies. In previous studies we can find that, studies about dwell time or other behaviors on content pages were mostly user studies, either in naturalistic settings or laboratory settings, and users' behaviors were logged on the client; while studies on click-through behaviors and other behaviors on result lists were mainly using search engine logs collected by the server. The advantage of the server log analysis is that it is easier to obtain by search engines and search engine logs often consist of large quantities of data; however, the drawback of such methods is that they lack the context of search, and lack data on users' interactions with documents. In contrast, client-logged data can provide more contexts and users' complete interactions with information objects during their search. But not much research on client-logged analysis has examined the behaviors on search result lists and/or combined these behaviors with dwell time or other interactions on content pages for implicit feedback. This calls for future studies to investigate users' behaviors on both search result pages and individual documents, and their relationships to the explicit ratings of interest and preference.

Another challenge for behavioral evidence is how to use multiple behavioral sources to predict document usefulness. Studies have demonstrated that a combination of behaviors on search result pages and content pages could improve the performance of prediction over that using any single behavioral evidence (e.g. Fox, et al., 2005; Agichtein, et al., 2006). However, as behaviors on search result pages are not directly related to the relevance of any single content page, further studies are needed to explore what we could infer from users' behaviors on result pages, and how to apply behaviors on search result pages and content pages in the prediction of document usefulness.

2.2 Contextual factors in information retrieval research

Contextual factors are important in affecting users' information behaviors, as well as the prediction of users' preferences and interests from implicit relevance feedback. We need to understand what contextual factors influence users' behaviors and users' document preferences; we also need to identify significant contextual factors without explicitly eliciting them, and then incorporate them in the design of information retrieval systems. Identification of contextual factors is another facet to personalizing users' search experiences. This section of the review focuses on studies about the relationship between contextual factors and user behaviors.

Cool and Spink (2002) identified four levels of context: information environment level—the social and environmental factors that influence human information behaviors, e.g. location and time; information seeking level—the goals and tasks that information seekers are trying to accomplish and how they influence their seeking behaviors; IR interaction level—the interaction between user and system within search sessions; and query level—linguistic context by attempting to understand, or disambiguate, users' context of meaning when they use a particular query term. Context on the information environment level is beyond the scope of this dissertation. Context on the query level has been discussed above in the users' preferences learning section. So this section focuses on context on information seeking level and IR interaction level. Research at these levels tries to understand how information goals and task influence users' behaviors, how behaviors within these contexts differ and how successive information seeking episodes represent contexts. The contextual factors reviewed below mainly include task type, topical familiarity and domain knowledge and individual differences.

2.2.1 Task type effect on information search behaviors

Tasks are defined as a motivation of information seeking and search. It has been shown by many studies that task type has an influence on users' search behaviors. In order to examine the effect of tasks, several ways to classify task types have been identified. Task types are often examined as independent variables in information seeking and retrieval research.

In examination of information seeking strategies of novices, Marchionini (1989) designed two types of tasks, i.e., **a closed task** and **an open-ended task**. The closed task required students to find three facets of a fact, and there is only one correct answer; the open-ended task required users to find information about one subject, and there may exist many related facts. These two types of tasks were designed to vary according to task difficulty, with the hypothesis that one task would take more time and be considered more difficult by users, and that search strategies would differ in the two types of task. It was found that users spent more time and performed more moves for the open-ended task than for the closed task. Qiu (1993) used two types of tasks to investigate the effect of task type on search strategies: **specific task** and **general task**. In the specific search task, users were asked to search for a specific fact that was known to exist; this type of task is very similar to the closed task as defined by Marchionini (1989). In the general search task, users searched for general information about a broad topic, and it is somewhat related to Marchionini's open-ended tasks. Qiu found that users preferred to use browsing for completing the general task, whereas they used analytical search strategies more frequently in the specific task. Kim (2001) examined the influence of cognitive style, online database search experience, and task type on users' search behaviors. The two types of tasks are **Known-item search task** and **Subject search task**. Kim pointed out that Known-item and Subject search tasks correspond to the closed and open-ended tasks of Marchionini (1989) and the specific and general tasks of Qiu (1993). For the Known-item task, there was one piece of target information; the Subject search is defined as a task requiring the searcher to retrieve information that is related to a given subject.

The task classifications in these three studies are closely related, and they are designed in such way as to differentiate the task difficulty of the search tasks. The common feature is that closed/specific/known-item search tasks only require users to find one specific fact, while open-ended/general/subject search tasks require users to find multiple unspecified facts or general information about a topic. Such classifications focus on the type of information users are searching for in the task, whether it is one fact or multiple facts or information. However, there are many other dimensions which could help define task type.

Li (2008) proposed a faceted classification of tasks, including work task and search tasks, in which she identified multiple dimensions that could define task type. She also conducted a semi-structured interview to validate this task classification. This classification of tasks not only considers the generic facets of tasks, but also includes common attributes of tasks. The 'generic facets of tasks' are external characteristics of tasks, including Source of task, Task doer, Time, Process, Product, and Goal. The common attributes describe the internal attributes of the task, including task complexity (subjective and objective), task difficulty, task interdependence, degree of structure of task, salience of task, degree of urgency of task, and knowledge of task. These attributes are labeled as 'Common attributes of task' because each type of task classified based on 'Generic facets of task' can be described by all of these attributes. In the following section, we introduce several dimensions in Li (2008)'s classification scheme: task product, task complexity and task goal (quality), and also discussed other related studies involving these task types. Another facet, Level of document judgment, which was not contained in Li's classification, but has been found to be significant in affecting users' search behaviors is also discussed.

1) Task product

Product as a facet of tasks refers to the outcomes or results of task completion. Many previously mentioned studies, including Marchionini (1989), Qiu (1993) and Kim (2001), classified tasks with respect to the task product facet. According to Li (2008), for work task, task product includes Physical (a task which produces a physical product), Intellectual (a task which produces new ideas or findings), and Decision/Solution (a task which involves decision making or problem solving); for search tasks, task product can be classified into Intellectual, Factual information (a task locating facts, data, or other similar information items in information systems), Image (a task locating images in information systems), and Mixed product (a task locating different types of information items in information systems). Li (2008) further investigated the relationship between work tasks and interactive information searching behaviors by conducting an experimental study. One facet that she examined in the experiment is task product. In the task design, she selected two types of work task product: Intellectual and

Decision/Solution tasks. The results indicated that intellectual tasks involved more retrieval systems consulted and result pages viewed, longer query length and higher self-rated success. This study demonstrated that work tasks with different products could shape users' search behaviors.

Freund (2008) classified information tasks into five types based on the user's intended use of information: learning about a topic, fact-finding, making a decision, solving a problem, and finding out how to do something. The purpose of making such classification is to examine whether users expect different types of genre for their work tasks. When examining carefully, we can see this classification is very similar to Li (2008)'s task product facet classification. In Freund (2008), "learning about a topic" is a type of task when users try to learn about an unfamiliar topic, and seek general orientation and an understanding of concepts, which is similar to "Intellectual"; "making a decision" is to identify and compare alternatives in order to determine a course of action, which corresponds to "Decision task"; "finding facts" is to find specific factual information, which is "factual information"; "finding a solution" is to solve a problem by finding information on similar scenarios; "finding out how to" is to find a procedure or work plan identifying the steps to take and issues involved, the latter two could refer to "Solution task". In this study, it is found that users expect different types of information object when completing different types of tasks.

Kim (2009) also classified search tasks with respect to the type of answer users are searching for in the task, and she identified three types of tasks: 1) factual task: an "asking a fact" task, such as naming, identifying or listing; 2) interpretive task: a task to configure an answer rather than simply and concisely locate one, which is rather open-ended but goal-oriented; 3) exploratory task: a task motivated by the searcher's desire to broaden his or her knowledge of a topic, which is completely open-ended and vaguely structured. In this classification, both the factual and interpretive tasks are goal-oriented and ask specific questions, but the factual task expects specific facts as the answer, while the interpretive task expects general and multiple objects as the answer. The experimental results demonstrated that in factual tasks, users' typical strategy was typing specific keywords in the search engine, scrolling through the results, opening one result, scanning a page, and then coming back to the list of results until they found the target information; in interpretive tasks, users' search pattern proceeded from the general to the particular; in exploratory tasks, users often stopped their searching when they found a page that had lots of links and frequently planned on a later use of information.

Murdock et al. (2006) focused on information search in Question and Answering systems, with the purpose of identifying and improving the procedural questions. The focus of this study was to compare the differences in question description and non-content features of relevant documents in two types of questions: the factual questions which ask for a statement of fact, and the procedural questions which ask for a description of a process. They identified several features that differ significantly between the relevant set and the non-relevant set of documents for procedural questions. They clustered documents based on the documents' structural similarity and then re-ranked the documents based on the results of clustering. Their results showed an improved performance of ranking using this method. This study indicated that the identification of task type and the non-content features of documents needed for different types of tasks are important in improving users' search experience.

Kellar, Watters, and Shepherd (2007) investigated information seeking behavior on the Web through a field study, and identified four task categories: fact finding (FF), information gathering (IG), just browsing, and transaction. Among these, only FF and IG are goal-driven information seeking tasks. FF is defined as a task in which users are looking for specific facts or pieces of information; IG involves the collection of information, often from multiple sources. These two task types are similar to Kim (2009)'s classification, in that FF is equivalent to factual task, and IG includes both interpretive and exploratory tasks. Kellar, Watters, and Shepherd (2007)'s results show the differences between these two tasks are duration time (FF is longer than IG), number of pages viewed (IG has more than FF), query length (FF has longer queries, IG has shorter queries).

Toms et al. (2008) examined how search behavior differs according to three types of task information goals and two types of task structure. The three types of task are decision making, fact finding and information gathering; and the two types of task structure are hierarchical and parallel. Their experiment found that users formulated fewer queries but took more time to process the result of a query for 'Hierarchical tasks' than for 'Parallel tasks', and it suggested that 'Hierarchical tasks' required more effort considering most metrics other than number of queries. As to the task type, 'Decision Making' and 'Fact Finding' tasks contained more queries than 'Information Gathering'; in addition, those in 'Information Gathering tasks' were more likely to add additional, unprompted terminology than those in 'Decision Making'.

A recent study by Liu et al. (2010) investigated the relationship between user behaviors and different task types. They designed four journalism assignments, varying according to three facets in Li (2008): task product, task complexity and search task goal (quality), and an extended facet, Level of document judgment. With respect to task product, they designed two types of tasks: Factual and Mixed. It was found that users spent significantly longer completion time and visited significantly more pages and more sources in Mixed tasks than in Factual tasks.

2) Task complexity

Besides task product, another important facet of task is objective task complexity. In Li (2008), objective task complexity reflects the number of Information Retrieval systems and types of sources used in accomplishing the work task, and it contains three levels: high, medium, and low. In her controlled study, she found that task complexity affected a number of aspects of information search behaviors, and it was associated with the total completion time, the number of retrieval systems consulted, the number of pages viewed, the number of queries, and the number of unique queries.

Complexity has also been examined by many other researchers, but the definition of complexity is different from Li (2008), and it is often related to task difficulty. For example, Byström and Järvelin (1995) investigated the relationship between task complexity, information types, and information sources. In this study, task complexity is defined in terms of "a priori determinability of, or uncertainty about, task outcomes, process and information requirements" (p.194). In this study, tasks are classified into five categories: automatic information processing tasks, normal information processing tasks, normal decision tasks, known, genuine decision tasks, and genuine decision tasks. They found that when the task complexity increased, more types of information were needed, and people were less likely to predict the information type they needed and relied more on experts to provide useful information. Such a definition of task complexity is related to users' pre-familiarity of the task; the more knowledge of the task the user has, the less uncertainty he/she may have in the searching. This definition of task complexity is explicitly subjective, and the same task may not be classified as the same type for different users, depending on their familiarity with the task. A definition of task complexity which is not individual-dependent, such as Li's, might be more useful for experimental purposes.

3) Task Goal (quality)

In Li's classification, Search goal contains two sub-facets: quality and quantity. Search goal (quantity) describes the number of sub-goals contained in the task. Search goal (quality) includes three values: specific goal, amorphous goal, and mixed goal. A task with specific goal is defined as "a task with explicit or concrete goals"; a task with amorphous goal refers to "a task with abstract goals"; and a task with a mixed goal is "a task with both concrete and abstract goals". This facet of task goal quality was first proposed in MacMullin and Taylor (1984), in which they identified 11 dimensions to define the attributes of user's work problem situation, including such categories as "design/discovery", "well/ill structured", "complex/simple", "specific/amorphous goal" and so on. With respect to the goal (quality), their definition is "specific goals are easy to be operationalized and measurable while amorphous goals are ambiguous and hard to be operationalized". In Li's thesis, the example of a task with a specific goal is "to get the grade for the class"; and the example of a task with an amorphous goal is "preparing exams", and the search task involves searching textbook about biology to get a better understanding of the biology issues; if a task contains both specific and amorphous goals, it is labeled as a "task with a mixed goal".

The effect of task goal (quality) has been investigated in Liu et al. (2010), in which they designed four tasks varying three types of task goal (quality): specific, amorphous and mixed. In the specific task, the goal of the task is locating documents on a well-defined topic or confirming facts; while the amorphous goal requires users to find some experts on a given issue, which is very abstract and should be determined by the users. The task with mixed goals contains a specific goal and an amorphous goal. The results of this study showed that goal (quality) affected the number of sources (i.e. the unique Internet domains visited by the user in a task), average decision time (i.e. the time taken during the search process to decide whether a document is useful), and the ratio of reading to scanning. In particular, users had significantly longer decision time for the specific task than for the amorphous task and the mixed task. Their results demonstrated that task goal (quality) does not have a strong effect on whole-session level behaviors, e.g. completion time, total number of pages and etc., but it did influence some within-session level behaviors, especially document examination, e.g. decision time. Although not many studies have examined this facet of task, it seems to be an important feature of task that could influence users' search behaviors, especially within-session behaviors.

4) Level of document judgment

The facet "level of document judgment" was proposed in Liu et al. (2010), and it includes two values: segment and document. Segment level tasks require locating specific information within a page, while document level tasks only require users to judge if a page is useful or relevant in general but do not necessarily require locating specific information.

When examining the dwell time on individual pages as evidence for implicit relevance feedback, Kellar, Watters, Duffy, and Shepherd (2004) conducted an experiment using three types of tasks: relevance judgment, simple question answering, and complex question answering. In relevance judgment tasks, users were asked to judge five news documents on a given topic; in simple question answering tasks, users were asked to find a fact in five given documents, when only one document contained the answer; the complex question answering tasks they designed were similar to simple question answering tasks, but required users to find multiple facts in the question, and the answer was contained in only one document. Even though the tasks in this experiment are not information search tasks, they can be differentiated with respect to the "Level of document judgment". In relevance judgment tasks, users only need to judge whether each article is about the given topic, and do not need to locate any specific fact in the document, so they can be labeled as "document level"; while in the question answering tasks, users need to read or scan the document carefully to locate the specific fact(s), so they can be labeled as "segment level". The results of this study showed that in relevance judgment tasks ("document level"), participants spent similar time on relevant and non-relevant documents, whereas in question answering tasks ("segment level"), participants spent significantly more time on documents that contain the answer for the question than those that do not contain the answer. In addition, participants took a longer time to read documents in complex question answering tasks, regardless of the relevance, when compared with the other two types of tasks. This result demonstrates that reading time on documents is a good indicator of document relevance when users need to find a specific answer on the documents (judge the relevance on segment), and it is not a good indicator in general relevance judgment tasks (document level).

In Liu et al. (2010), users' search behaviors in tasks with segment level and document level were compared. In the task with segment level, users are asked to find an authoritative page that either confirms or disconfirms the statement in the assignment. Therefore, users need to locate the specific piece of fact in the document to judge the usefulness of a page. In the tasks with document level, users only need to judge whether the page they find is relevant to the given topic or not. This study examined multiple search behaviors, and found the facet of “level of document judgment” showed significant effects on all the examined behaviors, including task completion time, number of pages visited, number of queries, the average decision time, the ratio of reading to scanning and so on. Compared with Kellar, et al. (2004), they have similar results on the effect of “Level” on decision time, i.e. users spent significantly longer time on deciding whether a document is useful or not for tasks with segment level than tasks with document level. However, this study did not compared the decision time, or dwell time on useful documents and non-useful documents, and it is also unclear whether “Level of document judgment” would affect users’ other within-session behaviors.

5) Task type prediction

There have been some studies on task type prediction, and many of them were based on Broder’s (2002) task classification. Broder (2002) classified searching tasks into three types: navigational, informational and transactional. Studies on identifying task types of this classification were through analyzing users’ post-query behaviors; for example, Rose and Levinson (2004), Lee, Liu, and Cho (2005), and Chang, He, Yu and Lu (2006). The results of these empirical studies indicate that most tasks by real users (over 60 percent in Rose and Levinson (2004)’s study) are informational tasks. Many information search tasks that we discussed above are informational tasks, and studies on the prediction of task types or task facets from users’ search behaviors are limited nowadays.

A number of studies have tried to identify search success or failures by observing users’ several or sequential within-session behaviors. Hassan, Jones and Klinkner (2010) used a supervised Markov model to predict the success of user search goal from the user’s sequential search behaviors, including the sequence of all queries and clicks in one search session as well as the time between actions. Their model indicated that users in successful search tend to spend longer transition time between a search result click and a query submission than unsuccessful search; unsuccessful sessions were more likely to end with an abandoned query, which contained no clicks after a query. Aula, Khan and Guan (2010) studied potential behavioral signals that suggested that a user was having trouble in a search task. They first conducted a laboratory study with 23 users to identify potential behavioral signals when users struggle to find the information they are looking for. Then they tested their previous finding in a large-scale study. Their results showed that when having difficulty in searching, users started to formulate more diverse queries, used advanced operators, and spent longer time on search result pages as compared to the successful tasks. These studies indicate that within-session behaviors could reflect some aspects of search contexts.

Liu, Gwizdka, Liu, and Belkin (2010) investigated the relationship between users' behaviors and task difficulty, and whether it varied in different types of tasks. With respect to user behaviors, they divided all behavioral signals into two categories: whole-task-session level and within-task-session level. The three types of search tasks are: single-fact finding, multiple-fact-finding and multiple-piece information gathering. The results of this study showed that task type affected the relationship between task difficulty and user behaviors; that is, behavioral variables showing significant differences between difficult and easy tasks were different across task types. In addition, the results also demonstrated that both whole-session level and within-session level behaviors can help predict task difficulty. While whole-session level variables showed higher prediction accuracy, within-session level behaviors had the advantage of enabling real-time prediction. The within-session level behavioral variables examined in this study include:

Number of pages per query, *Number of unique pages per query*, *First dwell time on content pages*, *Mean dwell time of all content pages*, *First dwell time on Search Result Pages (SERPs)*, and *Mean dwell time of all SERPs*.

Currently, many studies on task type prediction focus on task difficulty or search failures, or general task type (like informational or navigational tasks). Studies on task type effect have focused on user behaviors on the whole-session level, rather than within-session level, which cannot contribute much to the real-time prediction of task type. More studies are called for in the future to examine how task type or facets of task features influence within-session level behaviors, and whether such behaviors could help predict task type.

6) Summary of behavioral measures as task type predictors

The above review shows that task type has an important effect on user behaviors, and the behaviors that have been examined are mainly whole-session level behaviors. A few recent studies (e.g. Hassan, Jones and Klinkner, 2010; Aula, Khan and Guan, 2010) on task type learning demonstrate that user behaviors on search result pages might shed light on the task type. Therefore, behaviors on both whole-session level and within-session level can be indicators for predicting task type. These behavioral measures are summarized in Table 2.

Table 2 Classification of behavioral measures in information search tasks
(from the literature review)

	Whole-session level	Within-session level

Query-related	Number of queries;	Query reformulation type;
		Query reformulation interval time;
		Query reformulation speed;

SERP-related	Total dwell time on unique SERPs;
	Number of SERPs;
	Number of unique SERPs;
		First dwell time on unique SERPs;
		Accumulated average dwell time of SERPs to the current time point;
		Time before first click on unique SERPs;
		Number of clicks on unique SERPs;
		Click speed on SERPs;

Content page-related	Total dwell time on unique content pages; Number of content pages; Number of unique content pages;	First dwell time on unique content pages; Mean dwell time of all content pages till this point; Number of mouse-clicks on each content page; Amount of scrolling on each content page; Usage of "find" on each content page; Number of revisits to unique content pages;
Query and page related	Number of queries with no useful pages; Number of queries with useful pages; Ratio of queries with no useful pages to all queries; Ratio of queries with useful pages to all queries	Number of content pages per query; Number of unique content pages per query; (Note: these measures are only applicable to search sessions with more than one queries)
Others	Task completion time	

2.2.2 Users' knowledge effect on information search behaviors

Users' knowledge has been identified as another important contextual factor that can influence users search behaviors. Users' knowledge includes domain knowledge, topic familiarity and search knowledge.

Wildemuth (2004) investigated the effect of domain knowledge on search tactics, i.e., the changes of search terms and concepts represented by terms. In this study, participants came three times for searching, with each session representing an increasing level of domain knowledge. It was found that when domain knowledge was low, users did more moves per search, selected less efficient concepts in the search and made more errors in the reformulation; while when their domain knowledge was at the peak, they incorporated multiple concepts in their searches but made fewer changes to their searches. White, Dumais and Teevan (2009) also examined the effect of domain expertise on web search behavior in four different domains (medicine, finance, law and computer science) based on large-scale log-based data. They found experts generated more technically-sophisticated and longer queries than non-experts; experts spent more time and visited more pages than non-experts.

Topic familiarity has also been shown to have influence on users' search behaviors. For example, Kelly and Cool (2002) examined the effect of topic familiarity on users' information search behaviors and found that when one's familiarity with search topics increased, his/her reading time decreased while search efficacy (the ratio of saved documents to total viewed documents) increased.

The stage in the search can also reflect users' relative topic familiarity, since users' knowledge about a topic increase as they go through stages of searching. White, Ruthven and Jose (2005) examined the effect of search stages on the utility of implicit relevance feedback (IRF) and explicit relevance feedback (ERF) in two separate systems. Their results demonstrate that in their IRF system, IRF is used more in the middle of the search than at the beginning or end, whereas in their ERF system, ERF is used more towards the end. Liu and Belkin (2010) examined the interaction effect of search stage and topic knowledge on the prediction of document usefulness based on dwell time on content pages. They found the dwell time varies for useful documents in different search stages and with different levels of topic familiarity.

2.2.3 Other contextual factors

Besides task type and users' knowledge effect, other contextual factors include environment factors (such as location, time of search, devices on which information search is conducted, etc.) and individual differences (such as searchers' cognitive features, study approach, demographic features, etc.).

Context at the information environment level can be collected by the IP address or different kinds of devices that users are using to connect to the Internet. For example, many mobile devices can provide accurate geo-information about where users are, and the location information may help system to interpret users' search behavior in current situation.

Individual differences that have been investigated include cognitive style, study approach, gender, age, etc. Among these, cognitive style has been examined most extensively, and it was often examined with the interaction of other individual difference factors or task type factor. Cognitive style is defined as the individual's characteristic method of organizing and processing information (Goldstein & Blackman, 1978). Among different cognitive styles, field dependence (FD)/field independence (FI) is one of the most extensively researched approaches, because it reflects how well an individual is able to restructure information based on the use of salient cues and field arrangement (Weller, Repman, & Rooze, 1994). A study by Chen and Ford (1998) investigated the influence of cognitive ability on hypertext navigation. It was found that FD users made significantly greater use of the main menu, while FI users made more use of the relatively sequential Previous/Next Buttons. Wang, Hawk, and Tenopir (2000) examined the effect of cognitive ability on affective aspects, and found that FD students experienced more difficulty and confusion than FI students. Ford and Chen (2000)'s study showed that FD users tended to build a global picture with the hierarchical map when interacting with Web services.

Kim (2001) investigated how cognitive style, online database search experience and task type influenced users' search behavior on the Web. The results of this study showed that online database search experience had a strong impact on both search performance and navigational style. With respect to cognitive style, it was found that cognitive style had a significant influence on the user's information seeking behavior most for those with little or no online search experience. Although in general the FIs outperformed the FDs, this difference could be minimized when the FDs were equipped with substantial online search performance. The author also suggested that different layouts and structures of Web pages might help the FDs search and navigate the Web in a more efficient way.

Results from previous studies on cognitive styles and individual differences suggest that different cognitive style groups prefer different interface functionalities and display structures provided by the retrieval systems. Because this dissertation study only focuses on personalization on the content of search results, rather than the display functions or the presentation methods, these individual differences will not be investigated as major contextual factors.

2.3 The evaluation of personalization algorithms

After generating various personalization algorithms, researchers need to evaluate the performance of their algorithms in information retrieval systems. For algorithms to personalize search result content based on prediction of document preference, there are mainly two ways to evaluate the performance: one is to measure the prediction accuracy of document preferences from implicit relevance feedback by comparing with the explicit relevance judgment; the other is to measure the differences in the retrieval performance before and after implementing personalization algorithms.

2.3.1 Prediction accuracy

Some studies proposed methods to predict or identify relevant or user preferred documents, which are useful to personalize users' information retrieval process. In these studies, the goal is to predict users' preferences for the documents that have been returned from search systems to users. Therefore, the evaluation method they often adopted is the prediction accuracy; for example, Agichtein, Brill, Dumais and Ragno (2006), Dou et al. (2007) and Joachims et al. (2007) all used this measure in their evaluations.

As previously discussed, Agichtein, Brill, Dumais and Ragno (2006) used a user behavioral model to predict web search result preferences using both server side and client side logging to capture searchers' natural interactions with a commercial search engine. The behaviors they considered include: query features, browsing features, and clickthrough features. In order to evaluate the performance of their method, they compared the predicted preferences to the "correct" preferences derived from the explicit user relevance judgments. In particular, they computed the average recall (fraction of "correctly" predicted preferences among all explicit preferences) and precision (fraction of "correctly" predicted preferences among all predicted preferences) to measure the accuracy of prediction performance of their model. Their results show that considering the number of clicks in the result list and the position of clicks in the result list together with clickthrough behaviors had higher preference prediction accuracy over state-of-the-art clickthrough methods. Their method achieved precision of 0.648 and 0.717 at recall of 0.08, which was somewhat higher than precision than using other clickthrough methods.

Joachims et al. (2007) also used prediction accuracy to evaluate their personalization algorithms, which predicted document relevance based on click-based behaviors. In the evaluation part, they compared the agreement of implicit relevance judgments generated from their method with the explicit relevance judgments by users. The results indicated the implicit judgments from their method are reasonably accurate.

2.3.2 Personalization performance

Instead of measuring the accuracy of preference prediction, some studies have re-ranked the search results based on the inference of document relevance and then evaluated the change of the retrieval performance. For example, many studies used Discounted Cumulative Gain (DCG), which measures the usefulness, or gain, of a document based on its position in the result list, to examine the performance of their personalization algorithms.

Teevan, Dumais and Horvitz (2005) built models of uses' interests from both search-related information, such as previous queries and previously visited pages, and also the documents and emails the user has read and created. Then they conducted user studies to evaluate the performance of re-ranked search results based on the interest models they built. Users were asked to evaluate the relevance of documents and then the Discounted Cumulative Gain (DCG) was calculated for each personalization algorithm. They compared the results with several baselines in the evaluation of performance of each algorithm; and their results show that personalized results performed significantly better than explicit relevance feedback and the default ranking.

Chirita, Firan, and Nejdl (2006) used desktop documents to extract user specific information for information searching. At the stage of evaluation, they interviewed 15 subjects to evaluate the performance of their personalization algorithms. They asked subjects to choose 6 queries related to their everyday activities but with different features, and then asked them to assess the relevance of all top 10 URLs generated by 15 versions of the algorithms the system presented for each query. Then they evaluated the output quality in terms of Mean Average Precision (MAP) over the first 10 results, precision at the first 5 positions of the resulted ranking, as well as precision at the top 10 output rankings; and tested whether the improvement over the Google API output (the original results without query expansion) is statistically significant. Their results showed the three measures they used were consistent.

White and Kelly (2006) used three implicit relevance feedback (IRF) algorithms using display time: IRF personalized to users, IRF personalized to task information, and IRF personalized to both users and tasks; and then expanded initial queries to retrieve a new set of documents. In the evaluation, they compared the performance of the three methods with a baseline algorithm without any personalization to see how much improvement each algorithm got compared against a baseline algorithm with a single display time threshold across all subjects. Their results indicate that tailoring display time thresholds to the search task leads to an increase in the performance of IRF algorithms, but doing so based on user information worsens performance.

Kelly, Gyllstrom and Bailey (2009) compared query suggestions to term suggestions and examined differences between automatically generated suggestions and those generated by humans in a previous experiment. They conducted a user study in which participants were asked to search for some search tasks using a system with query suggestions. They used the number of documents saved and snDCG for performance evaluation. snDCG is the normalized session-based discounted cumulated gain. Their results showed that subjects who received user-generated suggestions saved more documents; the most were saved for query suggestions. snDCG showed that the best performance was achieved by those who received user-generated query suggestions and that there appeared to be an interaction with source of suggestions and suggestion type: those who received user-generated suggestions did better with query suggestions, while those with system-generated terms did better with term suggestions.

The goal of personalization is to improve search performance by each individual user, to help them accomplish their search tasks. A personalized IR system should first correctly predict users' information need, document preferences and search context, and then take such information to provide personalized search results to individual users. Therefore, we evaluated the prediction accuracy of our predictive models in this study and plan to evaluate the retrieval performance in future studies.

Chapter 3. Theoretical Framework

3.1 Research model

Personalization of the content of search results for individual users requires the IR systems to understand the user's information goal, the current search task or possible preferences of documents beyond the issued queries. Knowing which of the documents users viewed were useful to them is helpful in predicting which documents that users have not yet viewed will be useful for their current search task. Implicit relevance feedback is an important technique to obtain the user's preferences of documents by observing their search behaviors unobtrusively. As reviewed in Chapter 2, contextual factors have an interaction effect on the relationship between user behaviors and document usefulness, and they can also directly affect users' general search behaviors. Thus, it is necessary to understand the three way interactions among user's behaviors, contextual factors and document preferences. Figure 1 shows the theoretical model of this thesis, which illustrates the three important components of personalization in IR research: contextual factors, user behaviors and document usefulness.

On the top of the model is information goal, which leads the user to engage with the IR system when his/her own knowledge is insufficient to resolve the problematic situation. As shown in Figure 1, a user starts the search with an information goal, and there are multiple contextual factors influencing how he/she behaves during search. These contextual factors also influence how users judge useful documents for the information goal. In this model, the user's behaviors are the source of information for personalization, which could reflect some contextual factors that the search is undertaken, and could also indicate the possible useful document for the search. The highlighted elements of the model of Figure 1 are those which are considered in this dissertation.

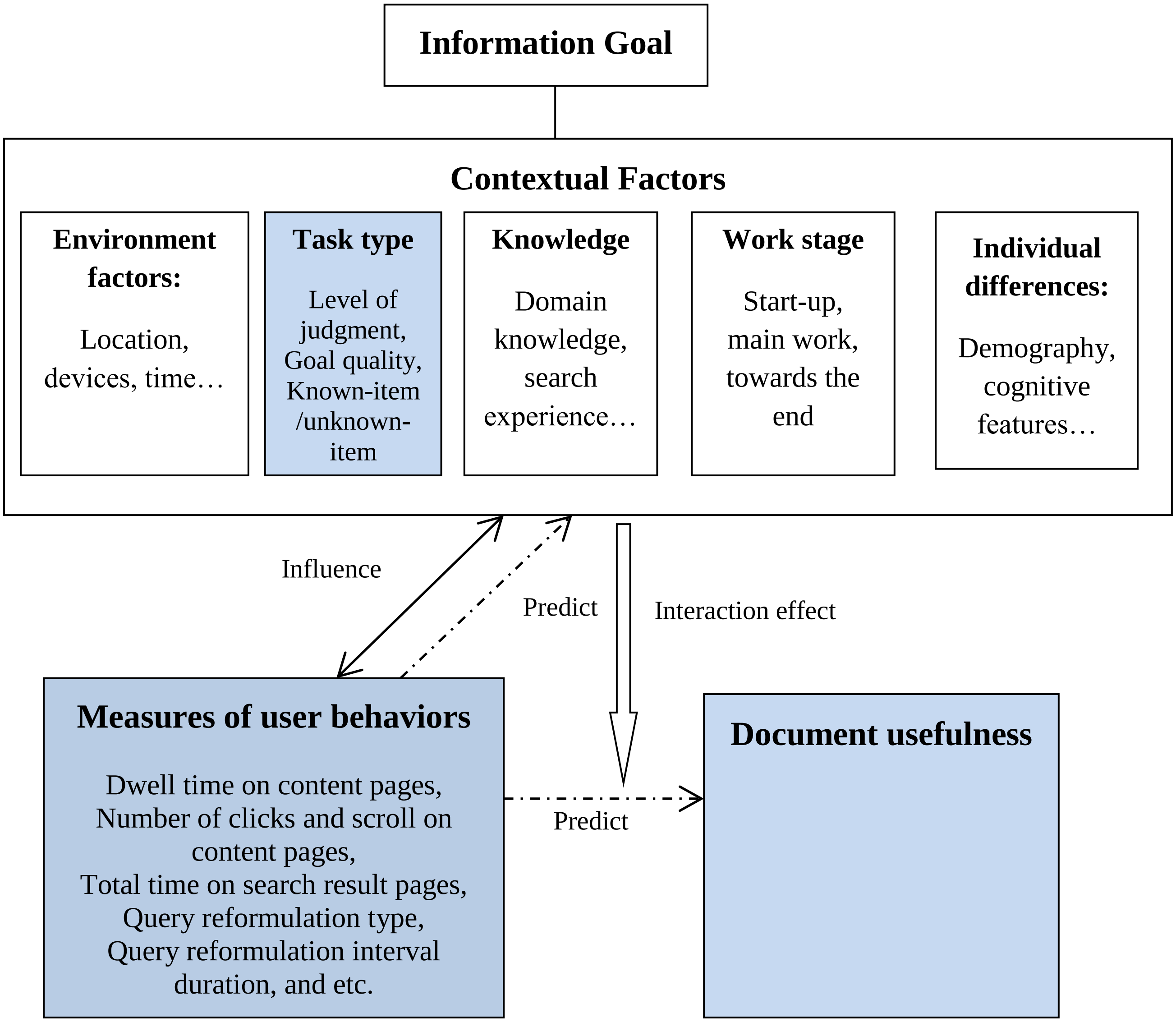

Information Goal
Contextual Factors
Environment factors:
Location, devices, time…
Task type
Level of judgment, Goal quality, Known-item /unknown-item
Knowledge
Domain knowledge, search experience…
Work stage
Start-up, main work, towards the end
Individual differences:
Demography, cognitive features…
Influence
Predict
Interaction effect
Measures of user behaviors
Dwell time on content pages, Number of clicks and scroll on content pages, Total time on search result pages, Query reformulation type, Query reformulation interval duration, and etc.
Predict
Document usefulness

Figure 1. The theoretical model of the study

3.2 Variables considered in the current study

The current thesis explores the interaction relationships between these three components: contextual factors, user behaviors and document usefulness.

A number of contextual factors have been identified in the literature review, including environment factors, task type, user's knowledge, work stage and individual differences. In this study, task type was selected to be examined in terms of its effect on users' general search behaviors, as well as the prediction of document usefulness.

Task type is selected in the current study for the following reasons. Task type has been found to be prominent contextual factor in influencing users' search behaviors (e.g. Li, 2008; Kellar, Watter, and Shepherd, 2007; Toms et al. 2008; Kim, 2009; Liu, et al. 2010). Task type can also influence the type of information objects users expect (Freund, 2008), as well as the interpretation of user behaviors for implicit relevance feedback (e.g. White and Kelly, 2006; Kellar, Watters, Duffy and Shepherd, 2004; Liu and Belkin, 2010). In addition, the performance of personalization algorithms has been found to vary for different types of information goals (Teevan, Dumais and Liebling, 2008; Teevan, Dumais and Horvitz, 2010).

User's search behaviors are investigated comprehensively in this study. User behaviors are examined mainly for two reasons: first is to examine which behaviors are related to users' document preferences; and the second is to explore which behaviors can be used to predict the task type that the user is working on. Measures of user behaviors that have been found to be related to document preferences are summarized in Table 1 in Section 2.1. The current study focuses on users' search behaviors in the current search episode without considering the content of users' queries to predict useful documents. Therefore, the potential significant measures of user behaviors as indicators of document usefulness in the current study include the Attention and Action categories of behavioral signals. Behavioral measures that were found to be influenced by task type are the sources of indicators of task type in the current study. These behavioral measures can be classified as behavioral measures on the whole-session level and behavioral measures on the within-session level. These behavioral measures are summarized in Table 2 in Section 2.2. The current study explored behavioral measures on both levels to identify important predictors of task type in our experiment.

3.3 Research questions

Within this general model, there are two main purposes of the proposed thesis: (1) Generate a predictive model of document usefulness based on users' multiple behaviors on search result pages and content pages in different types of tasks; (2) Generate a predictive model of task type based on users' behaviors. Accordingly, two specific research questions are proposed:

RQ 1. Does a predictive model of document usefulness on the basis of multiple user behaviors in different types of tasks result in more accurate prediction than a predictive model without considering different task types?

As reviewed in the literature, behaviors on content pages, search result pages and other behaviors during search episodes are indicative of document preferences by users. Through this research question, a model to predict document usefulness by combing these significant behavioral measures together without considering task type and four specific models that with task type considered is generated. Then the general predictive model is compared with the specific models with respect to the behavioral measures as predictors, the predictive rules and the prediction performance.

RQ 2. What user behaviors can be used to predict task type?

In this research question, user behaviors at the whole-session and within-session levels are examined and compared by task type and task facets. We then explore behavioral measures to generate the predictive models of task type using behavioral measures on the within-session level, the whole session level and both levels combined.

Chapter 4. Methodology

The research questions in this study were answered through analysis of the data collected in a controlled lab experiment conducted in the Personalization of the Digital Library Experience (PoODLE) project (http://comminfo.rutgers.edu/imls/poodle/). The author of this dissertation had been working as a research assistant for this PoODLE project, in collaboration with other project members on the experimental design and data collection for the project. But the data analysis for this dissertation research has been done by the author. This chapter first describes the design of the experiment, the experimental system, the participants, the search tasks and task types and the procedure of the experiment[1]. This is followed by the discussion of the behavioral measures the author extracted from the search logs and how they were analyzed in this dissertation study to address the research questions.

4.1 *Experimental Design*

The experiment was designed to investigate the effects of search task type and task facets (described in section 4.5) on searching behavior, such as saving and reading behaviors. Data was collected on a variety of searcher behaviors, such as Web page visits with time stamps, mouse and keyboard activities, eye gaze, and various interactions with the search systems and information objects.

4.2 *Experimental system*

The experiment was designed and conducted using a system that reduces the complexity of creating interactive information retrieval (IIR) experiments that log users' multidimensional

[1] A detailed description of this experiment design can be found in Liu, et al. (2010) in the JCDL proceeding.

interactive search behavior (Bierig, et al., 2010). The system has a client-server architecture where researchers configure IIR experiments from a range of extensible tasks. The current experiment configuration applied assigned work and search tasks, questionnaires to gather background information and perceptions before and after the tasks, and usefulness evaluation questionnaires. The system was used to rotate tasks into sequences and monitor the progress of the experiment. Users accessed the experiment through an interface that presented them with their task sequence and provided them with additional instructions. The system is able to log a wide range of user behaviors with an array of heterogeneous logging tools. For the experiment, logs were created for web traffic using UsaProxy (http://fnuked.de/usaproxy/) and Morae (http://www.techsmith.com), keyboard and mouse activity using RUI (http://ritter.ist.psu.edu/projects/RUI/), and eye movements using the Tobii T60 Eye-tracker with Tobii Studio (http://www.tobii.com). The experiment system framework is available as open source (http://sourceforge.net/projects/piirexs/).

The search interface in the experiment system (shown in Figure 2) has two frames: on the right side is the regular Internet Explorer (IE) window, with a blank starting page; on the left side is a panel that allows the users to save desired pages and also to delete them.

4.3 *Participants*

We recruited 32 university undergraduates majoring in journalism as participants for this study,. They were informed in advance that they would receive $20 for participation. To ensure they treated their assigned tasks seriously, they were told that the top 25%, who saved the best set of pages for all four tasks, as judged by an external expert, would receive an additional $20.

More detailed information about the background information of the participants can be found in Section 5.1.1.

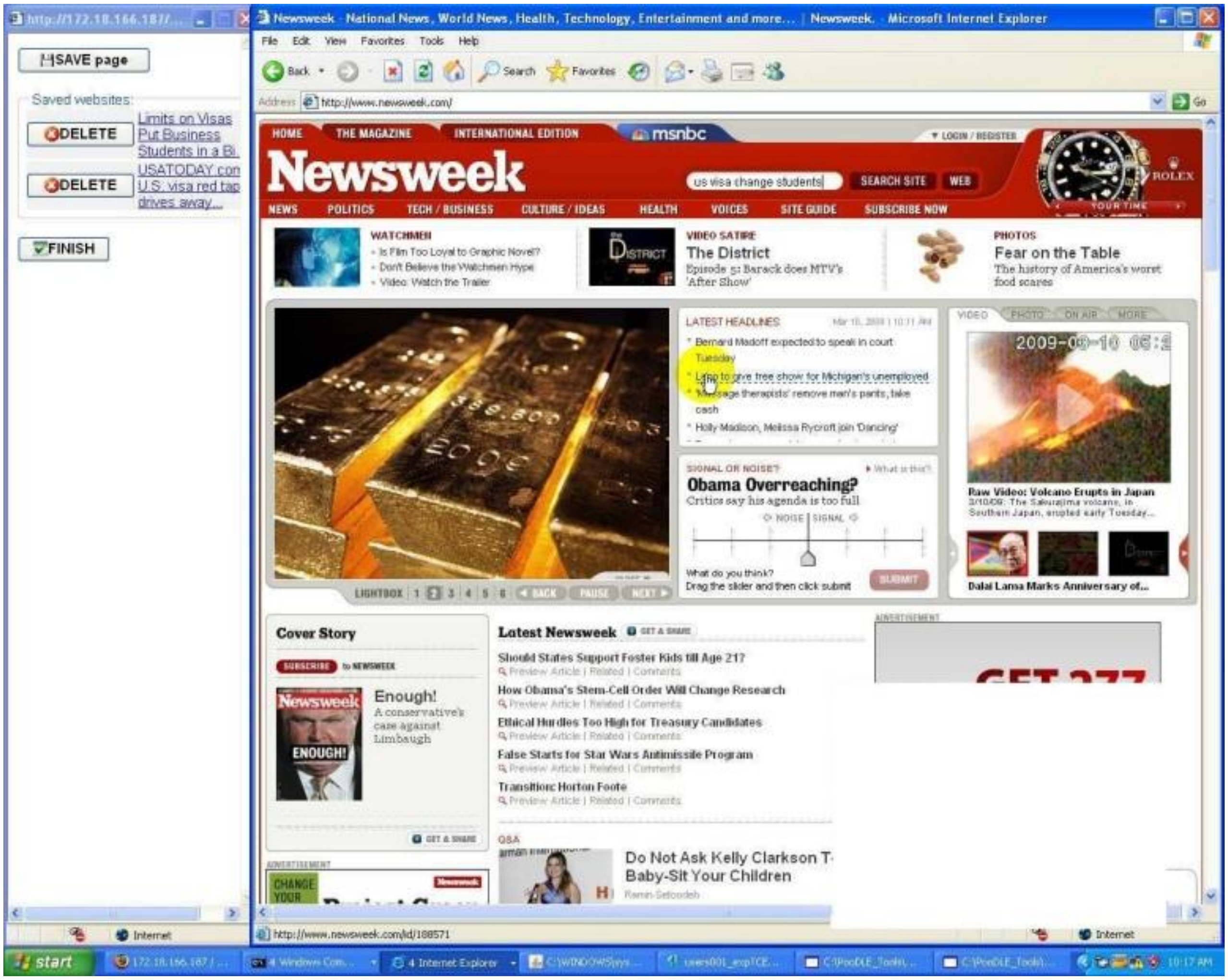

http://172.18.166.187/...
SAVE page
Saved websites:
DELETE
Limits on Visas Put Business Students in a Bi...
DELETE
USATODAY.com U.S. visa red tape drives away...
FINISH
Newsweek - National News, World News, Health, Technology, Entertainment and more... | Newsweek. - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.newsweek.com/
HOME THE MAGAZINE INTERNATIONAL EDITION msnbc
Newsweek
us visa change students
SEARCH SITE WEB
NEWS POLITICS TECH / BUSINESS CULTURE / IDEAS HEALTH VOICES SITE GUIDE SUBSCRIBE NOW
WATCHMEN
VIDEO SATIRE The District
PHOTOS Fear on the Table
LATEST HEADLINES
SIGNAL OR NOISE? Obama Overreaching?
Raw Video: Volcano Erupts in Japan
Dalai Lama Marks Anniversary of...
Cover Story
Enough! A conservative's case against Limbaugh
Latest Newsweek
Should States Support Foster Kids till Age 21?
How Obama's Stem-Cell Order Will Change Research
Ethical Hurdles Too High for Treasury Candidates
False Starts for Star Wars Antimissile Program
Transition: Horton Foote
Do Not Ask Kelly Clarkson To Baby-Sit Your Children
http://www.newsweek.com/id/188571
Internet
start
10:17 AM

Figure 2. The interface of the experiment system

4.4 *Procedure*

Each participant was given a tutorial as a warm-up task and then performed four Web search tasks (described in section 4.5). Participants were asked to search using IE 6.0 on the computer in our lab and they were free to go anywhere on the Web to search for information and were asked to continue the search until they had gathered enough information to accomplish the task (with a time limit of 20 minutes for each task).

For each task, participants were asked to **save** content pages that were useful to accomplish the assignment, or delete saved pages that were found to be not useful later. When participants decided they found and saved enough information objects for purposes of the task, they were then asked to evaluate the usefulness of the information objects they saved, or saved and then deleted, through replaying the search using a screen capture program. An online questionnaire was then administered to ask about their searching experience, including their subjective evaluation of their performance, and reasons for that evaluation. The order of the four tasks was systematically rotated for each participant following a Latin Square design. After completing four different tasks, an exit questionnaire was administered asking about their overall search experience.

4.5 *Search Tasks and Task Types*

This experiment was conducted in the work domain of journalism, for reasons of both validity and convenience. This domain was chosen because although journalism can be associated with any topics, it has a relatively small number of work task types. This means that we were able to have a range of topics for our tasks, while maintaining a good measure of control over realistic tasks, thus enhancing validity. At our institution we have ready access to a university journalism department, so that we had experts to help us define the work tasks, and we had access to participants trained for such professional journalism tasks. We began task identification by interviewing journalism faculty (including practicing journalists) about typical journalism work and searching tasks for which professional journalists receive training. The task descriptions were formalized from those interviews. The task type was defined as a combination of task facets proposed by Li & Belkin (2008). The attractive feature of this classification scheme for our purposes is the ability to vary and control the values of the different facets in the construction of work and search tasks to be performed by the participants in our study. This scheme allowed us to generalize the results in our study to other tasks that contain the same task facet values. Table 3 is an overview of the facets which we manipulated in this experiment.

Table 3. Facets of task which were varied in this study

Facets	**Values**	**Operational Definitions/Rules**
Product	Factual	Locating facts, data, or other similar items
	Intellectual	Produces new ideas or findings on the basis of locating facts
Goal (quality)	Specific goal	Goal is explicit and measurable
	Amorphous goal	Goal cannot be measureable
Naming	Named	Locating factual information about named fact
	Unnamed	Locating factual information about unnamed fact
Level	Document	Judgment is made on the document as a whole
	Segment	Judgment is made on part(s) of a document
Objective Complexity	High complexity	A work task involving at least five activities during engaging in the task; a search task involving searching at least three types of information sources
	Low complexity	A work task involving one or two activities during engaging in the task; a search task involving searching one type of information source

The four work tasks and associated search tasks that we identified are presented below. These tasks follow the normal scenario practice as proposed by Borlund (2003), and are couched in journalism terms; that is, journalists are typically given an assignment, and an associated task to complete.

Background Information Collection (BIC)

Your assignment: You are a journalist for the New York Times, working with several others on a story about “whether and how changes in US visa laws after 9/11 have reduced enrollment of international students at universities in the US”. You are supposed to gather background information on the topic, specifically, to find what has already been written on this topic. Your task: Please find and save all the stories and related materials that have already been published in the last two years in the Times on this topic, and also in five other important newspapers.

Interview Preparation (INT)

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

Copy Editing (CPE)

Your assignment: You are a copy editor at a newspaper and you have only 20 minutes to check the accuracy of the three underlined statements in the excerpt of a piece of news story below.

“New South Korean President Lee Myung-bak takes office

Lee Myung-bak is the 10th man to serve as South Korea's president and the first to come from a business background. He won a landslide victory in last December's election. He pledged to make the economy his top priority during the campaign. Lee promised to achieve 7% annual economic growth, double the country's per capita income to US$4,000 over a decade and lift the country to one of the topic seven economies in the world. Lee, 66, also called for a stronger alliance with top ally Washington and implored North Korea to forgo its nuclear ambitions and open up to the outside world, promising a better future for the impoverished nation. Lee said he would launch massive investment and aid projects in the North to increase its per capita income to US$3,000 within a decade "once North Korea abandons its nuclear program and chooses the path to openness"."

Your Task: Please find and save an authoritative page that either confirms or disconfirms each statement.

Advance Obituary (OBI)

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

The task facet values for each of the four search tasks are shown in Table 4. We held constant the values of the following facets (not in Table 3): Source of task; Task doer; Time (length) Process; Goal (quantity); Interdependence; and Urgency.

Table 4 Variable facet values for the search tasks

Task	Product	Level	Naming	Goal (quality)	Objective Complexity
BIC	Mixed	Document	Unnamed	Specific	High
CPE	Factual	Segment	Named	Specific	Low
INT	Mixed	Document	Unnamed	Mixed	Low
OBI	Factual	Document	Unnamed	Amorphous	High

The BIC task is a mixed product because identifying "important" newspapers is intellectual, but finding topical documents is factual. It is Document Level because whole stories are judged. It has the Specific Goal of finding documents on a well-defined topic, but Unnamed because the search targets are not specifically identified (compare with CPE).

INT is a Mixed Product, because defining expertise is intellectual, and contact information is a fact. It is at the Document Level, because expertise is determined by a whole page. The Goal Quality is Mixed, because determining expertise is amorphous but contact information is specific. It is Unnamed because the search targets are not specifically identified in the task.

CPE is a Factual Product, because facts have to be identified. It is at the Segment Level, because items within a document need to be found. It has the Specific Goal of confirming facts, and it is Named because the search targets are specified.

OBI is a Factual Product, because facts about the person are needed. It is at the Document Level because entire documents need to be examined. It is Unnamed because the search targets are not specifically identified in the task. The Goal Quality is Amorphous because "all the information" is undefined.

4.6 *Predictive Modeling Methods*

Both of the two research questions in this dissertation study involve generating predictive models. In this section, we briefly discuss the available methods for predictive modeling, and what methods were chosen in this dissertation study. There are several different mathematical ways to combine variables into a multivariable predictor. In RQ1, our goal is to generate predictive models of document usefulness, either useful or not-useful. For binary outcomes, the most often used model is logistic regression that is a form of generalized linear models and allows one to predict a discrete outcome, such as group membership from a set of dependent variables that may be continuous, discrete, dichotomous, or a mix of any of these. Logistic regression can be used as a predictive modeling method to estimate the likelihood of document usefulness for a given document with a set of observations that are particular to that content page. By conducting logistic regression, the relationships among a series of predictors with the outcome can be modeled, allowing for better understanding of their relationships, and variables are combined in a linear manner. However, non-linear models may be also pertinent for some situations. In contrast to logistic regression, recursive partitioning is a nonparametric type of analysis that repeatedly subdivides data into smaller and smaller subgroups based on characteristics that predict the desired endpoint. The goal is to construct subgroups that, ideally, consist entirely of subjects with one endpoint category or another. Recursive partitioning, unlike logistic regression is nonlinear in its parameters, and would have an advantage if the true relationship between the variables and the outcome of interest is nonlinear. Since this is an exploratory study, we used both logistic regression and recursive partitioning methods to identify the important predictors as document usefulness.

In RQ2, we need to generate predictive models of task type, and in our study, there are four categories in task type. Therefore, we need to explore methods that can generate predictive models of more than two discrete outcomes. There are several possible methods that we can explore, like multinomial logistic regression, neural networks, or support vector machines (SVM). Among them, the predictive models by neural networks or SVM may be difficult to interpret in order to explain the importance of each predictor variable. The goal of our study in RQ2 is to examine which behavioral measures are more important in distinguishing different task types, rather than achieving the best prediction performance. Thus, we decided to use multinomial logistic regression to generate the predictive models of task type.

4.7 *Modeling on Behavioral Measures*

During the search, all of the participants' interactions with the computer system were logged during the searches on the client side, using several multiple logging systems. This dissertation study is to explore what behavioral measures can be used to infer document usefulness and task type. Therefore, we need to extract or calculate some behavioral measures to generate the predictive models.

RQ1 is aimed at comparing the accuracy of the prediction of document preferences when combining multiple user behaviors, with all tasks considered and in different task types. To address this research question, first of all, we need to extract measures of user behaviors that are associated with individual content pages from users' search interaction logs. The behavioral variables examined in RQ1 are listed in Table 5.

Table 5. Behavioral measures examined in RQ1

Behavioral measures	**Definition**
Behavioral measures on clicked documents	
dwell time	the dwell time of the document, from the point that the page is opened till the point that the page is closed
number.of.mouse. clicks	the number of mouse clicks, including Left button down, Right button down, and scrolling, while the page is opened
number.of.keystrokes	the number of keystrokes while the page is opened
visit_id	the number of times a content page has been visited during one search session
Behavioral measures during query intervals	
time_to_first_click	the time before first click after issuing a query
content_mean	the average dwell time on content pages during that query interval;
content_sum	the total dwell time on content pages during that query interval;
content_count	the total number of content pages visited during that query interval;
serp_mean	the average dwell time on SERPs during that query interval;
serp_sum	the total dwell time on SERPs during that query interval;
serp_count	the total number of SERPs examined during that query interval;
prop_content	the proportion of time on content pages of the total dwell time during that interval
interval_time	the total time in each query interval
diff_content	the difference between the dwell time on a content page and the average dwell time on all content pages during its associated query interval

There are in general two groups of behavioral measures: one is the measures on the clicked documents, which describe how users interact on each of the content pages; the other is the measures during query interval[2], which describe what users do between issuing one query and the next. For RQ1, two statistical methods were considered in this study: logistic regression and recursive partitioning. In particular, we first used recursive partitioning to identify the most important predictors, and their relationships for predicting document usefulness; then we also

[2] Query interval is defined as the interval between two successive queries issued in one search session.

generated the predictive models using logistic regression. The results of the two models were compared and evaluated to decide which one to be selected as the final predictive model to use.

RQ2 is to investigate user behaviors that can be used to predict task type. The behavioral variables we extracted from the user-experiment log are listed in Table 6. These behavioral measures can be divided into two levels according to the time point when the variable can be observed by the system: whole-session variables, which cannot be captured by the system until the completion of the search session; within-session variables, which can be captured by the system during any ongoing search session. As mentioned in the literature review, previous studies mainly focused on the whole-session level, but the prediction of task type based on the within-session level is more useful for personalization.

Table 6. The behavioral measures examined in RQ2

Behavioral measures	Definition
behavioral measures on the whole-session level	
Task completion time	The total time users spent on each task session
Numbers of all documents	The number of all documents that the user viewed in the search session
Numbers of unique documents	The number of unique documents that the user viewed
Number of SERPs	The number of search result pages that the user viewed in the search session
Number of unique SERPs	The number of unique search result pages that the user viewed in the search session
Number of queries	The number of queries that the user issued in the search session
Total time spent on documents	The sum of dwell time users spent on all viewed documents
Total time spent on SERPs	The sum of dwell time users spent on all search result pages
Ratio of document time to all	The ratio of total time spent on documents to task completion time

Ratio of SERP time to all	The ratio of total time spent on search result pages to task completion time
Behavioral measures on the within-session level	
Mean dwell time of all documents	The average dwell time of all the visited documents in the session
Mean dwell time of unique documents	The average of the total dwell time of all unique documents visited in the session
Mean dwell time of all SERPs	The average dwell time of all the search result pages in the session
Mean dwell time of unique SERPs	The average of the total dwell time of all unique search result pages in the session
Number of documents per query	Average number of viewed documents per query
Number of unique documents per query	Average number of unique viewed documents per query
Number of SERPs per query	Average number of search result pages visited per query
Number of unique SERPs per query	Average number of unique search result pages visited per query
Average query interval	The average time of all query intervals in the task; the query interval time is the period from the time point after one query is issued and before the subsequent query is issued
Average Time To First Click	The average time on Time_to_first_click, which is defined as the time after a query is issued and before the first click on the search result of that query

Then we used the multinomial logistic regression method to generate the predictive models of task type on the basis of behaviors at the whole-session level, at the within-session level, and on both levels.

Chapter 5. Results for RQ1

RQ 1. Does a predictive model of document usefulness on the basis of multiple user behaviors in different types of tasks result in more accurate prediction than a predictive model without considering different task types?

In RQ1, we first conducted descriptive analysis of participants, the number of content pages visited in the experiment, and the behavioral measures that were selected for this RQ. Then the behavioral measures on useful and non-useful pages were compared, to examine the document usefulness effect on each of the behavioral measures in general. We then explored two methods to generate the predictive models of document usefulness: recursive partitioning and logistic regression. The general predictive model was first generated and then the specific predictive models for each type of tasks were generated. Finally, the general and specific predictive models were compared and evaluated with respect to the predictors and the prediction performance.

5.1 *Descriptive analysis*

5.1.1 Descriptive analysis of participants

Before participants worked on searching, they were asked to fill out a background questionnaire to collection personal information, e.g. demographic characteristics, experience and expertise in computer use and information search, using both search engines and online library catalogs. A descriptive analysis on the background information was conducted.

From the questionnaire, we know that the participants were between 18 and 27 years old. Among them, 29 were between 18 and 22 years old, and 3 were between 23 and 27 years old.

There were 26 female (81.3%) among all participants, and 6 were male. Most students were native English speakers (78%) with the remainder of the population stating a high degree of English knowledge. Participants reported to have been using a range of different browsers (IE, Firefox, Safari, Safari, Chrome, AOL Explore, Flock and Mozilla).

Participants were asked to indicate their levels of computer expertise on a seven point scale, where 1= novice and 7=expert. On average, participants rated their levels of expertise with computers as 4.75 (standard deviation, simplified as SD in the following, was 0.72). One participant rated himself as level 3 (the minimum in the sample), and four rated themselves as level 6 (the maximum in the sample), and the rest of the 27 participants rated themselves as level 4 or 5. On average, they had 8.84 years of online searching (SD=2.07). Among them, one claimed to have 4 years of online searching (the minimum in the sample), one had 5 years, and one had 14 years (the maximum in the sample).

Participants rated their levels of expertise with searching as high as 5.97 on a 7-point scale (SD= 1.09). Except for two participants, one of whom rated himself as level 2 and one who rated himself as level 4, all other 30 participants rated themselves higher than neutral. On average, they rated their levels of experience of WWW search engines as 5.06 on the 7-point scale (SD=0.88). Among them, one rated himself as level 3 (the minimum in the sample), and one rated himself as level 7 (the maximum in the sample). Their average level of experience with online library catalogs was 6.44 on the 7-point scale (SD=0.76). Compared with their self-ratings on the level of experience of WWW search engines (5.06), participants claimed that they had higher levels of online library catalogs searching (6.44). Among them, five rated themselves as level 5 (the minimum in the sample), eight rated themselves as level 6, and 19 rated themselves as level 7 (the maximum in the sample). Participants were asked to rate how often they could find what they look for on the 7-point scale. On average, their level of 4.41 with quite large variety (SD=1.34). One of them rated himself as level 1 (the minimum in the sample), two rated themselves as level 2, four rated themselves as level 3, eight rated as level 4, eleven rated as level 5, five rated as level 6, and one rated himself as level 7 (the maximum in the sample).

5.1.2 Descriptive analysis of content pages visited

In this experiment, participants were asked to save the best pages that could help them accomplish each of the search tasks. Data of users' interactions with the computer was logged by Morae software. In all 128 sessions (32 participants * 4 search sessions), it was observed that participants visited 3725 content pages. Among them, 1033 pages (about 27.7%) were saved as useful pages, and the other 2692 pages (about 72.3%) were not saved, and they were considered as non-useful pages.

Table 7. Frequency table of the saved and non-saved pages

	Frequency	Percentage	Average number of pages per session
Non-saved pages	2692	72.3%	21.03
Saved pages	1033	27.7%	8.07
Total content pages	3725	100%	29.10

On average, each participant visited about 29 content pages per session, and among them, about 8 content pages were saved as useful pages. Except for one participant in one search task[3], all participants saved at least one content page during each of the four search sessions. The frequency data of the content pages visited in this user experiment is shown in Table 7.

[3] User007 in OBI did not save any documents.

Table 8. Frequency table of the saved and non-saved pages in each of the tasks

		Non-saved pages	Saved pages	Total number of pages visited
BIC	Number of pages	890	271	1161
	Percentage of total pages	76.7%	23.3%	100%
CPE	Number of pages	301	212	513
	Percentage of total pages	58.7%	41.3%	100%
INT	Number of pages	839	262	1101
	Percentage of total pages	76.2%	23.8%	100%
OBI	Number of pages	662	288	950
	Percentage of total pages	69.7%	30.3%	100%
Total		2692	1033	3725

From this descriptive analysis we also found that the number of saved and non-saved pages in the experiment and in each of the tasks was not balanced, with too many of them non-saved. Therefore, we needed to balance the samples with the same number of saved and non-saved pages before generating the predictive models; please refer to section 5.3 and 5.4 for more detailed information on the balanced sampling.

5.1.3 Descriptive analysis of behavioral measures

From the logs, we extracted a variety of measures of the users' search behaviors on search result pages (SERPs) and content pages throughout the session, and during query intervals. In this part, we conducted a descriptive analysis for each of the behavioral measures.

The descriptive analysis of each of the behavioral measures on the whole-session level is shown in Table 9. The histograms and the Q-Q plots for each of the behavioral measures examined for RQ1 are shown in Figure 10 to Figure 21 (in Appendix A) to examine whether the distribution of each variable was normal or not. They demonstrate that none of these measures were normally distributed. Therefore in the univariate statistical analysis, we need to conduct non-parametric analysis to compare the distribution among saved and non-saved groups.

Table 9. Descriptive statistics of behavioral measures on the whole-session level

	Minimum	Maximum	Mean	Median	Std. Deviation
dwell time on content pages (seconds)	0.10	318.50	14.33	9.10	17.06
visit_id	1.00	18.00	1.53	1.00	1.20
number of mouse clicks	0.00	140.00	7.01	2.00	11.63
number of keystrokes	0.00	78.00	1.23	0.00	5.34
time_to_first_click (seconds)	0.20	262.50	8.97	6.80	10.45
content_mean (seconds)	0.10	318.50	14.33	11.41	12.18
content_sum (seconds)	0.10	520.20	98.89	68.40	94.80
content_count	1.00	34.00	8.16	6.00	7.41
serp_mean (seconds)	0.20	136.25	8.85	7.30	8.21
serp_sum (seconds)	0.20	272.50	25.89	19.20	25.50
serp_count	1.00	20.00	3.18	2.00	2.69
prop_content	0.00	1.00	0.72	0.77	0.22
interval_time (seconds)	3.60	539.40	124.79	93.80	99.41

5.2 *Univariate analyses of behavioral measures*

Before generating the predictive models, we used univariate analyses to compare each of the behavioral measures in the two conditions, on saved pages or on non-saved pages. As mentioned before, the distribution of all these variables were not normal, so Wilcoxon tests were conducted. The results of the Wilcoxon tests showed, that except three of these variables (time_to_first_click, content_count, prop_content), all other behavioral measures were significantly different between saved and non-saved pages (shown in

Table 10).

Table 10. Univariate comparisons of candidate predictor variables between saved and non-saved pages.

Behavioral variables	Non-saved Median (SD)	Saved Median (SD)	Wilcoxon test results
dwell time on content pages (seconds)	7.80 (13.76)	14.50 (22.40)	p<.001**
visit_id	1.00 (1.18)	1.00 (1.21)	p<.001**
number of mouse clicks	2.00 (10.51)	3.00 (13.97)	p<.001**
number of keystrokes	0.00 (5.64)	0.00 (4.48)	p<.001**
time_to_first_click (seconds)	6.80 (11.39)	7.00 (7.46)	p=.27
content_mean (seconds)	11.00 (9.71)	13.02 (16.67)	p<.001**
content_sum (seconds)	65.60 (95.77)	72.40 (92.23)	p=.001**
content_count	6.00 (7.52)	6.00 (7.08)	p=.15
serp_mean (seconds)	7.24 (9.06)	7.40 (5.42)	p=.02*
serp_sum (seconds)	17.90 (25.65)	21.30 (25.03)	p<.001**
serp_count	2.00 (2.66)	3.00 (2.74)	p<.001**
prop_content	0.77 (0.23)	0.78 (0.20)	p=.65
interval_time (seconds)	92.10 (99.96)	99.50 (97.88)	p=.001*

Note: The variables that had significant difference between two situations are highlighted in this table.

Specifically, compared with non-saved pages, users were more likely to have more dwell time, more mouse clicks and keystrokes, and visited more than once on saved content pages; in addition, during the intervals that contained saved pages, they were more likely to have more dwell time on content pages and SERPs, longer total time on content pages and SERPs, visited more SERPs, and the total interval time was longer than during query intervals that did not contain saved pages.

5.3 *General predictive modeling*

To better learn the behavioral measures, we made a training collection that balanced the number of saved pages and not-saved pages by sampling the larger not-saved pages pool. In each of the balanced samples, we randomly selected the same number of non-saved pages as the number of saved pages, and then combined the selected non-saved pages with the saved pages. In this way, ten balanced training sets were constructed, each sharing the same saved pages pool. The recursive partitioning and logistic regression methods were applied to each sample to generate predictive models. For the decision tree model generated from the recursive partitioning, we compared the variables and cutoff values in the generated models to identify the repeated variables and cutoff values; for the logistic regression model, we selected the repeated variables in the generated models and applied them on the whole dataset and the logistic regression model that achieve the best prediction accuracy is selected as the final logistic regression predictive model.

5.3.1 Predictive modeling using recursive partitioning

We generated predictive models using recursive partitioning based on each of the ten random samples, and these tree models are shown Figure 22 (in Appendix B). Among the ten tree models, three behavioral measures were consistently selected as the predictors in all the models: *dwell time*, *visit_id*, and *time to first click*; and all these three variables were positively related to document usefulness. First, dwell time on content pages was selected as the most important predictor in all the ten samples, and the cutoff points ranged from 15.95 seconds to 21.65 seconds. This is consistent with previous discussions and research on dwell time as a predictor of document usefulness. Even though in some studies, the dwell time was not found to be significantly different between useful and non-useful pages (Kelly & Belkin, 2005), it was identified as the most important factor in the tasks designed in our experiment. Secondly, visit_id, the time a page has been visited in the task was selected as an important predictor of document usefulness when the dwell time on the content pages was shorter than the cutoff point. That is, if a content page has been visited more than once in a task session, even if the dwell time on it is not longer than the cutoff point, it is very likely that this page is a useful page for the user in this task. Thirdly, time_to_first_click, the time the user spent on SERPs after issuing a query and before the first click on the SERP, was also selected as one of the important predictors. That is, if the time_to_first_click was shorter than 1.75 seconds, then the content pages clicked on that SERP were very likely to be non-useful pages.

The tree models from three of the ten samples identified other behavioral measures, besides the three most important variables we discussed above: sample 5, sample 6, and sample 9. In sample 5 and 6, prop_content and content_mean were identified to be negatively related to document usefulness. In sample 9, number of keystrokes was selected as an important predictor, and it was negatively related to document usefulness. Because these variables were selected as important predictors in a small portion of the samples we have, we do not consider them as important predictors in our final general prediction tree model from recursive partitioning.

Then the next step is to decide the cutoff point for each of the predictors: *visit_id*, *dwell time* and *time to first click*. The method we adopt is to calculate the mean cutoff points in the ten samples, and then use the mean as the cutoff point for the final model. The cutoff of the variable, visit_id was the same among the ten samples, (visit_id > 1), so this cutoff point is the same in the final model. The mean cutoff point for dwell time is 18 seconds, and the mean cutoff point for time to first click is 1.71 seconds. Therefore, the general prediction tree model can be shown as Figure 3.

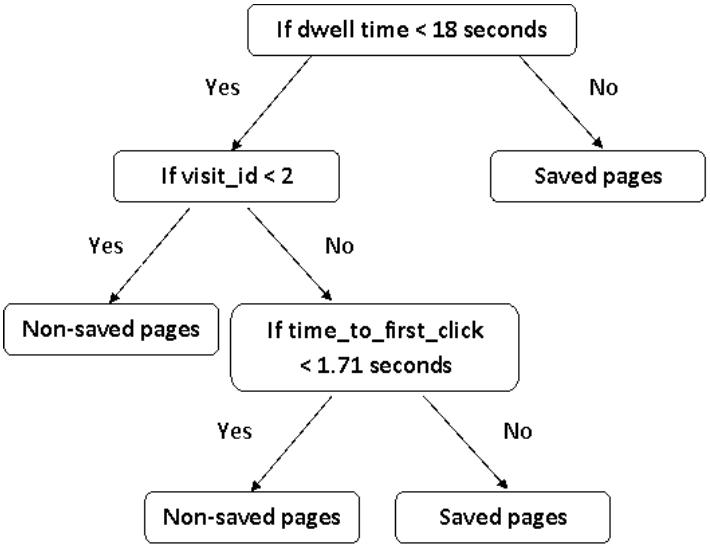


Figure 3 The general predictive model

5.3.2 Predictive modeling using logistic regression

We also generated logistic regression models for each of the ten samples, and each sample identified several important predictors of document usefulness. The selected predictors in each of the samples for the general model are shown in Table 61--Table 70 (in Appendix C). Table 11 lists all the selected predictors and the coefficients in the ten samples. It shows that three variables were consistently selected in the logistic regression model, they are dwell time, visit_id and number of keystrokes. Among them, dwell time and visit_id were positively related to document usefulness, while number of keystrokes was consistently negatively related to document usefulness. Six samples selected number of mouse clicks as one of the important predictors. Some of the samples also selected other variables as predictors of document usefulness, e.g. serp_sum, time_to_first_click and content_sum, but given that they were not consistently selected, and the small coefficients (which indicated less contribution to the predictive model), we do not include them in the final predictive model.

Table 11. The coefficients for selected predictors in ten samples (general model)

	1	2	3	4	5	6	7	8	9	10	total
dwelltime	0.05	0.04	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	10
visit_id	0.6	0.53	0.39	0.55	0.53	0.45	0.51	0.51	0.44	0.47	10
number of keystrokes	-0.04	-0.02	-0.04	-0.03	-0.05	-0.03	-0.03	-0.04	-0.05	-0.04	10
number of mouse clicks		-0.01		-0.01	-0.01	-0.01	-0.02		-0.01		6
serp_sum		0.01				0.01				-0.01	3
time to first click		-0.02								-0.02	2
content_sum					-0.003			-0.003	-0.002		3

Therefore, the general logistic regression model included four behavioral measures, dwell time, visit_id, number of keystrokes and number of mouse clicks. Then the ten logistic regression models were evaluated on the whole dataset, i.e. all the visited content pages in the PoODLE project, and the model that obtain the best performance on the whole data was selected as the final general logistic regression model. This model obtained the best performance among the ten logistic regression models based on the whole data, and its prediction performance is shown in Table 12.

Equation 1. General Logistic Regression (LR) predictive model of document usefulness

$$\ln\left(\frac{p}{1-p}\right) = -\ 1.47\ +\ 0.05\ *\ dwell\,time\ +\ 0.53\ *\ visit\_id - 0.04\ *\ number.of.keystrokes - 0.01\ *\ number.of.mouse.clicks$$

Table 12. The prediction performance using the general model

		Predicted		Overall
		Saved pages	Non-saved	
Observed	Saved pages	615	416	
	Non-saved pages	674	2018	
Prediction accuracy		47.71%	82.91%	70.72%

5.4 *Specific predictive modeling*

Specific predictive models were task specific, so we generated the specific predictive models for each of the task in our experiment. Similarly as the general predictive modeling, we also first constructed ten balanced random samples for each of the tasks as the training sets. In each of the balanced samples, we randomly selected the same number of non-saved pages as the number of saved pages in each task, and then combined the selected non-saved pages with the saved pages. In this way, ten balanced training sets were constructed for each task, each sharing the same saved pages pool. The recursive partitioning and logistic regression methods were applied to each sample to generate predictive models. For the decision tree model generated from the recursive partitioning, we compared the variables and cutoff values in the generated models to identify the repeated variables and cutoff values; for the logistic regression model, we selected the repeated variables in the generated models and applied them on the whole dataset of that type and the logistic regression model that achieve the best prediction accuracy is selected as the final specific logistic regression predictive mode for that task. In this part, we presented the predictive modeling process for the specific models in each of the four tasks.

5.4.1 Specific model for the BIC task

1) Predictive modeling using recursive partitioning

We generated predictive models using recursive partitioning based on each of the ten random samples in BIC tasks, and these tree models are shown in Figure 23 (in Appendix B). The behavioral variables that have been selected as predictors in these models are summarized in Table 13, which also counted the number of times each variable was selected in the models. The variables that had been selected more than five times were selected in the final recursive portioning predictive model, and there were four variables in the final model: *dwell time*, *visit_id*, *serp_sum*, and *number.of.mouse.clicks*. Similarly as in the general model, the cutoff point for each variable was calculated as the mean of the cutoffs in the models.

Table 13. The selected predictors from recursive partitioning in ten samples (BIC)

variables	1	2	3	4	5	6	7	8	9	10	total
dwelltime	x	x	x	x	x	x	x	x	x	x	10
visit_id	x	x	x	x	x	x	x	x		x	9
serp_sum	x	x	x	x	x	x	x	x		x	9
number.of.mouse.clicks		x	x	x	x	x	x	x		x	8
time_to_first_click	x			x				x	x	x	5
number.of.keystrokes			x			x			x		3
serp_mean	x				x		x		x		4
prop_content		x							x		2
content_sum			x								1
content_mean						x					1
diff_content							x				1
interval									x		1

Among these predictors, except number.of.mouse.clicks, which was negatively related to document usefulness, all other predictors were positively related to document usefulness in BIC.

The cutoff point for dwell time is 12.45 seconds, for visit_id, it is 2, for the number of mouse clicks is 2; there were two cutoff points for serp_sum, depending on the value of dwell time. If the dwell time on the page is more than 12.45 seconds, the cutoff point for serp_sum is 16.22 seconds; if the dwell time on the page is less than 12.45 seconds, the cutoff point for serp_sum is 8.71 seconds. The specific tree predictive model for BIC is shown in Figure 4.

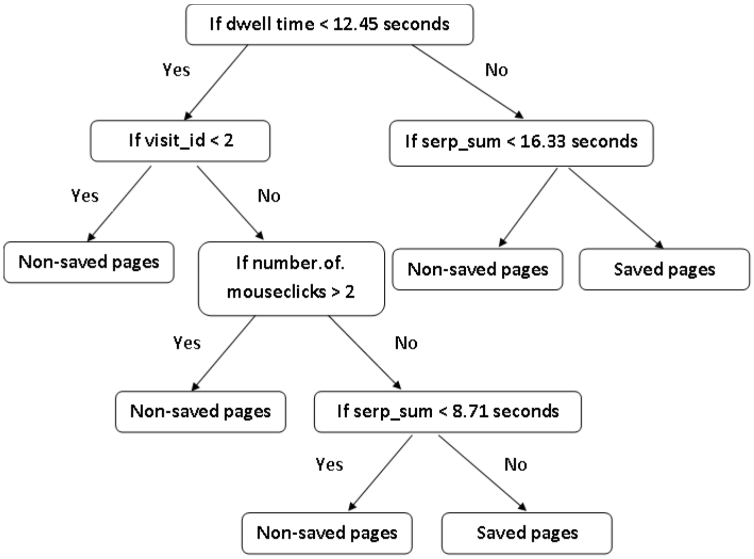


Figure 4. Specific decision-tree predictive model for BIC

2) Predictive modeling using logistic regression

We then generated logistic regression models on each of the ten samples, each sample identified several important predictors of document usefulness in BIC task. The selected predictors in each of the samples for BIC are shown in Table 71 --Table 80 (in Appendix C). Table 14 listed all the selected predictors and the coefficients in the ten samples. It shows that two variables were consistently selected in the logistic regression models: dwell time and number of keystrokes. Another two variables were selected in more than five models and they were also included in the final predictive model for BIC task: content_count and content_sum.

Table 14. The coefficients for selected predictors in the ten samples of BIC task

Samples	1	2	3	4	5	6	7	8	9	10
dwell time	0.06	0.05	0.06	0.06	0.05	0.06	0.04	0.05	0.05	0.05
number of keystrokes	-0.09	-0.07	-0.08	-0.09	-0.07	-0.07	-0.07	-0.06	-0.10	-0.07
content_count		0.10	0.10	0.15	0.09	0.11	0.13	0.11	0.18	0.12
content_sum				-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
time_to_first_click								-0.04	-0.04	-0.04
content_mean									0.05	0.06
prop_content									-1.62	
number of mouse clicks										-0.03

Therefore, the specific logistic regression model for BIC task included four behavioral variables: dwell time, number of keystrokes, content_count, and content_sum. The model that obtained the best performance on the whole dataset was selected as the final specific logistic regression model.

Equation 2. Logistic Regression (LR) predictive model of document usefulness for the BIC task

$$\ln(\frac{p}{1-p}) = -0.76 + 0.06 * dwell\ time - 0.08 * number.of.keystrokes + 0.07 * content\_count - 0.005 * content\_sum$$

5.4.2 Specific model for the CPE task

1) Predictive modeling using recursive partitioning

We generated predictive models using recursive partitioning based on each of the ten random samples in CPE tasks, and these tree models are shown in Figure 24 (in Appendix B). The behavioral variables that have been selected as predictors in these models are summarized in Table 15, which also counted the number of times each variable was selected in the models. The variables that had been selected more than five times were selected in the final recursive portioning predictive model, and there were five variables in the final model: dwell time, visit_id, prop_content, number.of.mouse.clicks, and diff_content.

Table 15. The selected predictors from recursive partitioning in ten samples (CPE)

variables	1	2	3	4	5	6	7	8	9	10	total
dwell time	x	x	x	x	x	x	x	x	x	x	10
visit_id	x	x	x	x	x	x	x	x	x	x	10
prop_content	x		x	x	x	x		x	x	x	8
number.of.mouse.clicks	x	x	x	x		x	x		x		7
diff_content		x		x			x	x	x		5
content_sum	x	x					x				3
serp_mean	x		x					x			3
content_count		x			x		x				3
time_to_first_click				x		x				x	3
content_mean			x							x	2
interval		x									1
serp_sum									x		1

Among these predictors, except number.of.mouse.clicks, which was negatively related to document usefulness, all other predictors were positively related to document usefulness in CPE. Similarly as before, the cutoff pint for each variable was calculated as the mean of the cutoffs in the models. There were two cutoff points for dwell time: the one on the top level is 32.56 seconds, and the other one on the lower level is 13.3 seconds. The cutoff point for visit_id is 2, for prop_content is 0.84, and for number.of.mouse.clicks is 11. The specific tree predictive model for CPE can be shown as Figure 5.

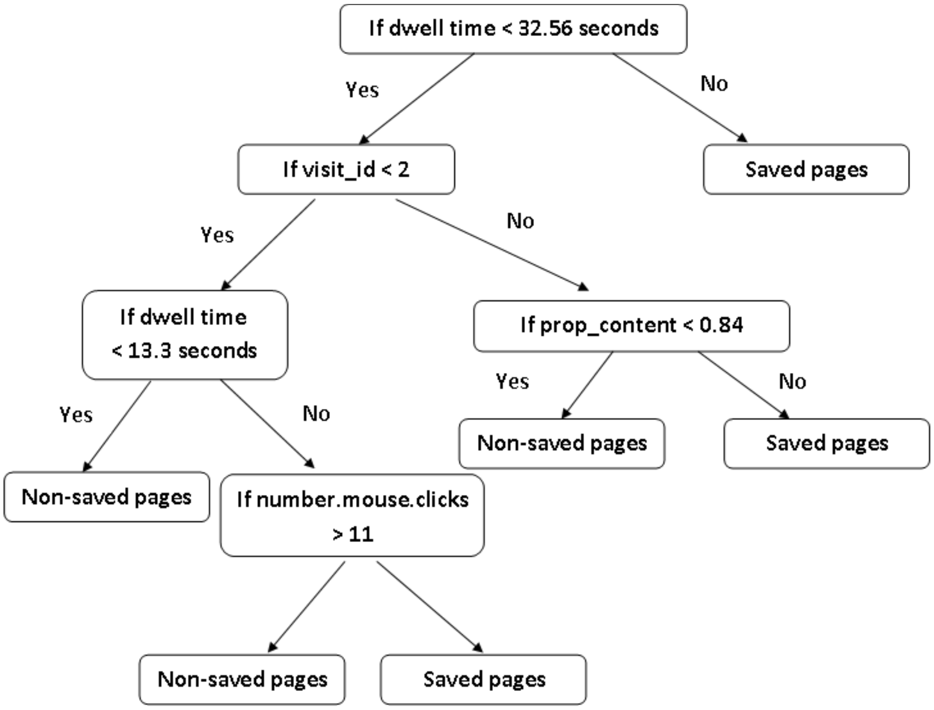


Figure 5. Specific decision-tree predictive model for CPE

2) Predictive modeling using logistic regression

We then generated logistic regression models on each of the ten samples, and each sample identified several important predictors of document usefulness in CPE task. The selected predictors in each of the samples for CPE are shown in Table 81--Table 90 (in Appendix C).

Table 16. The coefficients for selected predictors in the ten samples of CPE task

Samples	1	2	3	4	5	6	7	8	9	10
dwell time	0.05	0.05	0.05	0.04	0.04	0.04	0.05	0.06	0.05	0.05
visit_id	0.76	0.79	0.79	0.62	0.68	0.60	0.75	0.62	0.72	0.67
number of mouse clicks	-0.03	-0.02	-0.02	-0.02	-0.02	-0.03	-0.03	-0.03	-0.02	-0.03
content_mean							-0.03			-0.03

Table 16 lists all the selected predictors and the coefficients in the ten samples. It shows that three variables were consistently selected in the logistic regression models: dwell time, visit_id and number of mouse clicks. Another variable was selected in two models was content_mean, but since it only occurred twice among the ten samples, we do not include it in the final model. Therefore, the specific logistic regression model for CPE task included three behavioral variables: dwell time, visit_id and number of mouse clicks. The model that obtained the best performance on the whole dataset was selected as the final specific logistic regression model.

Equation 3. Logistic Regression (LR) predictive model of document usefulness for the CPE task

$$\ln(\frac{p}{1-p}) =- 1.59 + 0.03 * dwelltime + 0.07 * visit\_id - 0.02 * number.of.mouse.clicks$$

5.4.3 Specific model for the INT task

1) Predictive modeling using recursive partitioning

We generated predictive models using recursive partitioning based on each of the ten random samples in INT tasks, and these tree models are shown Figure 25 (in Appendix B). The behavioral variables that have been selected as predictors in these models are summarized in Table 17, which also counts the number of times each variable was selected in the models. The variables that had been selected more than five times were selected in the final recursive portioning predictive model, and there were three variables in the final model: dwell time, visit_id, and prop_content.

Table 17. The selected predictors from recursive partitioning in ten samples (INT)

variables	1	2	3	4	5	6	7	8	9	10	total
dwell time	x	x	x	x	x	x	x	x	x	x	10
visit_id	x	x	x	x	x	x	x	x	x	x	10
prop_content	x		x			x	x	x	x		6
content_sum	x	x	x				x				4
diff_content	x			x			x			x	4

time_to_first_click		x	x	x			x				4
content_count			x	x	x				x		4
number.of.mouse.clicks				x		x		x		x	4
serp_sum				x	x			x	x		4
serp_count		x				x				x	3
interval					x					x	2
content_mean								x		x	2
serp_mean										x	1

All the three predictors, dwell time, visit_id, and prop_content were positively related to document usefulness in INT. Different from BIC and CPE model, the prediction tree model for INT identified visit_id as the most important predictor, and it identified two cutoff points for visit_id. The top level specified visit_id < 3, and at the lower level, it specified visit_id < 2. Similarly as before, the cutoff pint for dwell time and prop_content was calculated as the mean of the cutoffs in the models. The cutoff point for dwell time is 21.55 seconds. The cutoff point for prop_content is 0.94. The specific tree predictive model for INT can be shown as Figure 6.

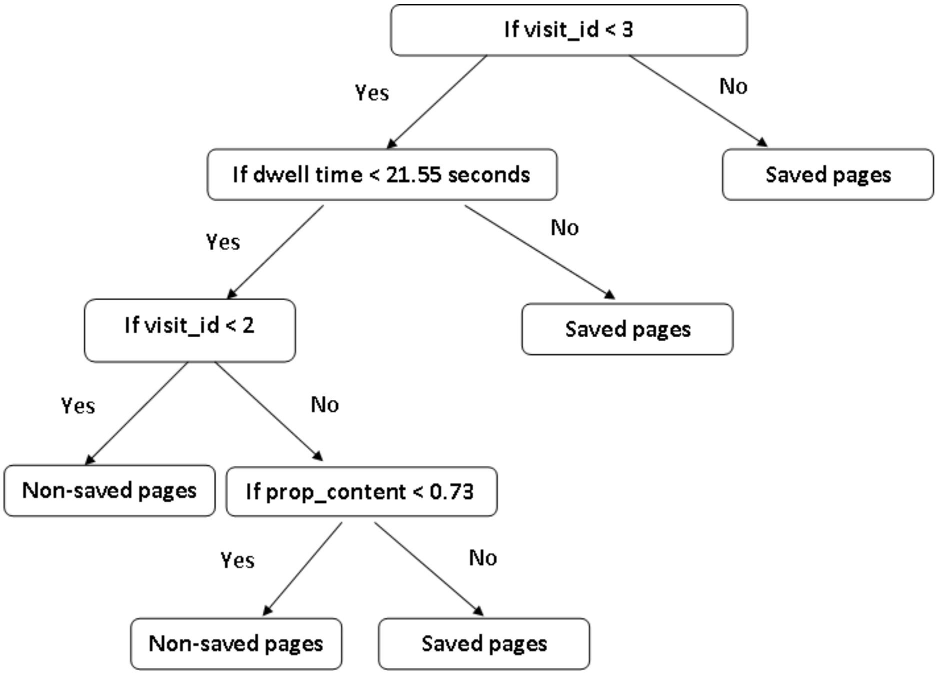


Figure 6. Specific decision-tree predictive model for INT

2) Predictive modeling using logistic regression

We then generated logistic regression models on each of the ten samples, and each sample identified several important predictors of document usefulness in INT task. The selected predictors in each of the samples for INT are shown in Table 91--Table 100 (in Appendix C). Table 18 listed all the selected predictors and the coefficients in the ten samples. It shows that four variables were consistently selected in the logistic regression models: dwell time, visit_id, number of mouse clicks and serp_sum. Another two variables were selected in eight of the ten models: serp_mean and serp_count.

Table 18. The coefficients for selected predictors in the ten samples of INT task

Samples	1	2	3	4	5	6	7	8	9	10
dwell time	0.06	0.05	0.06	0.06	0.05	0.06	0.07	0.06	0.04	0.06
visit_id	0.77	0.88	0.93	0.89	0.68	0.74	0.68	0.82	0.87	0.72
number of mouse clicks	-0.03	-0.03	-0.04	-0.05	-0.04	-0.05	-0.04	-0.04	-0.05	-0.05
serp_sum	0.05	0.04	0.06	0.06	0.07	0.07	0.05	0.06	0.06	0.06
serp_mean	-0.10	-0.10	-0.13		-0.13	-0.15		-0.15	-0.13	-0.13
serp_count			-0.30	-0.28	-0.26	-0.29		-0.37	-0.27	-0.22
number of keystrokes	-0.05									
content_mean								0.05		
content_count					0.06			0.08		
content_sum			-0.006		-0.006			-0.005	-0.004	

We then included only these six variables for logistic regression in the ten samples, and found that serp_count was not significant in many of the cases. Given the relationship among serp_sum, serp_count, and serp_mean, we decided to choose only two of them, and did not include serp_count into the final model. Therefore, the specific logistic regression model for INT task included five behavioral variables: dwell time, visit_id, number of mouse clicks, serp_mean, and serp_sum. The model that obtained the best performance on the whole dataset was selected as the final specific logistic regression model.

Equation 4. Logistic Regression (LR) predictive model of document usefulness for the INT task

$$\ln(\frac{p}{1-p}) =- 1.73 + 0.05 \ast dwelltime + 0.76 \ast visit\_id - 0.04 \ast number.of.mouse.clicks - 0.04 \ast serp\_mean + 0.02 \ast serp\_sum$$

5.4.4 Specific model for the OBI task

1) Predictive modeling using recursive partitioning

We generated predictive models using recursive partitioning based on each of the ten random samples in OBI tasks, and these tree models are shown in Figure 26 (in Appendix B). The behavioral variables that have been selected as predictors in these models are summarized in Table 19, which also counted the number of times each variable was selected in the models. The variables that had been selected more than five times were selected in the final recursive portioning predictive model, and there were three variables in the final model: dwell time, visit_id, and time_to_first_click.

Table 19. The selected predictors from recursive partitioning in ten samples (OBI)

variables	1	2	3	4	5	6	7	8	9	10	total
dwell time	x	x	x	x	x	x	x	x	x	x	10
visit_id	x	x	x	x	x	x	x	x	x	x	10
time_to_first_click		x	x	x	x	x		x	x	x	8
serp_mean	x						x			x	3
serp_sum	x			x							2
diff_content					x		x				2
number.of.mouse.clicks							x				1

content_mean			x								1
content_mean									x		1
content_sum	x										1
interval									x		1

All the three predictors, dwell time, visit_id, and time_to_first_click were positively related to document usefulness in OBI. Similarly as before, the cutoff point for predictors was calculated as the mean of the cutoffs in the models. The cutoff point for dwell time is 21.55 seconds, for visit_id is 2, and for time_to_first_click is 2.11 seconds. The specific tree predictive model for OBI can be shown as Figure 7.

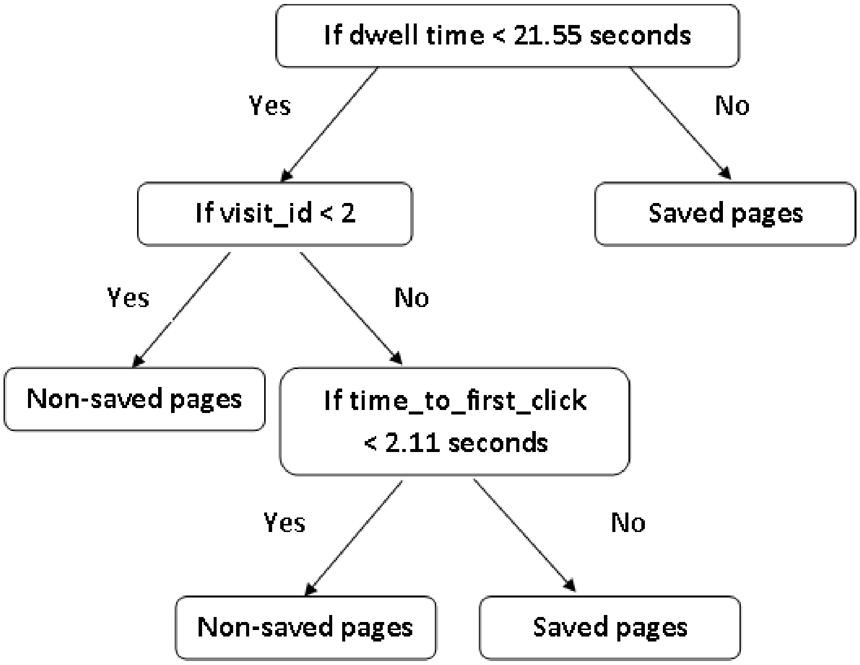


Figure 7. Specific decision-tree predictive model for OBI

2) Predictive modeling using logistic regression

We then generated logistic regression models on each of the ten samples, and each sample identified several important predictors of document usefulness in OBI task. The selected predictors in each of the samples for OBI are shown in Table 101--Table 110 (in Appendix C).

Table 20 listed all the selected predictors and the coefficients in the ten samples. It shows that only two variables were consistently selected in the logistic regression models: dwell time and visit_id. Four other variables were selected in some of the models: content_count, content_sum, serp_mean, and serp_count, but since none of them occurred in more than five samples, we do not include them in the final model.

Table 20. The coefficients for selected predictors in the ten samples of OBI task

Samples	1	2	3	4	5	6	7	8	9	10
dwell time	0.07	0.07	0.08	0.09	0.10	0.08	0.07	0.06	0.08	0.09
visit_id	0.86	0.74	0.75	0.84	1.10	1.01	0.89	0.60	0.93	0.97
content_count			-0.13	-0.11		-0.15	-0.11			-0.14
content_sum										0.007
serp_mean				0.15						
serp_count				0.21		0.22				

Therefore, the specific logistic regression model for OBI task included only two behavioral variables: dwell time and visit_id. The model that obtained the best performance on the whole dataset was selected as the final specific logistic regression model.

Equation 5. Logistic Regression (LR) predictive model of document usefulness for the OBI task

$$\ln(\frac{p}{1-p}) = -\ 2.41\ +\ 0.08\ *\ dwelltime\ +\ 0.84\ *\ visit\_id$$

5.5 *Comparison of predictive models*

We compare the predictive models in this section. We first compared the two general models, General_tree and General_LR, on their prediction performance and the selected predictors. Then we compared the predictive models in each of the four tasks, in particular, we compared the prediction performance and selected predictors by the two general models and two specific models in each task. Finally, the prediction performance was compared across different tasks to examine the task effect on the prediction performance.

The ultimate goal of predicting document usefulness is to personalize search results, thus the prediction performance of document usefulness should be evaluated in retrieval performance and user experience. However, since implementation of the predictive models is out of the scope of this dissertation study, we simply compared the prediction accuracy by the measures following measures.

- The precision of saved pages: the number of cases that were observed as “Saved” and predicted as “Saved” divided by the total number of predicted as “Saved” pages in this task; we want this measure to be as big as possible;

- The precision of non-savd pages: the number of cases that were observed as “Non-Saved” and predicted as “Non-Saved” divided by the total number of predicted as “Non-Saved” pages in this task; we want this measure to be as big as possible;

- The overall accuracy: the number of cases that were predicted the same as the observed situation divided by the number of all cases; we want this measure to be as big as possible;

Because in a Personalized IR system, only the documents that are predicted as useful will be taken to do relevance feedback, the prediction of “Saved” pages is more important than the prediction of “Non-saved” pages. Therefore, when comparing the prediction performance among four predictive models, the precision of “Saved” pages is given top priority in comparing the prediction performance by different models.

We also compared what behavioral measures are important predictors for each type of task, and their relative importance that contributed to the predictive models. In logistic regression models, the coefficients (B) and the associated odds ratio (Exp B) are highly dependent on the unit of measure. Menard discussed six ways to standardize B weights. In this study, we used the simple way to standardize the predictors by multiplying the unstandardized B weight by the predictor's standard deviation, and then we can compare the relative importance of the predictors using the standardized B weights in each of the models.

5.5.1 Discussion of the two general predictive models

The general predictive models are not task specific and we generated them by considering all the interactions in the four tasks. The General Tree model selected dwell time, visit id, and the time to first click as the three most important predictors, and the General LR model selected dwell time, visit id, number of keystrokes, and number of mouse clicks as the four most important predictors. It is apparent that the variables on content pages were more important since only one behavioral measure on search result pages was selected, while all the variables on the content pages were selected. This indicates that if the task type is unknown to the system, then it is more important to monitor users' search interactions on content pages.

Table 21. Prediction performance comparison by the two general predictive models

	General_tree	General_LR
precision of saved pages	43.06%	47.71%
precision of non-saved pages	86.96%	82.91%
overall accuracy	65.48%	70.72%

We also compared the prediction performance by the two general models in Table 21.The general LR model had better prediction than the general tree model. The reason for this might be that the General LR model contains more variables than the General Tree model.

5.5.2 Comparison of four predictive models in the BIC task

1) Comparison of prediction performance

The prediction performance by different models in BIC is listed in Table 22. Comparing between the general tree model and the specific tree model, the specific tree model for BIC had better prediction than the general tree model on BIC; specifically, the specific tree model for BIC had more correctly predicted saved pages, and more correctly predicted non-saved pages. Comparing between the general LR model and the specific LR model, the specific LR model for BIC correctly predicted more saved pages, but it correctly predicted fewer non-saved pages than the general LR model on BIC. If we compare all the models together, we see that the specific tree model for BIC achieved best prediction on saved pages, with the highest precision of saved pages (40.7%) and the highest recall of saved pages (66.8%). So the best predictive model for BIC is the Specific tree model.

Table 22. Prediction performance comparison by general and specific models in BIC

BIC	General_tree	General_LR	Specific_tree	Specific_LR
precision of saved pages	33.1%	34.9%	**40.7%**	36.6%
precision of non-saved pages	85.1%	82.0%	**87.4%**	84.8%
overall accuracy	61.0%	67.1%	**69.5%**	66.4%

2) Comparison of predictors in the model

For the BIC task, the specific predictive model is the decision tree model, which identified four important predictors: dwell time, visit_id, the total dwell time on search result pages during query interval (serp_sum) and the number of mouse clicks on content pages. Referring to Figure 23, dwell time was selected 10 times in the ten random samples, visit_id and serp_sum was selected 9 times, and the number of mouse clicks was selected 8 times. Therefore, we could conclude that for the BIC task, the most important predictor is dwell time, which is also put on the top node of the decision tree; then visit_id and serp_sum have similar contributions to the model, followed by the number of mouse clicks.

5.5.3 Comparison of four predictive models in CPE

1) Comparison of prediction performance

The prediction performance of different models in CPE is listed in Table 23. Comparing between the general tree model and the specific tree model, the specific tree model for CPE correctly predicted more saved pages, but correctly predicted fewer non-saved pages than the general tree model. Comparing between the general LR model and the specific LR model, the general LR model for CPE correctly predicted more saved pages, but it correctly predicted fewer non-saved pages than the specific LR model on CPE. If we compare all the models together, the specific LR model of CPE achieved the highest precision of saved pages (63.5%) and the highest overall accuracy (68.8%). So the specific model for CPE had better prediction performance than the general model, especially the specific LR model.

Table 23. Prediction performance comparison by general and specific models in CPE

CPE	General_tree	General_LR	Specific_tree	Specific_LR
precision of saved pages	52.0%	55.9%	57.6%	**63.5%**
precision of non-saved pages	**80.1%**	74.9%	79.6%	72.0%
overall accuracy	61.4%	64.9%	67.3%	**68.8%**

2) Comparison of predictors in the model

For the CPE task, three predictors were selected in the logistic regression model: dwell time, visit_id and the number of mouse clicks on the content page. As shown in Table 24, one standard deviation (SD) change in dwell time increases the odds of a useful document by a multiplicative factor of 2.48, and one SD increase of visit_id by 1.95, and a one SD increase in number.of.mouse.clicks decreases the odds of a useful document by 0.69. Therefore, comparatively, dwell time is the most important predictor for CPE, followed by visit_id, and then number of mouse clicks. Notice that all of the predictors were behavioral measures on content pages, and none of the behavioral measures on search result pages or related to search result pages were identified for CPE tasks. Thus, for this type of task, how the user interacts with each of the content pages contributes most to the prediction of document usefulness.

Table 24. Relative importance of the predictors in CPE

	Unstandardized		Standard Deviation	Standardized	
Predictor	B	Odds Ratio		Beta	Odds Ratio
dwell time	0.03	1.03	27.42	0.91	2.48
visit_id	0.07	1.07	0.96	0.67	1.95
number.of.mouse.clicks	-0.02	0.98	15.80	-0.37	0.69

5.5.4 Comparison of four predictive models in the INT task

1) Comparison of prediction performance

The prediction performance by different models in INT is listed in Table 25. Comparing between the general tree model and the specific tree model, the general tree model for INT had correctly predicted more saved pages, but correctly predicted fewer non-saved pages than the specific tree model on INT. Comparing between the general LR model and the specific LR model, the specific LR model of INT had better performance than the general LR model for INT; specifically, the specific LR model of INT correctly predicted more saved pages, and it correctly predicted more non-saved pages than the general LR model on INT. If we compare all the models together, the specific LR model of INT achieved the highest precision of saved pages (46.1%) and the highest overall accuracy (73.9%). So the specific model for INT had better prediction performance for the general model, especially the specific LR model.

Table 25. Prediction performance comparison by general and specific models in INT

INT	General_tree	General_LR	Specific_tree	Specific_LR
precision of saved pages	37.7%	43.7%	38.7%	46.1%
precision of non-saved pages	88.4%	85.4%	86.9%	86.5%
overall accuracy	64.7%	72.4%	66.8%	73.9%

2) Comparison of predictors in the model

For the INT task, five predictors were selected in the logistic regression model: dwell time, visit_id, the number of mouse clicks on the content page, the average dwell time on search result pages during the query interval (serp_mean), and the total time on search result pages during the query interval (serp_sum). As shown in Table 26, one standard deviation (SD) change in dwell time increases the odds of a useful document by a multiplicative factor of 2.39, one SD increase of visit_id by 2.44, one SD increase of serp_sum by 1.40, and a one SD increase in number.of.mouse.clicks decreases the odds of a useful document by 0.70, one SD increase of visit_id decreases the odds by 0.76. Therefore, comparatively, dwell time and visit_id were both most important predictors of document usefulness for INT, followed by serp_sum, and then serp_mean and the number of mouse clicks. Compared with CPE tasks, the visit_id showed to be more important, which is similar to dwell time of the page. It also identified two behavioral measures on the search result pages to be important predictors, the total time and the average time the user spends on search result pages. Therefore, for INT tasks, how the user interacts with the search result pages may also indicate the document usefulness.

Table 26. Relative importance of the predictors in INT

	Unstandardized		Standard Deviation	Standardized	
Predictor	B	Odds Ratio		Beta	Odds Ratio
dwell time	0.05	1.05	17.19	0.87	2.39
visit_id	0.76	2.14	1.17	0.89	2.44
number.of.mouse.clicks	-0.04	0.96	9.91	-0.35	0.70
serp_mean	-0.04	0.96	6.45	-0.27	0.76
serp_sum	0.02	1.02	21.87	0.34	1.40

5.5.5 Comparison of four predictive models in the OBI task

1) Comparison of prediction performance

Table 27 listed the prediction performance by different models in OBI. Comparing between the general tree model and the specific tree model, the general tree model for OBI had correctly predicted more saved pages, but correctly predicted fewer non-saved pages than the specific tree model on OBI. Comparing between the general LR model and the specific LR model, the specific LR model of OBI correctly predicted more saved pages, but it correctly predicted fewer non-saved pages than the general LR model on OBI. If we compare all the models together, it showed that general LR model for OBI achieved the highest precision of saved pages (60.1%), and the general tree model for OBI had the highest precision of non-saved pages (89.8%). However, there was not much difference among different models, and the specific models for OBI had a bit higher overall accuracy than the general model. Therefore, we can conclude that the specific model for OBI worked better than the general model.

Table 27. Prediction performance comparison by general and specific models in OBI

	General_tree	General_LR	Specific_tree	Specific_LR
precision of saved pages	54.9%	60.1%	59.0%	59.5%
precision of non-saved pages	89.8%	84.2%	87.7%	85.8%
overall accuracy	74.1%	76.3%	76.6%	76.4%

2) Comparison of predictors in the model

For the OBI task, only two predictors were selected in the logistic regression model: dwell time and visit_id. As shown in Table 28, one standard deviation (SD) change in dwell time increases the odds of a useful document by a multiplicative factor of 3.00, and one SD increase of visit_id by 2.92. Therefore, the two predictors differed little in their unique contributions to the prediction of document usefulness in OBI task. Compared with the logistic regression model for CPE and INT, the predictive model for OBI is very similar to CPE, except that it does not identify the number of mouse clicks as one of the important predictors. For both CPE and OBI, how the user interacts with each of the content pages contributes most to the prediction of document usefulness.

Table 28. Relative importance of the predictors in OBI

	Unstandardized		Standard Deviation	Standardized	
Predictor	B	Odds Ratio		Beta	Odds Ratio
dwell time	0.08	1.08	13.99	1.10	3.00
visit_id	0.84	2.32	1.27	1.07	2.92

5.5.6 Comparison of the prediction performance of the specific models in the four tasks

The prediction performance by the best specific predictive models for each of the four tasks is listed in Table 29. Among these tasks, the CPE task has the highest precision of saved pages, and OBI has the highest overall accuracy, and the highest recall of both saved pages and non-saved pages. Comparatively, BIC has the highest precision of non-saved pages, but the precision of saved-pages in BIC was the lowest.

Table 29. Prediction performance comparison among four tasks

	BIC	CPE	INT	OBI
precision of saved pages	**40.7%**	**63.5%**	**46.1%**	**59.5%**
false positive of saved pages	22.74%	13.65%	16.83%	14.42%
precision of non-saved pages	87.4%	72.0%	86.5%	85.8%
overall accuracy	69.5%	68.8%	73.9%	76.4%
recall of saved pages	66.8%	57.5%	60.8%	69.8%
recall of non-saved pages	70.3%	76.7%	77.9%	79.3%

Since only the predicted saved pages will be taken to do relevance feedback, the prediction performance on the saved pages will greatly affect the retrieval performance.

Therefore, our comparison mainly focuses on whether we could predict saved pages correctly. Comparing the precision of saved pages by the four tasks, we can see the performance varied by task and task facet. The prediction performance in CPE and OBI tasks was acceptable with relatively high precision (63.5% and 59.5%) and low false positives (13.65% and 14.42%); comparatively, the prediction performance in BIC and INT tasks was not very good, in either the precision or the false positive rate. Referring to the task facet values, this result can be explained by the task product facet. With respect to the product facet, CPE and OBI are factual tasks, and BIC and INT are tasks with mixed product (factual and intellectual). This result indicates that the prediction of document usefulness is more difficult in Intellectual tasks than in Factual tasks. However, since participants were asked to save all the best pages that were useful for their tasks, some useful pages might not be saved, or later were deleted, because they were not the best ones, but they were still useful to the tasks. Therefore, further research is needed to check whether the predicted useful pages will be more useful for users to accomplish the Intellectual tasks.

5.6 *Discussion for RQ1*

The first research question of this dissertation was concerned with what behaviors can be used to predict document usefulness, and whether the predictive models should be different for different types of tasks. The section begins with a discussion of the behavioral measures as the predictors of document usefulness. This is followed by the discussion of the task type effect on the predictive models.

5.6.1 Behavioral measures as predictors of document usefulness

In this study, we generated document usefulness predictive models based on a laboratory experiment of the influence of different types of tasks on task session behaviors. We constructed a general predictive model and several specific predictive models tailored to different types of tasks. Recursive partitioning analysis and logistic regression were used to generate the predictive models based on users' interaction behaviors during search sessions. The behavioral measures identified as predictors of document usefulness in our study are summarized in Table 30.

Table 30. The summary table of behavioral measures identified as predictors of document usefulness (+ = positive relationship, - = negative relationship)

Behavioral variables	General RP. Model[4]	General LR. Model[5]	BIC RP. Model	BIC LR. Model	CPE RP. Model	CPE LR. Model	INT RP. Model	INT LR. Model	OBI PR. Model	OBI LR. Model
dwell time on content pages (seconds)	+	+	+	+	+	+	+	+	+	+
number of mouse clicks		-	-		-	-		-		
number of keystrokes		-		-						
visit_id	+	+	+		+	+	+	+	+	+
time_to_ first_click	+								+	
content _mean										
content _sum				-						
content _count				+						
serp_mean								-		
serp_sum			+					+		
serp_count										
prop _content					+		+			
interval _time										

Our models identified several behavioral measures on content pages as important predictors of document usefulness: dwell time, the number of times a page has been visited in the

[4] RP Model stands for Recursive Partitioning Model

[5] LR Model stands for Logistic Recursive Model

session (visit_id), and number of mouse clicks on the content pages. Dwell time has been identified in all the predictive models that were generated from our study. Some previous studies, e.g. Kelly and Belkin (2004) and Liu and Belkin (2010), found the relationship between dwell time and document usefulness was not always significantly positively related, and it varied in different contexts. But in our study, it was found that the longer the dwell time, the more likely the page was useful to the task. One reason for this result is that participants were asked to save the best Web pages that could help them accomplish the task and the assignment, and we only consider Web pages that were saved by the participants as useful pages. Therefore, the useful Web pages in our study were actually the "extremely" useful pages, and it is reasonable to have the result that the extremely useful pages had significantly longer dwell time than other pages. Besides the positive relationship between dwell time and document usefulness, we found that the visit_id was positively related to document usefulness. This is reasonable because if a page has been visited more than once, it is very likely that the page is useful to the task, otherwise the user would not click on the same page the second or third time. Our study also found the number of mouse clicks was an important predictor of document usefulness, but they were shown to have a negative relationship, in all the models that identified number of mouse clicks as an important factor. Previous studies on sources for implicit relevance feedback demonstrated different results on the number of mouse clicks and scrolling as predictors of document usefulness. In a previous study, Claypool (2001) found the amount of scrolling was good indicator of document interests, while the number of mouse clicks was not a good indicator of interests. Fox, et al. (2005) did not observe a significant correlation between the amount of scrolling and relevance. A recent study by Guo and Agichtein (2012) found that the amount of scrolling was not strongly correlated with document usefulness, but there was a significant negative correlation of scrolling frequency and speed and document relevance. One explanation for our result is that, in our study, we counted all the mouse clicks to calculate the number of mouse clicks, including mouse clicks on pages and scrolling using mouse; and a majority of them (about 62%) were scrolling behaviors in our experiment. High frequency of scrolling a page may indicate that the users were scanning the page, but they would “read” the page more carefully if the page is useful.

As shown in our predictive models, some of the behavioral measures during query intervals were also identified as important predictors of document usefulness. These behavioral measures described how users interact with content pages and SERPs before and after each content page click. The identified behavioral measures during query intervals included: time to first click, the total time on content pages during the query interval (content_sum), the total number of content pages during the query interval (content_count), the average dwell time on SERPs (serp_mean), the total time on SERPs during the query interval (serp_sum), and the proportion of time on content pages (prop_content).

In particular, time_to_first_click, which was the time after the user issued a query and before he/she clicked on the first link on the SERP, was found to be positively related to document usefulness. The longer the time_to_first_click, the more likely that the user has made some initial judgment on the document to be clicked from the snippet, and the more likely the clicked document is useful to the task. The variable, content_sum, which was the total time the user spent on content pages during a query interval, was found to be negatively related to document usefulness in BIC task; in contrast, content_count, which was the total number of content pages, was found to be positively related to document usefulness in BIC task. This result is compatible with our previous finding that when there is any useful page found during a query interval, it is very likely that more than one useful page is found during this query interval, but the total time on content pages is not necessarily significantly different (Liu, Gwizdka & Liu, 2010).

The variable, serp_mean, which is the average dwell time on SERPs during a query interval, was found to be negatively related to document usefulness in the INT task. This is reasonable because the average time on SERPs has often been regarded as an indicator of task difficulty, because users tend to spend longer time on each SERP when there is not any useful content page found for that query (Liu, Gwizdka, Liu, Belkin, 2010; Liu et al., 2012). The variable, serp_sum, which is the total time the user spent on SERPs during a query interval, was found to be positively related to document usefulness in two tasks, BIC and INT. This is reasonable because when dwelling on SERPs, a user is reading through the snippets of the search result links before and after he/she clicked on any content pages, and longer reading time on SERPs could indicate there were some potentially useful pages on the SERPs. Otherwise the user would be very likely to leave that SERP and issue a new query if there was no potentially useful page on the SERP. Notice that this variable was only identified in two tasks but not in the other tasks, so there might be some task type effect, which is discussed in the next session. The variable, prop_content, which is the proportion of time a user spends on content pages during a query interval, was found to be positively related to document usefulness in two tasks: CPE and INT. This result also agrees with our previous findings in Liu, Gwizdka & Liu (2010), in which we found users tend to spend a greater proportion of time on content pages when the query interval contained at least one useful document. The task effect on this predictor is examined in the next section.

In summary, we have identified combinations of behavioral measures on content pages and during query intervals as important predictors of document usefulness. Comparatively, the behavioral measures on content pages were found to be more important than behavioral measures during query intervals. This indicates that it is more important for the search system to know how the user interacts with the content pages when predicting the usefulness of the clicked content pages. Consequently, a personalization assistant on the client-side would be more beneficial than on the server-side, because interactions on the content pages are not captured by the logger on the server-side. On the other hand, our results also demonstrate that the behavioral measures during query intervals could be important predictors of document usefulness, because when a useful document is found, it is very likely that more than one useful document is found during that query interval, and users' behaviors would be different before and after each click during that time. However, we should also be aware that different measures during query intervals were selected in different tasks. The reason for this might be because the information objects needed for different tasks varied and the ways to judge document usefulness were also different. The task effect is further discussed in the next section.

In this study, we examined several main behavioral measures on content pages, but it is also possible to extract more behavioral measures on content pages to describe users' interactions more comprehensively in order to make better predictions. For example, we count all the mouse clicks as one type of behavior, but we can divide them into mouse clicks for highlighting, scrolling, clicking on links and so on, and further examine whether separating them could help with better prediction. In addition, users' eye movement captured by eye-tracking may also help us better understand how users read and scan different parts of the page and may help us make better predictions.

5.6.2 Task effect on the predictive models

The specific models for different types of tasks show major differences by task type for both the predictors and the rules. In the specific predictive model for BIC, the total time on content pages (content_sum) was identified to be negatively related to document usefulness, and the total number of content pages (content_count) was identified to be positively related to document usefulness; however, neither of these two variables was identified in any of the other tasks. One explanation for this is that in the BIC task, users were searching to find related news stories on the issue, and they were asked to find as many as possible; so the user could read the content pages briefly and quickly decide whether the page was related or not. Therefore, the content_count was positively related to document usefulness but the content_sum was not. In addition, the total time on SERPs (serp_sum) was identified to be positively related to document usefulness in the RP model, which also indicates that for BIC tasks, users were taking the time on SERPs to help make usefulness judgments of the content pages.

Another task that identified behavioral measures during query intervals is INT, which identified the average time on SERPs (serp_mean), the total time on SERPs (serp_sum) and the proportion of time on content pages (prop_content) during a query interval as important predictors of document usefulness. In particular, the query interval that had shorter average dwell time on SERPs but more total time on SERPs during the query interval was more likely to have useful documents in INT. The INT task is an Intellectual task with amorphous goals, so when users were searching for this task, they explored different sources and different types of information in order to understand the issue and disambiguate the goal through searching. Therefore, they might not spend much time on each SERP and instead they clicked on multiple content pages and read these pages to understand the issue.

In the CPE task, the behavioral measures on content pages were identified to be the most important predictors, and the threshold for dwell time in CPE was much longer than that in other tasks. The only measure identified during query intervals was prop_content, which described the proportion of time spent on content pages during a query interval. This shows that how long the user dwelt on the content pages and a greater proportion of time on content pages were more important in predicting useful pages in CPE than in the other tasks. This is reasonable because the CPE task required users to find information for specific factual information, and the user would read each of the content pages very carefully in order to find the segment information. Therefore, the dwell time was significantly longer on useful pages than non-useful pages.

Similar to the CPE task, the OBI task identified behavioral measures on content pages as important predictors of document usefulness, and only one variable during query interval (time_to_first_click) was shown to be positively related to document usefulness in the RP model. This may be explained by the product facet of the task. OBI and CPE tasks are both factual tasks, and from our models we may infer that for factual tasks, the behavioral measures on the content pages were more important in predicting document usefulness than behavioral measures during query intervals.

In sum, the examination of task type effect on the specific predictive models showed that different combinations of predictors and rules were generated for different types of tasks. The reason for this is that task type affects the type of information objects needed for the task, and it also affects how users search for and judge the information during the search episode. In particular, the task facet of Product was found to significantly influence the prediction predictors. In factual tasks, the behavioral measures on the content pages were more important in predicting useful documents; while in intellectual tasks, how users behaved during query intervals also was indicative of document usefulness.

In practice, the general model would be applied at the beginning of a search episode, and after the information of task type is predicted, the specific predictive models can be applied (as shown in Figure 8). RQ2 of this dissertation study explores what behavioral measures can be used to predict task type so that we can apply these specific predictive models.

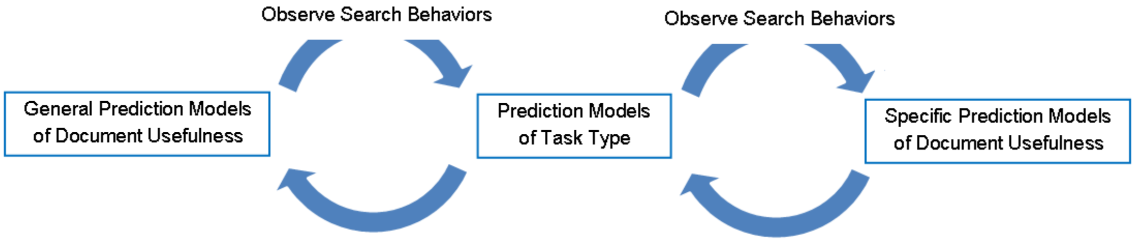

Observe Search Behaviors
Observe Search Behaviors
General Prediction Models of Document Usefulness
Prediction Models of Task Type
Specific Prediction Models of Document Usefulness

Figure 8. The process of implementing predictive models of document usefulness

5.7 *Summary for RQ1*

In RQ1, we generated the predictive models of document usefulness. The predictive models were generated when all tasks were considered as the general predictive model, and for each type of task, as the specific predictive models, using both recursive partitioning and logistic regression modeling. The specific models that achieved the best prediction performance on saved documents were selected as the final predictive models. From the specific models, it is clear that different types of tasks identified different behavioral measures as predictors of useful documents. In particular, the CPE and OBI tasks identified only behavioral measures on each of the content pages as important predictors, and none of the behavioral measures on the search result pages were selected. In contrast, the BIC and INT tasks both identified some behavioral measures on search result pages as important predictors of useful documents, in addition to the behavioral measures on content pages. In addition, the prediction performance in CPE and OBI tasks was much better than that in BIC and INT tasks, especially in correctly predicting the useful pages. This result is reasonable in that users only needed to find factual information in Factual tasks and they only needed to judge the document relevance to decide the document usefulness; therefore, the behaviors on content pages themselves would affect the prediction results more in factual tasks than in tasks with intellectual information. In addition, it might be easier for users to judge document usefulness in Factual tasks than in Intellectual tasks, so our prediction performance on useful documents was better in Factual tasks than in Intellectual tasks. This also indicates that it is important for the systems to learn about the task type, through observing user behaviors or other ways, so as to achieve the best prediction performance.

Chapter 6. Results for RQ2

RQ 2. What user behaviors can be used to predict task type?

In RQ2, we first conducted descriptive analysis of behavioral measures on the whole-session level and the within-session level from the user experiment. Then these behavioral measures were compared among different types of tasks and task facets, to examine the effect of task types on each of the behavioral measures. Finally, we conducted multinomial logistic regression to generate the predictive models of task type using the important behavioral measures as predictors.

6.1 *Descriptive analysis of the selected behavioral measures*

6.1.1 Descriptive analysis of behavioral measures on the whole-session level

The descriptive analysis of each of the behavioral measures on the whole-session level is shown in Table 31. From this table, we can see that when all tasks and all users were considered, it took users an average of 819.7 seconds (about 13.66 minutes) to complete a search task. In each search session, users visited 29 documents on average, 20 unique documents, 28 SERPs, 15 unique SERPs; users issued 14 queries on average. With respect to the total task completion time, a user spent on documents 418.3 seconds on average (about 52% of the task completion time), and spent on SERPs 283.5 seconds on average (about 32% of the task completion time).

The histograms and the Q-Q plots for each of the variables on the whole-session level are shown in Figure 27 to Figure 36 (in Appendix D) to examine whether the distribution of each variable was normal or not. They demonstrate that except for the two variables: the ratio of document time to all, and the ratio of SERP time to all, all the other variables were not normally distributed. Therefore in the following statistical analysis, we need to conduct non-parametric analysis to compare the distribution among different groups.

Table 31. Descriptive statistics of behavioral measures on the whole-session level

	Minimum	Maximum	Mean	Median	Std. Deviation	Shapiro. Test
Task completion time (seconds)	121.9	1743	819.7	758.5	423.2	p<.05
Numbers of all documents	2	86	29.38	27.5	18.28	p<.05
Numbers of unique documents	1	61	20.12	19	11.68	p<.05
Number of SERPs	2	85	28.64	25	18.67	p<.05
Number of unique SERPs	1	56	15.25	12	11.12	p<.05
Number of queries	1	46	14.11	11	9.72	p<.05
Total time spent on documents (seconds)	21.7	973.9	418.3	399.1	228.40	p<.05
Total time spent on SERPs (seconds)	9.5	1068.7	283.5	241.9	203.37	p<.05
Ratio of document time to all	0.14	0.90	0.52	0.51	0.14	p=0.41
Ratio of SERP time to all	0.03	0.68	0.33	0.32	0.14	p=0.46

6.1.2 Descriptive analysis of behavioral measures on the within-session level

Within-session level behavioral measures were studied because these measures can be calculated at any point during the search session, which is necessary for personalization during the search. For example, the mean dwell time of all documents can be calculated by averaging the dwell time on all the documents that have been visited at the point of calculation. In the current study, we only calculated these measures at the point when the task is completed for reasons of simplicity; but, in principle, they could be computed during the search session. The descriptive analysis of each of the behavioral measures on the within-session level is shown in

Table 32. From this table, we can see that when all tasks and all users are considered, users spent an average of 17.76 seconds on each document they visited, and spent totally 24.63 seconds on each unique document; users spent 10.07 seconds on each of the SERPs they visited, and totally 19.44 seconds on each of the unique SERPs they visited. During each search session, the average query interval time was 78.59 seconds; and on average, after each query, users visited 2.66 documents, 1.82 unique documents, 2.16 SERPs, and1.08 unique SERPs. In addition, after issuing one query, it took users an average of 10.49 seconds to click on the first link (either on document, the next page of the SERP or issued another query).

The histograms and the Q-Q plots for each of the variables on the within-session level are shown in Figure 37 to Figure 46 (in Appendix E) to examine whether the distribution of each variable was normal or not. They demonstrate that all the variables were not normally distributed. Therefore in the following statistical analysis, we need to conduct non-parametric analysis to compare the distribution among different groups.

Table 32. Descriptive statistics of behavioral measures on the within-session level

	Minimum	Maximum	Mean	Median	Std. Deviation	Shapiro. Test
Mean dwell time of all documents (seconds)	4.34	147.8	17.76	13.95	14.66	p<.05
Mean dwell time of unique documents (seconds)	5.425	147.8	24.63	20.27	16.73	p<.05
Mean dwell time of all SERPs (seconds)	3.043	31.9	10.07	9.041	4.57	p<.05
Mean dwell time of unique SERPs (seconds)	5.38	63.8	19.44	17.64	8.35	p<.05
Number of documents per query	0.5652	20	2.66	2	2.29	p<.05
Number of unique documents per query	0.5	11	1.82	1.407	1.41	p<.05
Number of SERPs per query	1	5.4	2.16	2	0.74	p<.05
Number of unique SERPs per query	0.64	2.4	1.08	1	0.26	p<.05
Average query interval (seconds)	20.67	508.5	78.59	60.57	67.5	p<.05
Average Time To First Click	2.49	34.53	10.49	10.49	5.13	p<.05

6.2 *Comparison of search behaviors by tasks*

6.2.1 Overall comparison by task

We first used univariate analyses to compare each of the behavioral measures in the four tasks. Kruskal-Wallis tests were selected because the data were not normal distributed. Significant differences were found for all the measures on the whole-session level across the four tasks. However, the Kruskal-Wallis test did not specifically indicate which pairs of tasks were significantly different, so we conducted Tamhane's test to determine such pairs because it is a suitable post-hoc test for the Mann-Whitney test. [6]. The Tamhane's post-hoc analyses found that users spent a significantly longer time to accomplish BIC than the other three tasks, while there is no difference in time in the other three. Users visited significantly fewer content pages in CPE than in the other three tasks, while there was no difference among the other three. They visited significantly more unique content pages in BIC and INT than in CPE and OBI. Number of SERPs, number of unique SERPs, and number of queries had a similar pattern: searchers issued significantly more queries, visited significantly more SERPs and more unique SERPs in BIC and OBI than in INT and CPE. There was no significant difference among different tasks on the total time on content pages. Users spent significantly longer total time on SERPs in BIC than other tasks, and followed by OBI, which also had significantly longer total time on SERPs than CPE and INT. Users spent a significantly larger ratio of time on content pages in CPE and INT than in BIC and OBI; while they spent a significantly larger ratio of time on SERPs in BIC and OBI than in CPE and INT.

[6] http://privatewww.essex.ac.uk/~scholp/kw_posthoc.htm

Table 33. Comparison of behavioral variables on whole-session level among tasks

Whole-session variables	BIC	CPE	INT	OBI	Kruskal-Wallis Test
Task completion time (seconds)	1251.70 (381.27)	469.35 (420.34)	718.05 (377.91)	816.35 (343.85)	<.001
Numbers of all documents	37.50 (16.51)	12.50 (13.04)	34.00 (19.07)	27.50 (17.22)	<.001
Numbers of unique documents	30.00 (9.53)	10.00 (7.96)	23.00 (12.41)	21.50 (11.43)	<.001
Number of SERPs	37.50 (18.14)	13.50 (11.09)	19.50 (13.23)	36.00 (18.42)	<.001
Number of unique SERPs	20.50 (11.41)	6.00 (5.49)	10.00 (6.56)	19.50 (12)	<.001
Number of queries	18.50 (10.93)	6.00 (5.46)	10.50 (6.57)	15.00 (9.66)	<.001
Total time spent on documents (seconds)	458.95 (204.82)	284.65 (253.56)	383.55 (242.89)	413.20 (192.49)	0.05
Total time spent on SERPs (seconds)	489.00 (214.22)	112.55 (138.89)	206.85 (126.99)	267.55 (179.69)	<.001
Ratio of document time to all	0.42 (0.12)	0.62 (0.12)	0.60 (0.14)	0.46 (0.13)	<.001
Ratio of SERP time to all	0.42 (0.12)	0.24 (0.12)	0.26 (0.12)	0.41 (0.13)	<.001

Significant differences were found for the measures on the within-session level across the four tasks, except Number of SERPs per query and Average time to first click. Post-hoc analyses using Tukey's test found that users spent significantly longer dwell time on content pages and unique content pages in CPE than in the other three tasks on average. There was no significant difference in dwell time on SERPs or unique SERPs among tasks. Users visited the most content pages per query in INT, while they visited the fewest content pages per query in OBI, and there was no significant difference in BIC and CPE and among others. Even though there was no significant difference among tasks on the number of SERPs per query, users visited significantly more unique SERPs per query in BIC and OBI than in CPE and INT. Users had significantly longer query interval time in CPE than in OBI; and there was no significant difference on the average time to first click among tasks.

Table 34. Comparison of behavioral variables on within-session level among tasks

Within-session variables	BIC	CPE	INT	OBI	Kruskal-Wallis Test
Mean dwell time of all documents (seconds)	13.69 (5.31)	23.26 (24.29)	12.32 (7.35)	11.95 (4.66)	<.001
Mean dwell time of unique documents (seconds)	17.25 (7.01)	31.6 (25.6)	17.86 (8.51)	15.98 (7.09)	<.001
Mean dwell time of all SERPs (seconds)	10.76 (3.74)	9.02 (5.4)	8.51 (5.45)	7.93 (2.89)	<.001
Mean dwell time of unique SERPs (seconds)	20.26 (5.13)	17.2 (10.34)	16.19 (10.42)	15.01 (5.81)	<.001
Number of documents per query	1.72 (2.17)	2.00 (1.73)	2.92 (3.29)	1.67 (1.17)	<.001
Number of unique documents per query	1.37 (1.59)	1.47 (1.07)	1.93 (1.95)	1.21 (0.9)	<.001
Number of SERPs per query	2.07 (0.92)	2.00 (0.68)	1.84 (0.5)	2.27 (0.78)	0.07
Number of unique SERPs per query	1.06 (0.33)	1.00 (0.14)	1.00 (0.09)	1.07 (0.3)	<.001
Average query interval (seconds)	54.89 (49.87)	82.78 (85.22)	69.11 (80.87)	44.29 (22.54)	<.001
Average Time To First Click	11.23 (4.82)	11.05 (5.51)	10.01 (5.92)	9.66 (4.16)	0.23

6.2.2 Comparison by task facets

The values of the varied facets for each of the four search tasks we designed are shown in Table 4. In this part, we compare the behavioral measures on both whole-session level and within-session level by each of the task facets.

1) Products of search tasks

First we examined the task Product facet. With respect to the whole-session level, it was found that users spent a significantly longer time to complete Mixed-Product tasks than Factual tasks. They visited significantly more documents and more unique documents, and issued more queries in Mixed-Product tasks than in Factual tasks. However, the number of SERPs and the number of unique SERPs did not show differences, nor did the number of search sources they used. Users spent longer times on documents and SERPs in Mixed-Product tasks than in Factual tasks, but there was no significant difference in the ratio of document time to all or the ratio of SERP time to all.

Table 35. Comparison of behavioral variables on whole-session level by product

	Task Product		Wilcoxon signed-rank test
	Factual (CPE&OBI)	Mixed (BIC & INT)	
Task completion time (seconds)	701.72 (390.18)	937.71 (424.88)	<.001
Numbers of all documents	22.98 (16.61)	35.77 (17.74)	<.001
Numbers of unique documents	15.73 (10.79)	24.5(10.93)	<.001
Number of SERPs	26.47 (18.77)	30.81 (18.46)	.125
Number of unique SERPs	13.89 (11.28)	16.61 (10.87)	.073
Number of queries	12.52 (9.25)	15.70 (9.99)	.043
Total time spent on documents (seconds)	371.81 (223.32)	464.86 (225.59)	.019
Total time spent on SERPs (seconds)	236.83 (179.86)	330.22 (215.85)	.011
Ratio of document time to all	0.53 (0.14)	0.51 (0.15)	.399
Ratio of SERP time to all	0.32 (0.15)	0.34 (0.13)	.526

With respect to the within-session level, it was found that users had significantly longer mean dwell times on all documents and on unique documents, while they had significantly shorter mean dwell times on all SERPs in Factual tasks than in Mixed-Product tasks. In addition, users visited significantly more documents and more unique documents per query in Mixed-Product tasks than in Factual tasks.

Table 36. Comparison of behavioral variables on within-session level by product

	Task Product		Wilcoxon signed-rank test
	Factual (CPE&OBI)	Mixed (BIC & INT)	
Mean dwell time of all documents (seconds)	21.23 (19.19)	14.29 (6.36)	.004
Mean dwell time of unique documents (seconds)	29.22 (21.26)	20.04 (8.34)	.002
Mean dwell time of all SERPs (seconds)	9.30 (4.35)	10.84 (4.69)	.019
Mean dwell time of unique SERPs (seconds)	18.48 (8.47)	20.39 (8.19)	.073
Number of documents per query	2.24 (1.49)	3.08 (2.83)	.036
Number of unique documents per query	1.56 (1.00)	2.07 (1.69)	.017
Number of SERPs per query	2.23 (0.73)	2.1 (0.76)	.265
Number of unique SERPs per query	1.09 (0.25)	1.07 (0.26)	.181
Average query interval (seconds)	80.42 (68.58)	76.76 (66.89)	.696
Average Time To First Click	9.77 (4.52)	11.20 (5.62)	.137

2) Objective task complexity

We then compared the behaviors by task complexity. First, task complexity affected most of the behavioral measures on the whole-session level. In particular, users took significantly longer to complete tasks with high complexity than tasks with low complexity. In addition, users visited more unique documents, more SERPs, more unique SERPs, issued more queries for tasks with high complexity than for tasks with low complexity. We also found that users spent a greater portion of time on documents for tasks with low complexity than tasks with high complexity, while spending a lower portion of time on SERPs in tasks with low complexity than tasks with high complexity.

Table 37. Comparison of behavioral variables on whole-session level by complexity

	Task Complexity		Wilcoxon signed-rank test
	Low (CPE & INT)	High (BIC & OBI)	
Task completion time (seconds)	680.24(401.43)	959.2(400.47)	<.001
Numbers of all documents	25.39(18.66)	33.36(17.12)	.006
Numbers of unique documents	16.27(11.19)	23.97(10.94)	<.001
Number of SERPs	18.31(12.46)	38.97(18.19)	<.001
Number of unique SERPs	9.2(6.24)	21.3(11.66)	<.001
Number of queries	9.42(6.28)	18.8(10.32)	<.001
Total time spent on documents (seconds)	399.77(248.21)	436.9(207.02)	.179
Total time spent on SERPs (seconds)	179.21(134.44)	387.84(207.82)	<.001
Ratio of document time to all	0.58(0.13)	0.45(0.13)	<.001
Ratio of SERP time to all	0.26(0.12)	0.41(0.12)	<.001

With respect to the behavioral measures on the within-session level, we found that users had significantly longer mean dwell times on all documents in tasks with low complexity than in tasks with high complexity; while users had significantly longer mean dwell times on unique documents in tasks with low complexity than in tasks with high complexity. There was no significant difference in the mean dwell time of SERPs between tasks with low and high complexity. In addition, the average query interval was significantly longer in tasks with low complexity than in tasks with high complexity. Users visited more documents and more unique documents per query, while visiting fewer SERPs and unique SERPs per query in tasks with low complexity than in tasks with high complexity.

Table 38. Comparison of behavioral variables on within-session level by complexity

	Task Complexity		Wilcoxon signed-rank test
	Low (CPE & INT)	High (BIC & OBI)	
Mean dwell time of all documents (seconds)	21.8(19.36)	13.72(5.00)	.001
Mean dwell time of unique documents (seconds)	30.57(21.00)	18.69(7.20)	<.001
Mean dwell time of all SERPs (seconds)	10.08(5.38)	10.06(3.62)	.394
Mean dwell time of unique SERPs (seconds)	19.82(10.3)	19.05(5.86)	.708
Number of documents per query	3.09(2.67)	2.22(1.75)	.001
Number of unique documents per query	2.02(1.50)	1.61(1.29)	.007
Number of SERPs per query	2.01(0.60)	2.32(0.84)	.016
Number of unique SERPs per query	0.98(0.12)	1.18(0.31)	<.001
Average query interval (seconds)	96.15(83.56)	61.04(39.70)	<.001
Average Time To First Click	10.09(5.66)	10.88(4.56)	.133

3) Level of document judgment

When comparing behavioral measures on the whole-session level by the level of document judgment, we found nearly all the measures, except the total time on documents, were significantly different between tasks on Document level and Segment level. In particular, users spent longer time to complete the tasks, visited more documents, more SERPs, and issued more queries in tasks on Document level than in tasks on Segment level. However, these results may also be due to the task complexity facet. Interestingly, we found that users spent a significantly greater proportion of time on documents in Document-level tasks than in Segment-level tasks;

while they spent a significantly lower proportion of time on SERPs in Document-level tasks than in Segment-level tasks.

Table 39. Comparison of behavioral variables on whole-session level by level

	Level of judgment		
	Segment (CPE)	Document (BIC, INT, OBI)	Wilcoxon signed-rank test
Task completion time (seconds)	617.99(420.34)	886.96(404.36)	.001
Numbers of all documents	16.22(13.04)	33.76(17.7)	<.001
Numbers of unique documents	10.94(7.67)	23.18(11.19)	<.001
Number of SERPs	15.38(11.09)	33.06(18.63)	<.001
Number of unique SERPs	7.5(5.49)	17.83(11.33)	<.001
Number of queries	7.56(5.46)	16.29(9.87)	<.001
Total time spent on documents (seconds)	369.33(253.56)	434.67(218.36)	.072
Total time spent on SERPs (seconds)	154(138.89)	326.7(203.61)	<.001
Ratio of document time to all	0.59(0.12)	0.49(0.14)	.001
Ratio of SERP time to all	0.24(0.12)	0.36(0.14)	.001

With respect to the within-session level variables, we found that users spent significantly longer mean dwell time on all documents and on unique documents in Segment-level tasks than in Document-level tasks. In addition, the average query interval was significantly longer in Segment-level tasks than in Document-level tasks.

Table 40. Comparison of behavioral variables on within-session level by level

	Level of judgment		Wilcoxon signed-rank test
	Segment (CPE)	Document (BIC, INT, OBI)	
Mean dwell time of all documents (seconds)	29.37(24.29)	13.89(5.85)	<.001
Mean dwell time of unique documents (seconds)	39.7(25.26)	19.61(8.01)	<.001
Mean dwell time of all SERPs (seconds)	9.97(5.4)	10.1(4.29)	.672
Mean dwell time of unique SERPs (seconds)	20.07(10.34)	19.22(7.63)	.702
Number of documents per query	2.51(1.73)	2.71(2.46)	.989
Number of unique documents per query	1.75(1.07)	1.84(1.51)	.901
Number of SERPs per query	2.11(0.68)	2.18(0.77)	.491
Number of unique SERPs per query	1(0.14)	1.11(0.28)	.065
Average query interval (seconds)	109.84(85.22)	68.18(57.29)	<.001
Average Time To First Click	10.24(5.31)	10.57(5.1)	.768

4) Task Goal (quality)

The comparison of behavioral measures on the whole-session level revealed that the number of SERPs, the number of unique SERPs, the number of queries, ratio of document time to all and ratio of SERP time to all were significantly different in the different categories of goal (quality). The Tamhane’s post-hoc analysis found that users visited significantly more SERPs and more unique SERPs, and issued more queries in tasks with Amorphous goal(s) than in tasks with Mixed goal(s). In addition, users spent a significantly lower percent of time on documents and greater percent of time on SERPs in tasks with Amorphous goal(s) than in tasks with Mixed goals(s).

Table 41. Comparison of behavioral variables on whole-session level by goal (quality)

	Task goal (quality)			
	Specific (BIC & CPE)	Mixed (INT)	Amorphous (OBI)	Kruskal-Wallis Test
Task completion time (seconds)	875.47 (475.2)	742.48 (377.91)	785.45 (343.85)	.339
Numbers of all documents	26.59 (18.09)	34.56 (19.07)	29.75 (17.22)	.118
Numbers of unique documents	19.17 (11.89)	21.59 (11.71)	20.53 (11.42)	.567
Number of SERPs	27.88 (19.52)	21.25 (13.23)	37.56 (18.42)	.001
Number of unique SERPs	14.91 (11.6)	10.91 (6.56)	20.28 (12)	.003
Number of queries	13.84 (10.66)	11.28 (6.57)	17.47 (9.66)	.034
Total time spent on documents (seconds)	434.42 (237.87)	430.2 (242.89)	374.29 (192.49)	.661
Total time spent on SERPs (seconds)	305.01 (235.03)	204.42 (126.99)	319.67 (179.69)	.053
Ratio of document time to all	0.52(0.14)	0.57(0.14)	0.47(0.13)	.012
Ratio of SERP time to all	0.32(0.15)	0.28(0.12)	0.41(0.13)	.001

When comparing behavioral measures on the within-session level by goal (quality), we found that all these measures were significantly different in the different categories of goal (quality). The Tamhane’s post-hoc analysis found that users had significantly longer mean dwell times on all documents and on unique document in tasks with Specific goals than in tasks with Mixed or Amorphous goals. In addition, users had significantly longer mean dwell times on all SERPs and on unique SERPs in tasks with Specific goals than in tasks with Amorphous goals. Users visited significantly more documents and more unique documents per query, but visited significantly fewer SERPs and unique SERPs in tasks with Mixed goals than in tasks with

Amorphous goals. The average query interval time was significantly longer in tasks with Specific goals than in tasks with Amorphous goals.

Table 42. Comparison of behavioral variables on within-session level by goal (quality)

	Task goal (quality)			
	Specific (BIC & CPE)	Mixed (INT)	Amorphous (OBI)	Kruskal-Wallis Test
Mean dwell time of all documents (seconds)	21.85(19.01)	14.24(7.35)	13.09(4.66)	<.001
Mean dwell time of unique documents (seconds)	29.17(21.26)	21.43(9.27)	18.75(7.34)	.007
Mean dwell time of all SERPs (seconds)	10.73(4.67)	10.18(5.45)	8.63(2.89)	.027
Mean dwell time of unique SERPs (seconds)	20.64(8.12)	19.57(10.42)	16.89(5.81)	.011
Number of documents per query	2.49(1.95)	3.68(3.29)	1.97(1.17)	<.001
Number of unique documents per query	1.8(1.34)	2.3(1.8)	1.38(0.9)	.001
Number of SERPs per query	2.2(0.8)	1.91(0.5)	2.34(0.78)	.044
Number of unique SERPs per query	1.09(0.27)	0.97(0.09)	1.18(0.3)	<.001
Average query interval (seconds)	90.46(71.97)	82.45(80.87)	51(22.54)	<.001
Average Time To First Click	11.36(5.2)	9.94(6.07)	9.3(3.58)	.046

6.3 *Predicting individual task types*

In the univariate analyses, we found that many of the behavioral measures, on both whole-session level and the within-session level, were significantly affected by task type and task facets. However, it is not clear which variables are significant in predicting task type in the previous analysis. In this part, we conducted forward-stepwise multinomial logistic regression to answer the research question of which behavioral measures can be used to predict task type.

Multinomial logistic regression (MLR) is a categorical data analysis method used when there are three or more unordered categories in the dependent variable. Stepwise selection involves analysis at each step to determine the contribution of the predictor variable entered previously in the equation. The forward stepwise method begins with the model that would be selected by the forward entry method. From there, the algorithm alternates between backward elimination on the stepwise terms in the model and forward entry on the terms left out of the model. This continues until no terms meet the entry or removal criteria. In this way it is possible to understand the contribution of the previous variables now that another variable has been added. Variables can be retained or deleted based on their statistical importance. Stepwise MLR is used because this is exploratory research, and the goal of this research is to discover relationships between the variables.

Multinomial logistic regression is often considered an attractive analysis method because it does not assume normality, linearity, or homoscedasticity. MLR requires selecting one category in the dependent variable as the reference group, and in our study, we selected CPE as the reference group because CPE is a Known-item type of task, which requires users to find the factual information with specific goal(s), on the Segment level and for Named information objects. Current search systems and search engines are most helpful for this type of task, but it will be very helpful if we can identify the other three types of tasks. We used the MLR method to explore whether we can distinguish the other three types of tasks from the Known-item search task. In our analysis, we first conduct forward-stepwise MLR using the whole-session level behavioral measures, then conduct forward-stepwise MLR using the within-session level behavioral measures, and finally conduct forward-stepwise MLR using behavioral measures on both whole-session and within-session levels.

6.3.1 Prediction using whole-session level measures

We considered ten behavioral measures on the whole-session variables to generate the predictive model for task type: task completion time, number of all documents, number of unique documents, number of SERPs, number of unique SERPs, number of queries, total time spent on documents, total time spent on SERPs, ratio of document time to all, and ratio of SERP time to all. Using a guideline provided by Hosmer and Lemeshow (2000), the minimum number of cases per independent variable is 10. In our dataset, the independent variable, task type, includes four levels: BIC, CPE, INT, OBI; and we have 32 sessions for each task type, so the dataset meets this requirement.

Table 43. The Step Summary table for whole-session level measures

Model	Action	Effect(s)	Model Fitting Criteria	Effect Selection Tests		
			-2 Log Likelihood	Chi-Square^{a,b}	df	Sig.
Step 0	Entered	Intercept	354.891	.		
Step 1	Entered	Number of unique SERPs	299.538	55.353	3	.000
Step 2	Entered	Task completion time	277.011	22.528	3	.000
Step 3	Entered	Numbers of all documents	250.823	26.188	3	.000
Step 4	Entered	Numbers of unique documents	240.233	10.589	3	.014
Step 5	Entered	Ratio of document time to all	230.317	9.916	3	.019

Stepwise Method: Forward Stepwise
a. The chi-square for entry is based on the likelihood ratio test.
b. The chi-square for removal is based on the likelihood ratio test

The Step Summary table (Table 43) shows the variable added at each step. The entry order of the variables included in the stepwise logistic regression demonstrates the order of variable importance. Therefore, the most important whole-session level variable in distinguishing the four types of tasks in our experiment is number of unique SERPs, followed by task completion time, number of documents, number of unique documents, and ratio of time spent on documents to all.

Table 44. Model Fitting Information in MLR on whole-session level

	Model Fitting Criteria	Likelihood Ratio Tests		
Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	354.891			
Final	230.317	124.574	15	.000

Table 44 presents the likelihood ratio test, which assessed the model fit in MLR. A greater amount of change between the two models suggests a greater improvement in model fit. Here we see that the final model is significantly different from the intercept-only model (p <.001). Thus, the independent variables we identified for the whole-session level, as a group, contribute significantly to prediction of task type.

Table 45. Likelihood Ratio Tests in MLR on whole-session level

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	230.647	.330	3	.954
Numbers of all documents	239.288	8.971	3	.030
Task completion time	253.215	22.898	3	.000
Numbers of unique documents	242.210	11.893	3	.008
Number of unique SERPs	249.723	19.406	3	.000
Ratio of document time to all	240.233	9.916	3	.019

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

Likelihood ratio tests (shown in Table 45) present the significance of the independent variables in the model. It tests the improvement in the model fit with each of the predictor variables added. Statistical significance (<.05) was found for each of the independent variables we selected on the whole-session level, therefore, significant relationships were found between these variables and task type.

Table 46. Parameter Estimates in MLR on the whole-session level

task[a]		B	Std. Error	Sig.	Exp(B)
BIC	Intercept	1.029	1.962	.600	
	Numbers of all documents	-.064	.069	.356	.938
	Task completion time	-.002	.002	.125	.998
	Numbers of unique documents	.367	.121	.002	1.444
	Number of unique SERPs	.117	.091	.196	1.124
	Ratio of document time to all	-10.946	3.772	.004	1.763E-005
INT	Intercept	.786	1.689	.642	
	Numbers of all documents	.082	.060	.172	1.085
	Task completion time	-.005	.002	.003	.995
	Numbers of unique documents	.166	.104	.108	1.181
	Number of unique SERPs	.039	.094	.676	1.040
	Ratio of document time to all	-3.891	2.906	.181	.020
OBI	Intercept	.560	1.778	.753	
	Numbers of all documents	.031	.065	.640	1.031
	Task completion time	-.007	.002	.000	.993
	Numbers of unique documents	.194	.115	.091	1.214
	Number of unique SERPs	.244	.093	.008	1.276
	Ratio of document time to all	-4.828	3.255	.138	.008

*CPE as the reference category

Table 46 presents the parameter estimates for the final model. MLR computed different estimates for all paired groupings of the tasks; specifically the comparison between BIC and CPE, between INT and CPE, between OBI and CPE. With respect to the comparison between BIC and CPE, it was found that the BIC and CPE tasks were different in two variables, number of unique documents and the ratio of document time to all, but no other variables. Specifically, search sessions that had more unique documents visited but a lower percent of time on documents were more likely to be in the BIC task, rather than in the CPE task. For every one more unique document visited in one search session, the odds of being in BIC increased by 44.4% (1.444-1.0=0.444), shown in the value of Exp (B). In addition, for every one percent increase of time spent on documents in the search session, the odds of being in BIC decreased by 98 % (0.02-1.0= -0.98).

With respect to the comparison between INT and CPE, it was found that the INT and CPE tasks were different in one variable, task completion time, but no other variables. Specifically, search sessions that had shorter task completion time were more likely to be in the INT task, rather than in the CPE task. For every one second increase in the task completion time of the search session, the odds of being in INT decreased by 0.5% (0.995-1.0=-0.005).

With respect to the comparison between OBI and CPE, it was found that the OBI and CPE tasks were different in two variables: task completion time and number of unique SERPs, but no other variables. Specifically, search sessions that had shorter task completion time, but more unique SERPs were more likely to be in the OBI task, rather than in the CPE task. For every one second increase in the task completion time, the odds of being in OBI decreased by 0.7% (0.993-1.0=-0.007); for every one more unique SERP visited in a search session, the odds of being in OBI increased by 27.6% (1.276-1.0 = 0.276). The MLR whole session predictive models are shown as below:

Equation 6. Predictive models of task type on the whole-session level

$$\text{odds for BIC} = \ln\left(\frac{Probability\ of\ BIC}{Probablity\ of\ CPE}\right)$$
$$= 1.029 - 0.064 * Number\ of\ all\ documents - 0.002 * Task\ completion\ time + 0.367 * Number\ of\ unique\ documents + 0.117 * Number\ of\ unique\ SERPs - 10.946 * Ratio\ of\ document\ time\ to\ all$$

$$\text{odds for INT} = \ln\left(\frac{Probability\ of\ INT}{Probablity\ ofCPE}\right)$$
$$= 0.786 + 0.082 * Number\ of\ all\ documents - 0.005 * Task\ completion\ time + 0.166 * Number\ of\ unique\ documents + 0.039 * Number\ of\ unique\ SERPs - 3.891 * Ratio\ of\ document\ time\ to\ all$$

$$\text{odds for OBI} = \ln\left(\frac{Probability\ of\ OBI}{Probablity\ of\ CPE}\right)$$
$$= 056 + 0.031 * Number\ of\ all\ documents - 0.007 * Task\ completion\ time + 0.194 * Number\ of\ unique\ documents + 0.244 * Number\ of\ unique\ SERPs - 4.828 * Ratio\ of\ document\ time\ to\ all$$

Another indicator of the usefulness of the final model is the classification table. As shown in Table 47, the observed versus the predicted groupings are compared. Overall, the final model accurately predicted 63.3% of the cases. We can compare this accuracy with the rate of accuracy achievable by chance alone, which can be calculated as: $0.25^2 + 0.25^2 + 0.25^2 + 0.25^2 =$ 25%). Therefore, the final model based on the whole-session variables had an improvement of 153% over the proportional by chance accuracy.

Table 47. Classification table in MLR on whole-session level

Observed	Predicted				
	BIC	CPE	INT	OBI	Percent Correct
BIC	21	2	5	4	65.6%
CPE	1	26	4	1	81.2%
INT	5	7	17	3	53.1%
OBI	7	3	5	17	53.1%
Overall Percentage	26.6%	29.7%	24.2%	19.5%	63.3%

In summary, we found a statistically significant overall relationship between the behavioral measures on the whole-session level and the task type, especially for the following four variables: task completion time, number of unique documents, number of unique SERPs and ratio of time spent on documents to all. Compared with CPE, the BIC task was more likely to have more unique documents visited but a lower percentage of time on documents in a search session; the INT task was more likely to have shorter task completion time in a search session; and the OBI task was more likely to have more unique SERPs visited but shorter task completion time in one search session.

6.3.2 Prediction using within-session level measures

We considered ten behavioral measures on the within-session level to generate the predictive model for task type: mean dwell time of all documents, mean dwell time of unique documents, mean dwell time of all SERPs, mean dwell time of unique SERPs, number of documents per query, number of unique documents per query, number of SERPs per query, number of unique SERPs per query, average query interval, average time to first click. The *Step Summary* table (Table 48) showed which variable was added at each step. The entry order of the variables included in the stepwise logistic regression also demonstrates the order of variable importance.

Therefore, the most important within-session level variable in distinguishing the four types of tasks at the within-session level in our experiment is mean dwell time of unique documents, followed by number of unique SERPs per query, average query interval, average time to first click and number of documents per query.

Table 48. The Step Summary table for within-session level measures

Model	Action	Effect(s)	Model Fitting Criteria	Effect Selection Tests		
			-2 Log Likelihood	Chi-Square[a,b]	df	Sig.
Step 0	Entered	Intercept	354.891	.		
Step 1	Entered	Mean dwell time of unique documents	307.764	47.128	3	.000
Step 2	Entered	Number of unique SERPs per query	285.491	22.272	3	.000
Step 3	Entered	Average query interval	267.230	18.261	3	.000
Step 4	Entered	Average Time To First Click	253.331	13.899	3	.003
Step 5	Entered	Number of documents per query	244.607	8.724	3	.033

*Stepwise Method: Forward Stepwise

a. The chi-square for entry is based on the likelihood ratio test.

b. The chi-square for removal is based on the likelihood ratio test.

The likelihood ratio test is shown in Table 49. Here we see that the final model is significantly different from the intercept-only model (p <.001). Thus, the independent variables we identified on the within-session level, as a group, contributed significantly to prediction of task type.

Table 49. Model Fitting Information in MLR on within-session level

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	354.891			
Final	244.607	110.285	15	.000

Likelihood ratio tests (shown in Table 50) present the significance of the independent variables in the model. It tests the improvement in the model fit with each of the predictor variables added. Statistical significance (<.05) was found for all the pre-selected variables, therefore, significant relationships were found between these variables and task type.

Table 50. Likelihood Ratio Tests in MLR on within-session level

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	252.227	7.620	3	.055
Average Time To First Click	254.509	9.902	3	.019
Number of unique SERPs per query	268.314	23.707	3	.000
Mean dwell time of unique documents (seconds)	258.261	13.654	3	.003
Number of documents per query	253.331	8.724	3	.033
Average query interval (seconds)	255.606	10.999	3	.012

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

Table 51 presents the parameter estimates for the final model. MLR computed different estimates for all paired groupings of the tasks, namely BIC versus CPE, INT versus CPE, and OBI versus CPE. With respect to the comparison between BIC and CPE, it was found that the BIC and CPE tasks were different in two variables: average time to first click and mean dwell time of unique documents. Specifically, search sessions that had longer average time to first click and shorter average dwell time on unique documents were more likely to be in the BIC task than in the CPE task. For every one-second increase in average time to first click, the odds of being in BIC decreased by 27% (1.27-1.0=0.27), shown in the value of Exp (B). For every one-second increase in mean dwell time of unique documents, the odds of being in BIC decreased by 16.6% (0.834-1.0=0.166).

With respect to the comparison between INT and CPE, it was found that the INT and CPE tasks also were different in only one variable: the number of documents per query, but not other variables. Specifically, search sessions that had more documents per query were more likely to be in INT tasks, rather than in CPE tasks. For every one document per query increase, the odds of being in INT decreased by 158.1% (2.581-1.0=1.581).

With respect to the comparison between OBI and CPE, it was found that the OBI and CPE tasks were different in two variables: number of unique SERPs per query and average query interval time, but not other variables. Specifically, search sessions that had more unique SERPs visited per query but a shorter query interval were more likely to be in OBI, rather than in CPE. For every one more unique SERP visited per query, the odds of being in OBI increased by 938.16% (939.16-1.0 = 938.16); for every one second increase in average query interval, the odds of being in OBI decreased by 7.1% (0.929-1.0=-0.071).

Table 51. Parameter Estimates in MLR on the within-session level

task[a]		B	Std. Error	Sig.	Exp(B)
BIC	Intercept	-2.444	2.903	.400	
	Average Time To First Click	.239	.086	.005	1.270
	Number of unique SERPs per query	4.409	2.493	.077	82.166
	Mean dwell time of unique documents (seconds)	-.181	.059	.002	.834
	Number of documents per query	.237	.476	.618	1.268
	Average query interval (seconds)	-.013	.023	.568	.987
INT	Intercept	3.011	3.044	.323	
	Average Time To First Click	.136	.079	.085	1.146
	Number of unique SERPs per query	-2.460	2.794	.379	.085
	Mean dwell time of unique documents (seconds)	-.085	.048	.076	.918
	Number of documents per query	.948	.418	.023	2.581
	Average query interval (seconds)	-.028	.018	.116	.972
OBI	Intercept	-4.361	2.952	.140	
	Average Time To First Click	.099	.097	.310	1.104
	Number of unique SERPs per query	6.845	2.577	.008	939.160
	Mean dwell time of unique documents (seconds)	-.033	.055	.547	.967
	Number of documents per query	.847	.491	.084	2.332
	Average query interval (seconds)	-.073	.028	.009	.929

*CPE as the reference category

The MLR within session predictive models are shown as below:

Equation 7. Predictive models of task type on the within-session level

$$\text{odds for BIC} = \ln\left(\frac{Probability\ of\ BIC}{Probablity\ of\ CPE}\right)$$
$$= -2.444 + 0.239 * Average\ time\ to\ first\ click + 4.409$$
$$* Number\ of\ unique\ SERPs\ per\ query - 0.181$$
$$* Mean\ dwell\ time\ of\ unique\ documents + 0.237$$
$$* Number\ of\ documents\ per\ query - 0.013 * Average\ query\ interval$$

$$\text{odds for INT} = \ln\left(\frac{Probability\ of\ INT}{Probablity\ of\ CPE}\right)$$
$$= 3.011 + 0.136 * Average\ time\ to\ first\ click - 2.46$$
$$* Number\ of\ unique\ SERPs\ per\ query - 0.085$$
$$* Mean\ dwell\ time\ of\ unique\ documents + 0.948$$
$$* Number\ of\ documents\ per\ query - 0.028 * Average\ query\ interval$$

$$\text{odds for OBI} = \ln\left(\frac{Probability\ of\ OBI}{Probablity\ of\ CPE}\right)$$
$$= -4.361 + 0.099 * Average\ time\ to\ first\ click + 6.845$$
$$* Number\ of\ unique\ SERPs\ per\ query - 0.033$$
$$* Mean\ dwell\ time\ of\ unique\ documents + 0.847$$
$$* Number\ of\ documents\ per\ query - 0.073 * Average\ query\ interval$$

Table 52 shows the classification table by MLR on the within-session level. Overall, the final model accurately predicted 62.5% of the cases. We can compare this accuracy with the rate of accuracy achievable by chance alone, which can be calculated as: $0.25^2 + 0.25^2 + 0.25^2 + 0.25^2 = 25\%$). Therefore, the final model based on the whole-session variables had an improvement of 150% over the proportional by chance accuracy.

Table 52. Classification table in MLR on within-session level

Observed	Predicted				
	BIC	CPE	INT	OBI	Percent Correct
BIC	19	2	6	5	59.4%
CPE	4	21	5	2	65.6%
INT	4	5	19	4	59.4%
OBI	4	2	5	21	65.6%
Overall Percentage	24.2%	23.4%	27.3%	25.0%	62.5%

In summary, we found a statistically significant overall relationship between the behavioral measures on the within-session level and the task type, especially the following three variables: Average time to first click, mean dwell time of unique documents, mean dwell time on all SERPs, number of content pages per query, number of unique SERPs per query and average query interval. Compared with CPE, the BIC task was more likely to have longer average time to first click but shorter mean dwell time on unique documents; the INT task was more likely to have more documents visited per query; the OBI task was more likely to have more unique SERPs visited per query, but shorter average query interval time.

6.3.3 Prediction using both within-session and whole-session level measures

We then considered whether adding a few whole-session level measures could improve the prediction by the within-session level measures. The parameter estimates for the model (shown in Table 53) show the variables selected in the final model: Mean dwell time of unique documents, Numbers of all documents, Task completion time, Number of unique SERPs, and Number of unique SERPs per query. In general, this model identified similar behavioral differences among different task types.

Table 53. Parameter Estimates in MLR on both whole-session and within-session levels

task[a]		B	Std. Error	Sig.	Exp(B)
BIC	Intercept	-.769	2.094	.714	
	Mean dwell time of unique documents	-.255	.070	.000	.775
	Numbers of all documents	-.021	.043	.626	.979
	Task completion time	.005	.002	.047	1.005
	Number of unique SERPs	.132	.090	.142	1.141
	Number of SERPs per query	.662	.524	.207	1.938
INT	Intercept	2.403	1.709	.160	
	Mean dwell time of unique documents	-.073	.039	.063	.929
	Numbers of all documents	.080	.039	.039	1.084
	Task completion time	-.002	.002	.402	.998
	Number of unique SERPs	.021	.091	.815	1.022
	Number of SERPs per query	-.661	.523	.206	.516
OBI	Intercept	-1.910	1.918	.319	
	Mean dwell time of unique documents	-.061	.048	.209	.941
	Numbers of all documents	.051	.042	.219	1.052
	Task completion time	-.004	.002	.096	.996
	Number of unique SERPs	.277	.092	.003	1.319
	Number of SERPs per query	.719	.506	.155	2.053

This exploration also revealed that the prediction performance by combining features on the whole-session level with the within-session level could be as good as 67.2% on overall accuracy (shown in Table 54). Compared with the overall accuracy of the model on the within-session level (62.5%), and the overall accuracy of the model on the whole-session level (63.3%), we can conclude that combining features on both the within-session level and whole-session level could improve the prediction accuracy over that of the model on either single level.

Table 54. Classification table in MLR on both within-session and whole-session levels

Observed	Predicted				
	BIC	CPE	INT	OBI	Percent Correct
BIC	23	2	3	4	71.9%
CPE	1	23	5	3	71.9%
INT	3	7	19	3	59.4%
OBI	4	2	5	21	65.6%
Overall Percentage	24.2%	26.6%	25.0%	24.2%	67.2%

6.4 *Discussion for RQ2*

The second research question of this dissertation was concerned with which behaviors could be used to predict task type. The section begins with a discussion of the task and task facet effects on behavioral measures. This is followed by the discussion of the predictive models of task types on the basis of behavioral measures. Finally, the application of the predictive models of task type is discussed.

6.4.1 Task and task facet effect on behavioral measures

The user experiment was designed to examine task type effects on behaviors using realistic tasks in an unconstrained setting. Our results demonstrate that the tasks and several task facets were distinguishable by various measurements on the whole-session level and the within-session level.

Table 55. Comparison of predictive models on the whole-session level

	Task	Task Facets			
		Product	Objective complexity	Level	Goal (quality)
Task completion time	X	X	X	X	
Numbers of all documents	X	X		X	
Numbers of unique documents	X	X	X	X	
Number of SERPs	X		X	X	X
Number of unique SERPs	X		X	X	X
Number of queries	X	X	X	X	X
Total time spent on documents	X	X			
Total time spent on SERPs	X	X	X	X	
Ratio of document time to all	X		X	X	X
Ratio of SERP time to all	X		X	X	X

*Note: the character "X" indicates significant associations of behaviors with the facet values.

First, the behavioral measures on the whole session level mainly described the amount of effort the user spent on different Web page types for the whole search session. The comparison of behavioral measures on the whole session level (shown in

Table 55) reveals that the effort that a user spent on the whole task session, and the distribution of the effort on queries, content pages and SERPs could indicate the type of task the user is searching for. Among the task facets that we varied in this experiment, the level of objective complexity affected most of the behavioral measures on the whole-session level. In particular, users visited more documents, more SERPs, issued more queries and spent longer time on tasks in high-complexity tasks than in low-complexity tasks. This is what we expected from controlling the task complexity level. In addition, when the complexity level increased, the percentage of time on documents decreased, while the percentage of time on SERPs increased.

Even though it was shown that most of the whole-session level measures were significantly different between tasks with Segment level and Document level, this may not be due to the effect of the Level of document judgment. In the experiment, only CPE was characterized by Segment level, and CPE was also a Factual task with low complexity. Therefore, the effect of Level on the whole-session level behavioral measures may due to the effect of task complexity. Further studies are needed to confirm this.

It is shown that **the facet Product** influenced the task completion time, number of documents visited and total time spent on documents. In particular, users visited more documents and spent more time on documents in Mixed-product tasks than in Factual tasks. This is reasonable because in tasks involving an Intellectual product, users needed not only to find the information object, but also to understand and interpret the information in order to create new ideas or thoughts on the issue.

The facet Goal (quality) influenced the number of SERPs, the number of unique SERPs, the number of queries, the ratio of time on documents and the ratio of time on SERPs. In particular, users visited more SERPs, more unique SERPs, issued more queries, spent a lower percentage of time on documents, but a greater percentage of time on SERPs in tasks with Amorphous goals than in tasks with Specific and Mixed of Specific and Amorphous goals. It is reasonable that when the task goal is amorphous, users tend to spend more efforts on SERPs exploring what information is available for the tasks and then decide what information objects to be collected during the search.

Table 56. Comparison of predictive models on the within-session level

	Kruskal Wallis-Test	Task Facets			
		Product	Objective complexity	Level	Goal (quality)
Mean dwell time of all documents	X	X	X	X	X
Mean dwell time of unique documents	X	X	X	X	X
Mean dwell time of all SERPs	X	X			X
Mean dwell time of unique SERPs	X				X
Number of documents per query	X	X	X		X
Number of unique documents per query	X	X	X		X
Number of SERPs per query			X		X
Number of unique SERPs per query	X		X		X
Average query interval	X		X	X	X
Average Time To First Click					X

*Note: the character "X" indicates significant associations of behaviors with the facet values.

We then compared the behavioral measures at the within session level, which described how users interacted with all kinds of Web pages and the search system during search sessions. The comparison of behavioral measures on the within session level (shown in Table 56) revealed that the facet Goal (quality) significantly influenced all the within-session variables. In particular, when searching for tasks with Amorphous goal(s), users had significantly shorter dwell times on documents and SERPs, visited fewer documents but more SERPs per query, and reformulated queries more frequently than when searching for tasks with relatively Specific goal(s). This is reasonable because when users had Specific search goal(s), they had concrete ideas of what expected information objects should be like, and they read the search results and documents more carefully; when they had Amorphous search goal(s), they tended to explore the search results to see what kinds of information were available for the task, and then decide what to follow-up.

Level of document judgment had a significant influence on three within-session level measures: mean dwell time on documents, mean dwell time on unique documents and average query interval. Particularly, users spent significantly longer dwell time on documents and significantly longer average query interval time in tasks on the Segment level than in tasks on the Document level. This is because users needed to read the documents more carefully to find the specific piece of information for tasks on the Segment level, and they only needed to read the document roughly to decide make usefulness judgment while searching for tasks on the Document level since we did not require them to use the found information to accomplish the work task. In addition, neither the dwell time on the SERPs nor the number of pages visited per query was significantly influenced by this facet.

The level of objective complexity also affected some of the behavioral measures on the within-session level. In particular, when the tasks became more complex, users were more likely to spend shorter dwell time on documents; this may because when they had more information objects to collect, they had relatively shorter time for each document they clicked. In addition, when the complexity level increased, users visited fewer documents but more SERPs per each query, and tended to reformulate queries more frequently. This also echoed the need of collecting more information objects, and examining more SERPs, which would allow them to find more information objects.

Even though we found in Mixed-product tasks that users spent significantly more total effort on documents on the whole-session level, they spent shorter average dwell time on each document they clicked than in Factual tasks. Therefore, they visited more documents per query in Mixed-product tasks than in Factual tasks. A possible reason for this is that in Factual tasks, users were searching to find the factual information, so they read the information more carefully; while in tasks with Intellectual work, users' goal(s) were to create new ideas from the information they found, rather the factual information itself, so they would search to read more information to generate their own ideas.

In sum, the results of the analysis of the relationship between behavioral measures on the whole session level and within session level demonstrate significant differences among different tasks. The differences on the whole session level described differences in the amount of effort users spent on the whole task and the distribution of search effort on queries, content pages and SERPs; while the differences on the within session level revealed the differences on the search process. The examination of these variables demonstrates that many of them could serve as important indicators of task type.

6.4.2 Predicting task type from behavioral measures

We used multinomial logistic regression method to generate the predictive models of task type. The results of this study demonstrate that some of the behavioral measures, on both the whole-session level and the within-session level, could be indicators of task type.

Table 57. The important predictors of task type on the whole-session level

	BIC	INT	OBI
Number of unique SERPs			+
Task completion time		-	-
Numbers of all documents			
Numbers of unique documents	+		
Ratio of document time to all	-		

* CPE as the reference category
+ indicates the measure is positively related to the predicted task type
- indicates the measure is negatively related to the predicted task type

First, our study found the following measures as a combination of important predictors on the whole-session level could be used to predict task type: Task completion time, Number of all documents, Number of unique documents, Number of unique SERPs, and Ratio of document time to all. Even though each of the whole-session level measures described the amount of effort during search,

Table 57 demonstrates the effort distribution on different types of Web pages and searches in different types of tasks; therefore, a combination of whole-session level measures was able to predict task type. In particular, we found that compared with the CPE task (Factual task on the Segment level, with Specific goal and low complexity), users visited more unique documents and more unique SERPs, but spent relatively lower percentage of time on documents in the BIC task (Mixed-product on the Document level, with Specific goal and High complexity); users had lower task completion time, but visited more unique documents in the INT task (Mixed product on the Document level, with Mixed goal(s) and Low complexity); and, users had lower task completion time but visited more unique SERPs in the OBI task (Factual task on the Document level, with Amorphous goal and High complexity). These features were possibly due to the complexity, product and goal (quality) facets.

The behavioral measures on the within session level were able to describe users' search processes and are able to be acquired before users finished the search sessions. Since the goal of personalization is to provide personalized search results and assistance to better users' search experience, it is important for the search system to monitor users' search experience during search. The selected important predictors of task type on the within-session level are summarized in Table 58.

Table 58. The important predictors of task type on the within-session level

	BIC	INT	OBI
Mean dwell time of unique documents	-		
Number of unique SERPs per query			+
Average query interval			-
Average Time To First Click	+		
Number of documents per query		+	

* CPE as the reference category
+ indicates the measure is positively related to the predicted task type
- indicates the measure is negatively related to the predicted task type

In particular, we found that compared with the CPE task (Factual task on the Segment level, with Specific goal and low complexity), users spent shorter average dwell time on unique documents, but longer average time to first click after issuing a query in BIC task (Mixed-product on the Document level, with Specific goal and High complexity); users visited more documents per query in INT tasks (Mixed product on the Document level, with Mixed goal(s) and Low complexity); and visited more unique SERPs per query but had lower average query interval times in the OBI task (Factual task on the Document level, with Amorphous goal and High complexity). The differences between BIC and CPE mainly came from the Product and Level facets, indicating that users did not need to read the documents as carefully in BIC as they did in CPE, but it took them longer to first click on documents since they needed to create new ideas from the search results in BIC (Mixed product). The differences between INT/OBI and CPE mainly came from the Goal (quality) facet. In particular, when they were searching for Factual information with amorphous goals, they visited more unique SERPs per query, as in OBI; and when they were searching for Mixed product with amorphous goals, they visited more documents per query, perhaps to create new ideas from the content of documents, as in INT.

Our results demonstrate that the multinomial logistic regression modeling is a helpful method in understanding how a combination of behavioral measures is affected by task features. Furthermore, it is possible to generate predictive models of task using a combination of a small set of behavioral measures. This is reasonable because task type influences a number of behaviors during search, and any single behavioral measures is not able to describe all of these effects. Therefore, identifying a combination of behavioral measures should be helpful in predicting task type, as the multinomial logistic regression methods showed in our study.

6.4.3 Application of the predictive models of task type

The second research question of this dissertation was concerned with understanding which behavioral measures could be used as predictors of task type, and understanding how to automatically apply the task-specific predictive models for implicit relevance feedback. The findings from this research address these questions and have implications for personalized information retrieval experience in different types of tasks.

In our study, we investigated behavioral measures on both the whole session level and the within session level, and generated predictive models on both the levels. A primary concern for a proposal to use behavioral measures to personalize search is ensuring the predictions can be made before or during searching, not after the task is finished, because if the user has finished the task or left the search system, there is no need for the system to provide any personalized assistance. Therefore, between the behavioral measures on the whole session level and the within session level, the predictive models on the within session level are applicable for the system to predict task type during searching, since the predictive model on the whole session level can be applicable only after the task is completed in practice.

However, that does not indicate the model on whole session level is not useful at all. For one thing, the prediction accuracy was higher by the model on the whole-session level (63.3%) than the model on the within-session level (62.5%). For another, the search system can calculate the behavioral measures during search by assuming that the task is just finished, and then apply the predictive model on the whole session level to make predictions of the task type at different points of the searching episode. Therefore, we also found that the predictive model based on both whole-session and within-session levels (67.2%) may achieve much more accurate prediction than the model on any single level. Even though the predictions based on the "assumed" whole-session level behavioral measures may not be very accurate at the beginning, we could expect that such predictions will become increasingly accurate as time goes on during the search episode. Therefore, the search system can monitor users' search behaviors, calculate all the behavioral measures that are available at the point they are calculated, on both the whole-session level and the within-session level, and then make predictions of task type through voting by multiple predictive models. It is also necessary for the search system to adjust the prediction results based on the up-to-date behavioral measures and by comparing the results from different models and at different stages of the search. In addition, it is also possible to apply the predictive model of task type across search sessions in long-term personalization of information retrieval.

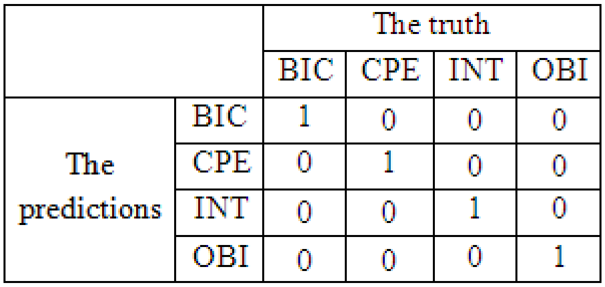
Studies have shown that re-finding is common in online search (Tyler and Teevan, 2010), so users may continue working on the same task or need to re-find information for the task again sometime later. Therefore, when implementing the predictive models of task type, we can also apply the whole-session level model and the within-session level model after the current search session is completed, save the prediction results, and apply the predicted results if the user returns to the same task. Thus the predictive models of task types can be implemented both during the search sessions and across the search sessions.

Another question is when to apply the predictive model of task type during the search session. Many of the behavioral measures can only be calculated after the user has been searching for a while, not right at the beginning of the search, so the predictive models can only be applied after some search behaviors have occurred in the search episode. For example, the mean dwell time of content pages can be calculated when the user has visited at least one content page; the mean dwell time of SERPs can be calculated when the user has visited at least one SERP; the average query interval can be calculated when the user has issued at least two queries in the session, and so on. Therefore, it is recommended that the predictive model first be calculated only after the second query is issued. Such timing may seem a bit insensitive at the beginning of search, however, if a search session is completed before the second query is issued, then the user may not need any personalization of the search results from the system. Thus, starting to apply the predictive model of task type after the second query is not too late for users.

Another issue to consider when applying the predictive model of task type is the cost and benefit of each prediction of task type, which can be described using a cost-benefit matrix (Lewis, 1995). By default the cost-benefit matrix has a value of one (1.0) for correct predictions and zero

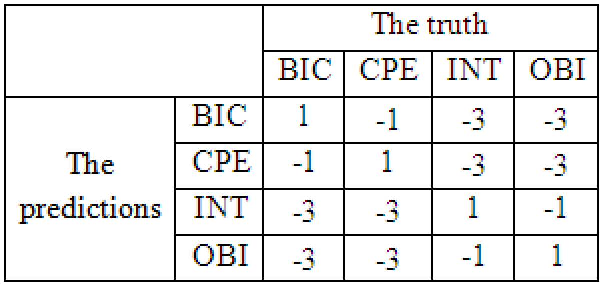
(0.0) for incorrect predictions (see Table 59). In most real-world situations, however, an incorrect prediction has negative benefit or has a cost (less than zero), and a correct prediction has a positive benefit (see Table 60). The purpose of the Cost-Benefit Matrix is to weigh the possible outcome of each prediction based on the performance of the actions for the prediction. If there is a significant difference in cost and benefit between different predictions then predictive model selection should not be based on the raw accuracy but on the optimizing benefit. If the predictive model is a logistic regression model that generates the probabilities of being each type, the benefit can be calculated by multiplying the predictive probabilities by the actual "Cost-Benefit Matrix".

Table 59. The default Cost-Benefit Matrix



		The truth			
		BIC	CPE	INT	OBI
The predictions	BIC	1	0	0	0
	CPE	0	1	0	0
	INT	0	0	1	0
	OBI	0	0	0	1

Table 60. The "actual" Cost-Benefit Matrix



		The truth			
		BIC	CPE	INT	OBI
The predictions	BIC	1	-1	-3	-3
	CPE	-1	1	-3	-3
	INT	-3	-3	1	-1
	OBI	-3	-3	-1	1

Recall that in our study, the reason to generate the predictive models for task type is because we found the specific predictive models of document usefulness for different types of tasks were able to achieve more accurate prediction than the general predictive model when no task information is provided. Therefore, correctly predicting a BIC task as a BIC type of task and then applying a BIC-specific predictive model of document usefulness has positive benefit for personalization; but if it predicts a BIC task as another type of task, and then applies the other specific model of document usefulness, this is likely to be harmful for personalization. In addition, it might be the case that mistakenly predicting a BIC task as an INT type of task is even more harmful than predicting a BIC task as a CPE type of task. If we know how much benefit and cost we could get by each prediction, when the cost-benefit matrix has new non-default values assigned (as shown in Table 60), the predictive model can optimize the net benefit (profit) associated with each prediction. The net benefit can be calculated by multiplying the predicted probability and the Cost-Benefit Matrix. For example, if the predicted probability of being each task type for a case is:

Probability of being BIC: 0.3; CPE: 0.6; INT: 0.05; OBI: 0.05.

then we can calculate the possible benefit of predicting the case as each task type as below:

Benefit of predicting as BIC: (1*0.3) + (-1*0.6) + (-3*0.05) + (-3*0.05) = -0.6

Benefit of predicting as CPE: (-1*0.3) + (1*0.6) + (-3*0.05) + (-3*0.05) = 0

Benefit of predicting as INT: (-3*0.3) + (-3*0.6) + (1*0.05) + (-1*0.05) = -2.7

Benefit of predicting as OBI: (-3*0.3) + (-3*0.6) + (-1*0.05) + (1*0.05) = -2.7

Therefore, the maximum benefit we can get is by predicting this case as CPE. As we can see from this example, the cost-benefit matrix input is essential to optimize personalization performance. However, the cost-benefit matrix cannot be obtained until testing the specific predictive models of document usefulness on different tasks is performed, and this will be left for future work.

Figure 9 shows the process of the application of the predictive models of documents and predictive models of task type generated from this study. Before the task type information is predicted, the system can simply apply the general predictive model for document usefulness, and then apply a specific model when the task type information is predicted. As shown in the figure, the prediction of task type can be adjusted through consistently observing users' search behaviors and the application of the specific model might also be adjusted as the task type prediction alters. Therefore, such a personalization search system is an interactive information retrieval system, which could adjust the prediction of task type and document usefulness on the basis of users' interactions during search process, then to provide the personalized search results to each user.

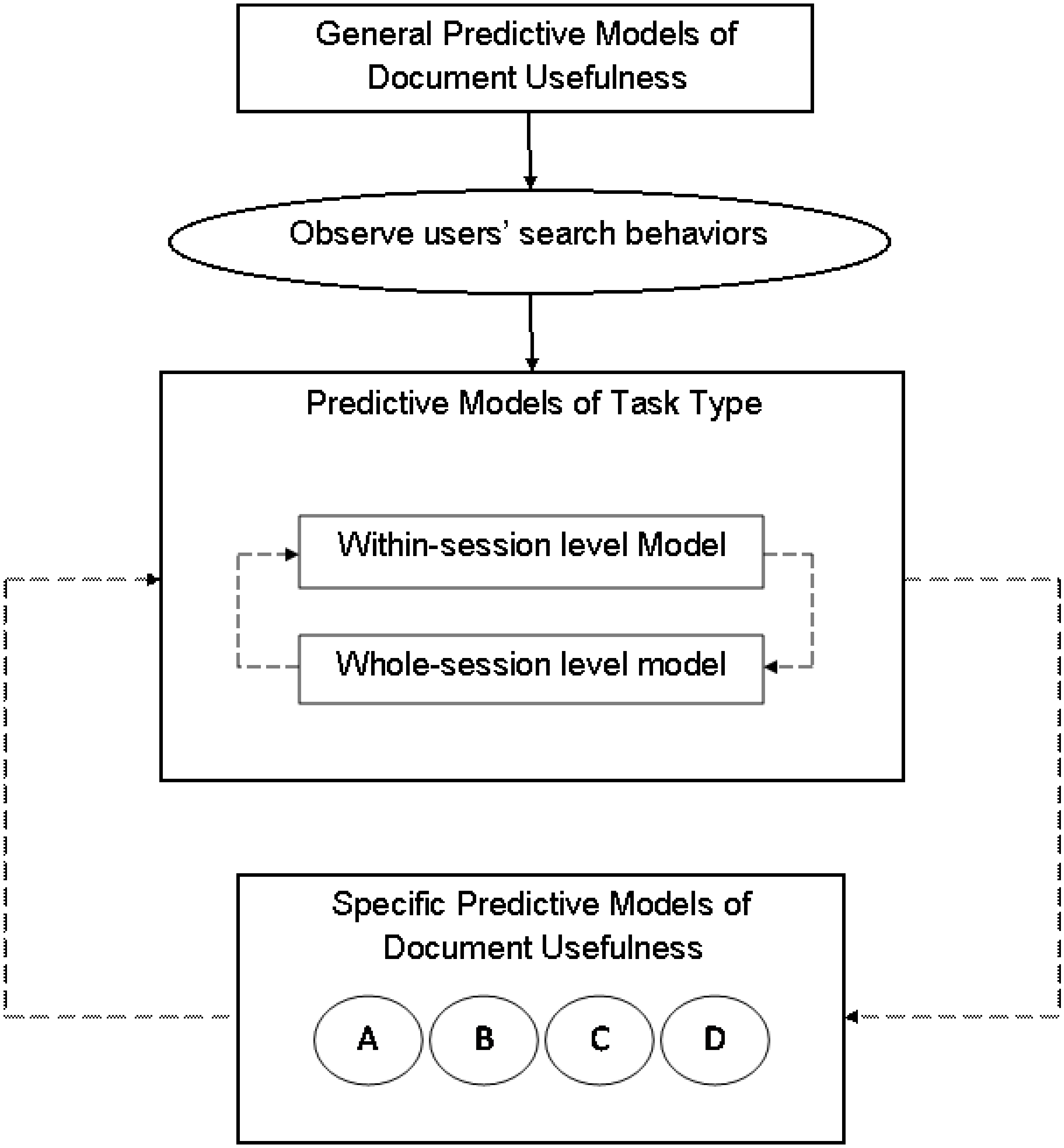

General Predictive Models of Document Usefulness
Observe users' search behaviors
Predictive Models of Task Type
Within-session level Model
Whole-session level model
Specific Predictive Models of Document Usefulness
A
B
C
D

Figure 9. Application graph of the personalization models

Chapter 7. Conclusion

7.1 *Summary of this dissertation research*

The overall goal of this dissertation was to understand how to personalize information retrieval systems to tailor search results to individual users through the inference of useful documents during their search episodes in different types of tasks. Two research questions were developed to address the general research goal, particularly generating predictive models of document usefulness for different types of tasks, as well as generating predictive models of task type on the basis of users' interaction behaviors.

The goals of this dissertation research were accomplished by analyzing results from a controlled user experiment with 32 participants. Participants were asked to search for four different types of tasks that varied by task facets, and all of their interactions with the computer were logged on the client side, using multiple loggers. During the search, participants were asked to save content pages that were useful for helping them to accomplish the assigned search tasks and these saving behaviors were considered as explicit judgments of document usefulness. In this study, we generated predictive models of document usefulness and task type on the basis of users' search behaviors.

The research generated several important findings. First, it was found that behavioral measures on both the content pages and the search result pages could be indicators of document usefulness. Behavioral measures on content pages help understand how users interact on the clicked documents, and they are indicative of document usefulness. For example, dwell time on a page was found to be among the most important predictors, as well as the visit_id of a page, the number of mouse clicks and the number of keyboard activities on a page. On the other hand, behavioral measures during query intervals help describe what the user does between issuing one query and the next. For example, the time to first click after issuing a query, the total time spent on content pages, the number of content pages visited, the average dwell time on SERPs, the total time on SERPs, and the proportion of time spent on content pages were all found to be important indicators of document usefulness in our predictive models. Previous studies have either focused on the behaviors on content pages or behavioral measures on SERPs. This study demonstrates that users' behaviors in a search episode are not isolated, and that a combination of behavioral measures on both content pages and SERPs during the search process should be considered in generating predictive models.

Secondly, our results demonstrate a significant effect of task type on predictive models of document usefulness. Task type not only affected the selection of behavioral measures as predictors of document usefulness, it also affected the thresholds or the weights for each of the behavioral measures. In addition, when task information is available, the task-specific model gives better prediction of document usefulness than the general model that is not task specific. Therefore, we can conclude that it is important for personalized IR systems to detect the context in which a search is conducted, especially the task type, and then to apply the relevant specific personalization model to individual users.

Thirdly, task type and task facets influenced how users interacted with search systems during search. Previous studies have shown that task type could influence users' search behaviors on the whole search session level, e.g. the amount of effort to accomplish the task, the search strategies or the types of information objects that would be useful. Our results further demonstrate that different task facets influence different aspects of users' search behaviors. For example, the task facet of objective complexity mainly affects the amount of effort a user spends on searching; the facet product mainly affects the interactions on content pages; the facet level mainly affects the dwell time on individual content pages; and the facet goal (quality) mainly affects how users examine the search result pages.

Fifth, based on our findings on the effects of task type on user behaviors, we made an important step in predicting task types on the basis of behavioral measures, collected on both the within-session level and the whole-session level. Even though many behavioral measures were influenced by the task type, a combination of a small portion of them was sufficient to build predictive models of task type. In addition, the predictive models on the within-session level achieved similar prediction performance to the predictive models on the whole-session level measures, and combining behavioral measures on both levels achieved the best prediction performance.

Last, but not least, the study has explored two classification methods to generate predictive models based on users' search behavioral measures: recursive partitioning and logistic regression. Previous studies have mainly focused on examining the differences in behavioral measures by contextual factors, and very few of them tried to generate models to predict the contextual factors based on the behavioral measures. The methods we employed in this study have been shown to be effective in selecting important behavioral measures for predictive models of task, an important contextual feature.

Due to the results being based on a controlled user experiment with college students working on four given tasks, care should be taken when generalizing the findings. Nevertheless, the task types in our experiment were defined by a combination of task facets, following the task classification scheme proposed by Li and Belkin (2008). We varied the values in several facets while keeping the other facets that we can control constant, like Source of task; Task doer; Time (length) Process; Goal (quantity); Interdependence; and Urgency, which makes it possible to generalize the findings relatively safely to other tasks with the same task facet values without concerns of the topicality issue.

7.2 *Implications for Personalized IR system design*

The results of this study have both theoretical and practical implications. Most notably, the results demonstrate that behaviors are affected by the context in which the user seeks information, and, in particular, that task type affects how users interact with search systems. The results further demonstrate that users' behavioral measures are indicative of document preferences and search context, and that it is important to combine multiple behavioral measures to infer document usefulness and task type, rather than taking account of any single behavioral measure. Finally, the results indicate that task type should be taken into consideration when inferring document usefulness from behavioral measures.

The results of this study have several practical implications for the design of personalized IR systems on the basis of users' interaction behaviors. First, it is important to observe users' search interactions during the search episode, especially on the client-side, rather than only using server-side logging, because our results demonstrated that the users' interactions on the content pages were the most important predictors of document usefulness. In addition, understanding how users interacted on the page would help the system understand how users would further use the information after searching.

Another practical implication for the design of personalized IR systems is that users' behavioral measures are not isolated, but instead, are related to each other. We not only need to investigate the effect of context and other factors on individual behavioral measures, but more important, we need to investigate the effects of contextual factors on a combination of behavioral measures, and how to integrate multiple measures to optimize the prediction performance.

Finally, personalization of search results should be an interactive and adaptive process during and across search episodes. Our results show that it is important for a retrieval system to learn and predict individual users' information needs and search contexts, and then to apply context-specific personalization models to individual users. In addition, the predictive models may not be acquired or be very accurate at the beginning of a search, so the personalized system should consistently monitor users' behavioral measures and adjust the predictive models on the basis of users' current interactions to tailor the results to the users.

7.3 *Limitations and Future Studies*

There are some limitations to our study as well as need for future studies. First, as always in user studies of this type, the experiments were constrained by a small sample of participants. But since participants were given 20 minutes to search for each of the tasks, and they visited a large quantity of online documents during the experiment, the sample size for generating predictive models of document usefulness may not be a big issue in our study. But the predictive models of task type may be a bit constrained by the small sample size, and the realistic way to address this issue is to do more studies, and then compare with the predictive models we generated in this study.

Second, the task facets were not balanced in this experiment. For certain facets, like the Level of document judgment, we had only one task representing one specific facet value (the Segment level). In addition, the task topic and the available information for the topic might also influence users' search behaviors and search experience. Future studies are needed to have enough task and topic examples to represent each value of the task facets controlled in the experiment, to fully examine the effect of each facet and any possible interaction among them.

Only saved pages were considered as useful documents for RQ1 in this study. It is possible that some of the documents were also useful in understanding the task or judging other information, but they were not saved because they did not directly provide information useful to the tasks. Therefore, we suggest future studies to ask users to evaluate the usefulness of all the visited documents during search to fully understand what usefulness means to users for each task. The experiment was designed to ask participants to conduct the search task, rather than complete the work task that led to the search task; in other words, we asked participants to save useful documents for completing their assignments later, and did not ask them to actually work on the tasks. Sometimes this is what people do, but there are also cases in which users actually read and use the collected information to complete the task while searching. There may be some differences in these two cases, and more studies are needed to compare the differences and examine any differences in behavioral measures.

Other behavioral measures, which could be logged during the search process, may also be potential sources for predictive models of document usefulness or task type. For example, eye-movement from eye-tracking data, and mouse or cursor movements on pages may provide more detailed information about how users interact on each of the Web pages (Cole, Gwizdka, Liu & Belkin 2011; Huang, White, Buscher, & Wang, 2012). We can expect more accurate predictions if such behavioral measures are included in our predictive models.

The behavioral measures on the within-session level for RQ2 in this study were calculated at the end of the search episode for simplicity reasons. But these measures can be calculated at any point during search session in practical. Future studies should examine the prediction performance by the real-time behavioral measures on the within-session level, and to explore at what point could these real-time measures are able to make reasonable and relatively accurate predictions of task type.

More work needs to be done to evaluate the performance of implementing the predictive models of document usefulness and task type. In particular, future studies should investigate how to collect behavioral measures and when the system could make relatively accurate predictions of task types, and how retrieval should be modified as a result of applying predictive models of document usefulness. The TREC Session Track provides a dataset that contains users' search behavior logs in search sessions, rather than only single queries, which could be a potential dataset to evaluate the predictive models. However, it is more important to implement the predictive models in real settings and evaluate the personalization performance by real users, in particular whether their search experience is improved in a personalized system compared with a non-personalized system.

Finally, more work needs to be conducted on when to personalize search results, especially under which context or which task type could we achieve the best retrieval performance through personalization. Teevan, Dumais & Horvitz (2010) examined users' clicks and explicit judgments for the same queries and found they differed greatly from one another. They proposed *the potential for personalization*, which could be defined by the gap between how well search engines could perform if they were to tailor results to the individual, and how well they currently perform by returning results designed to satisfy everyone. It is possible that personalization could improve retrieval performance in some task types, and that it does not work well in other task types. Since some behavioral measures are indicative of task types, it is reasonable to expect that some behavioral measures are also indicative of when to personalize.

In conclusion, this research has contributed to a better understanding of how to make use of multiple information-seeking behavioral measures as implicit evidence of document usefulness and task type, as well as how the task type as a contextual factor significantly affects the relationship between behavioral measures and document usefulness. The research findings have both theoretical and practical implications for designing personalized IR system on the basis of users' interaction behaviors. Future studies are suggested on making use of the findings in this study and issues to be further addressed.

APPENDICES

A. Descriptive analysis results on behavioral measures in RQ1

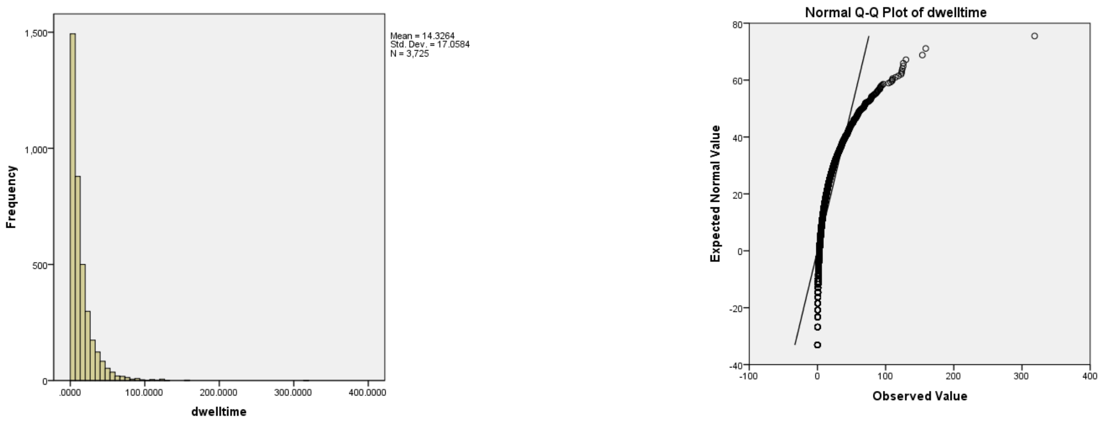


Figure 10. The distribution and the normal Q-Q Plot of dwell time

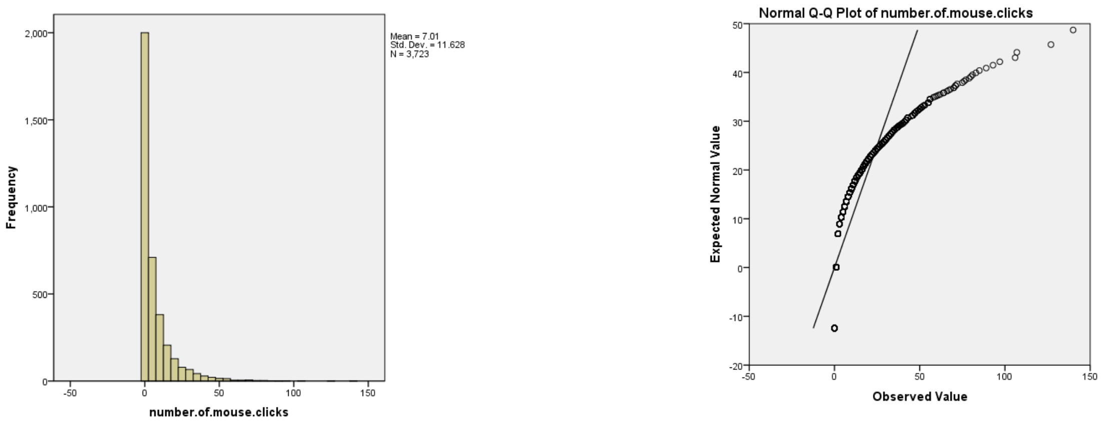


Figure 11. The distribution and the normal Q-Q Plot of number of mouse clicks

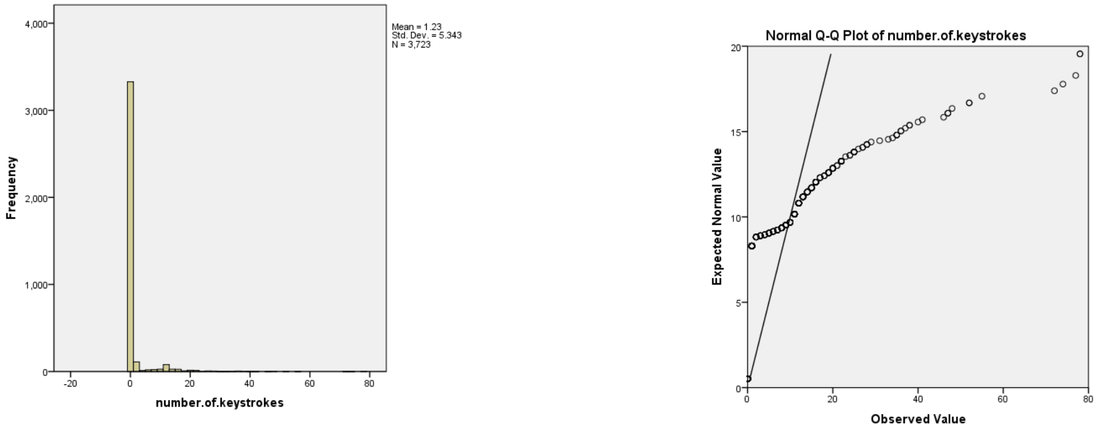


Figure 12. The distribution and the normal Q-Q Plot of number of keystrokes

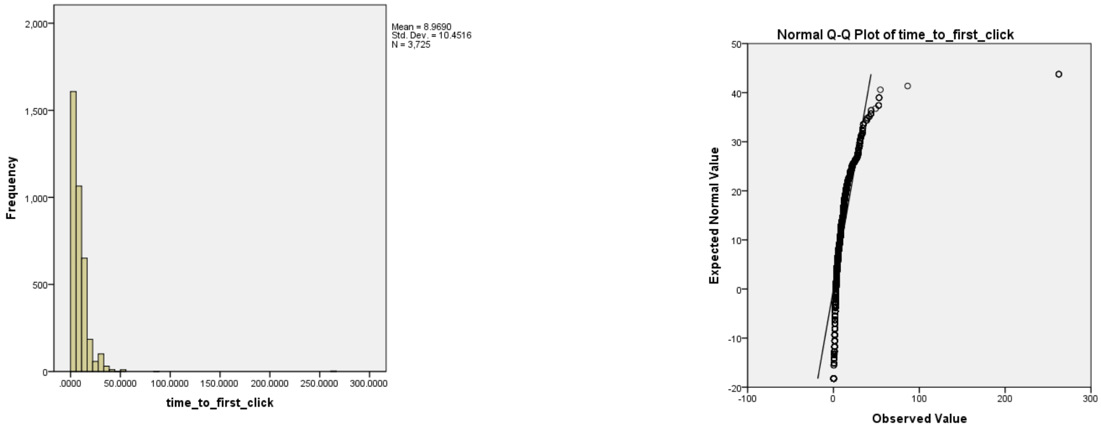


Figure 13. The distribution and the normal Q-Q Plot of time to first click

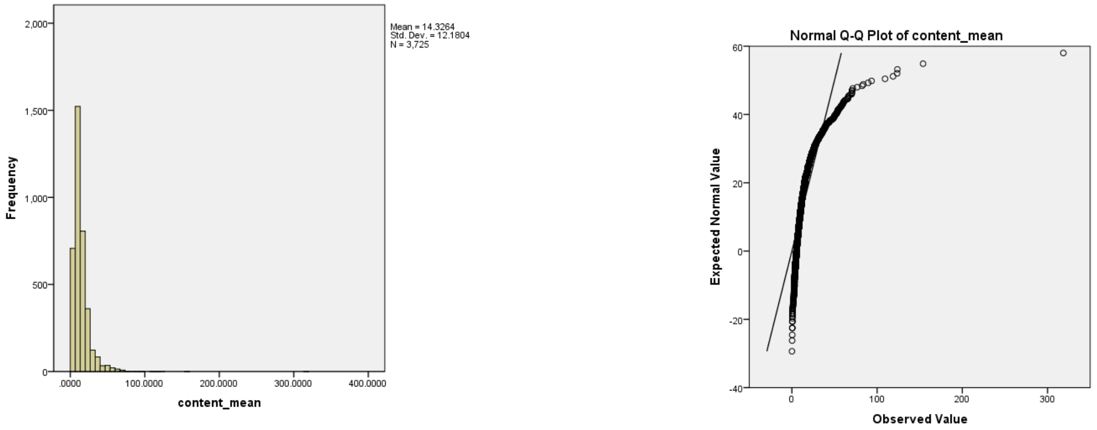


Figure 14. The distribution and the normal Q-Q Plot of content _mean

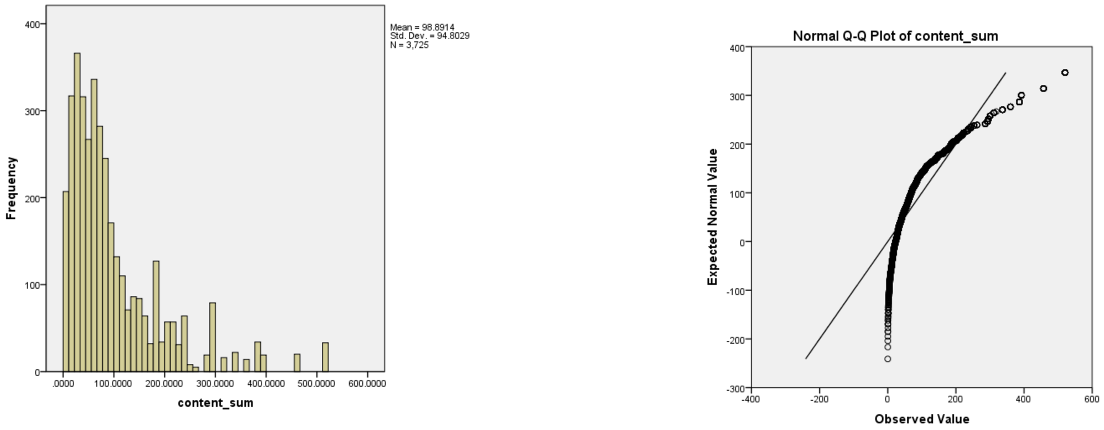


Figure 15. The distribution and the normal Q-Q Plot of content_sum

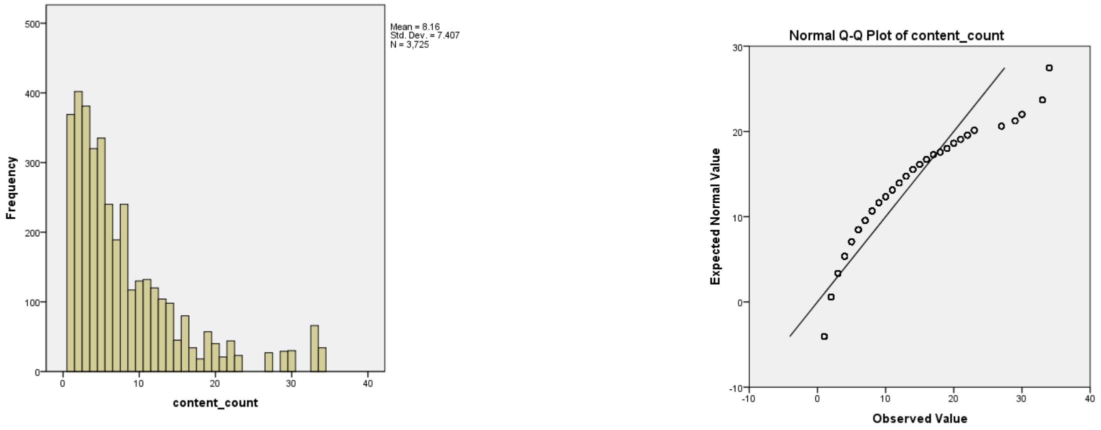


Figure 16. The distribution and the normal Q-Q Plot of content_count

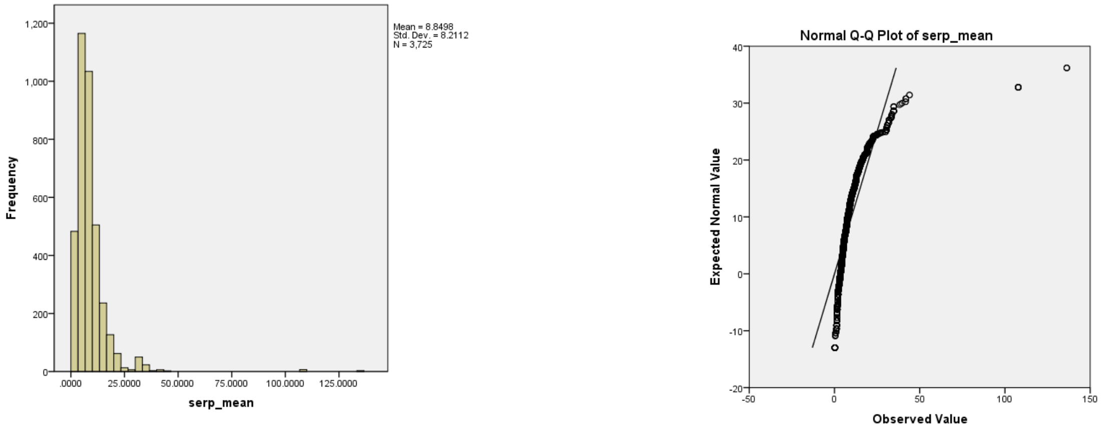


Figure 17. The distribution and the normal Q-Q Plot of serp_mean

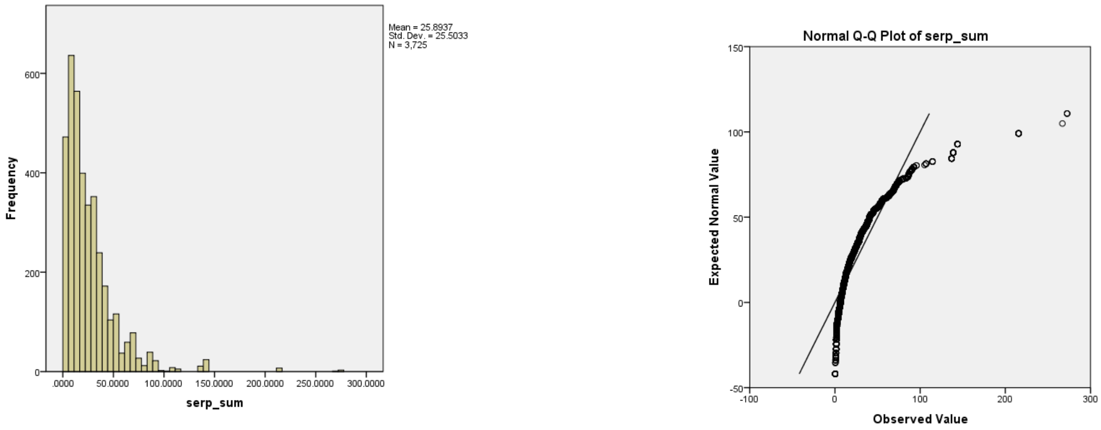


Figure 18. The distribution and the normal Q-Q Plot of serp_sum

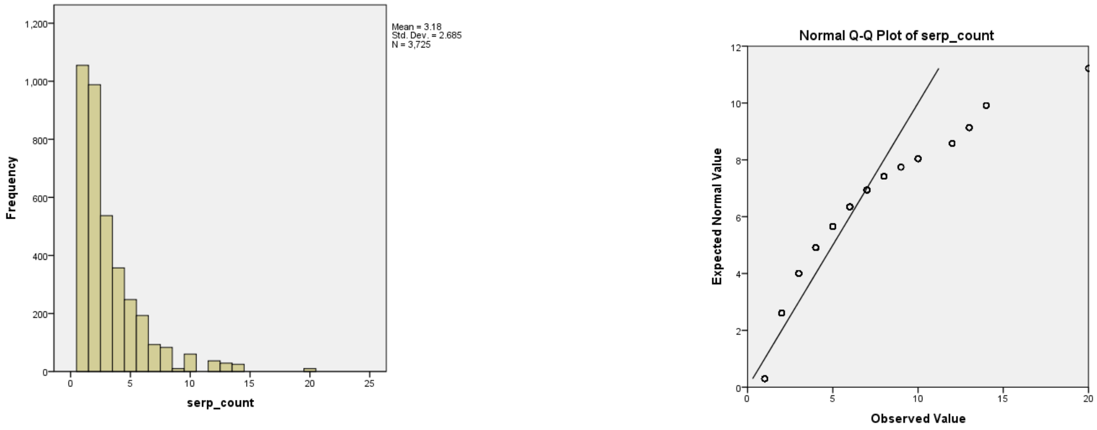


Figure 19. The distribution and the normal Q-Q Plot of serp_count

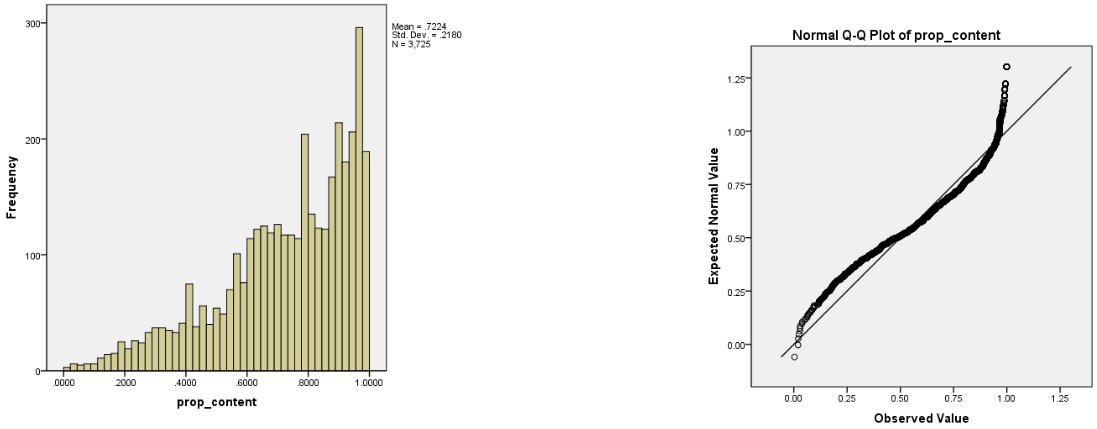


Figure 20. The distribution and the normal Q-Q Plot of prop_content

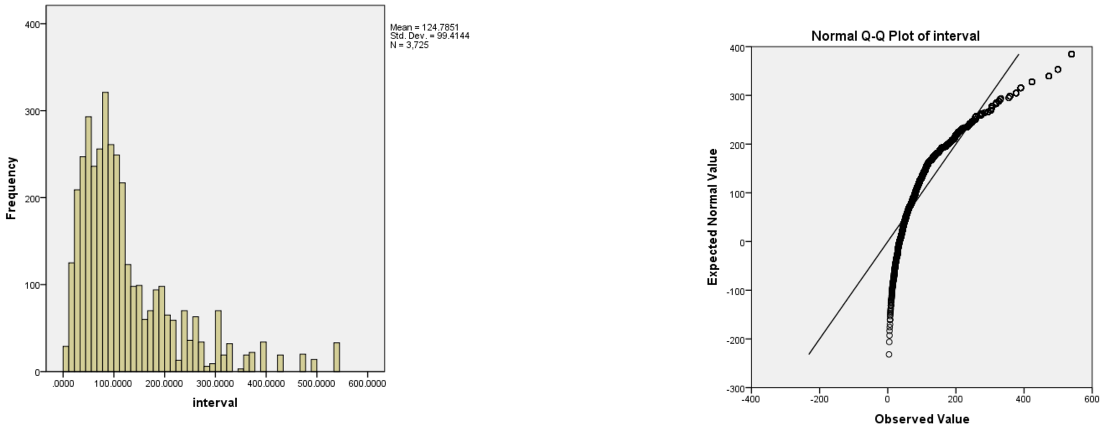
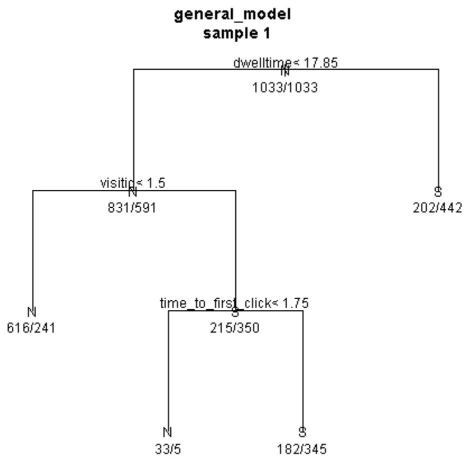
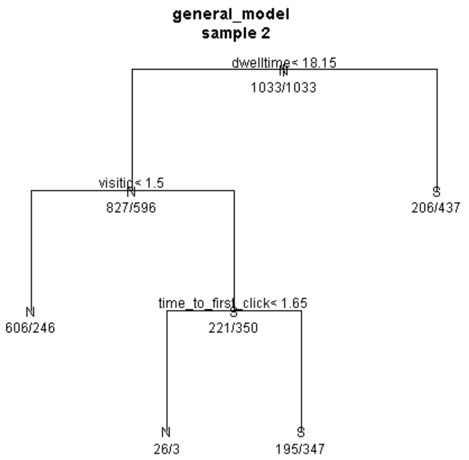


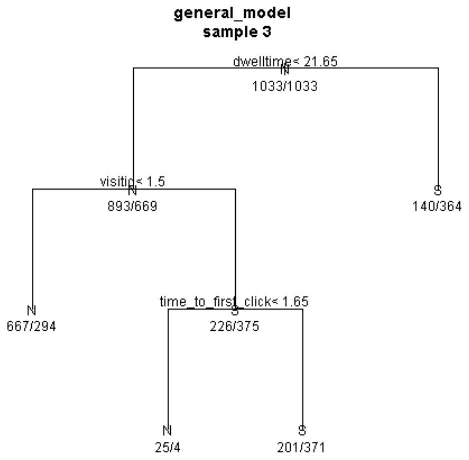
Figure 21. The distribution and the normal Q-Q Plot of query interval

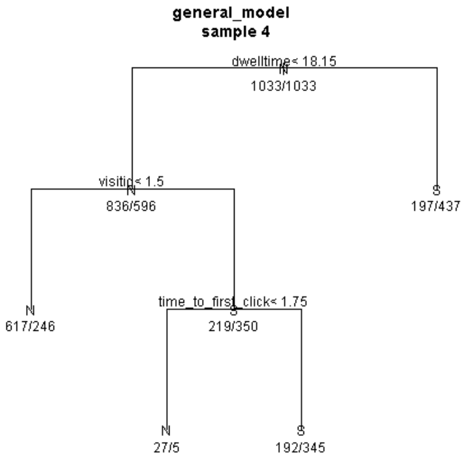
B. Predictive modeling on random samples using recursive partitioning in RQ1

a. For the general model









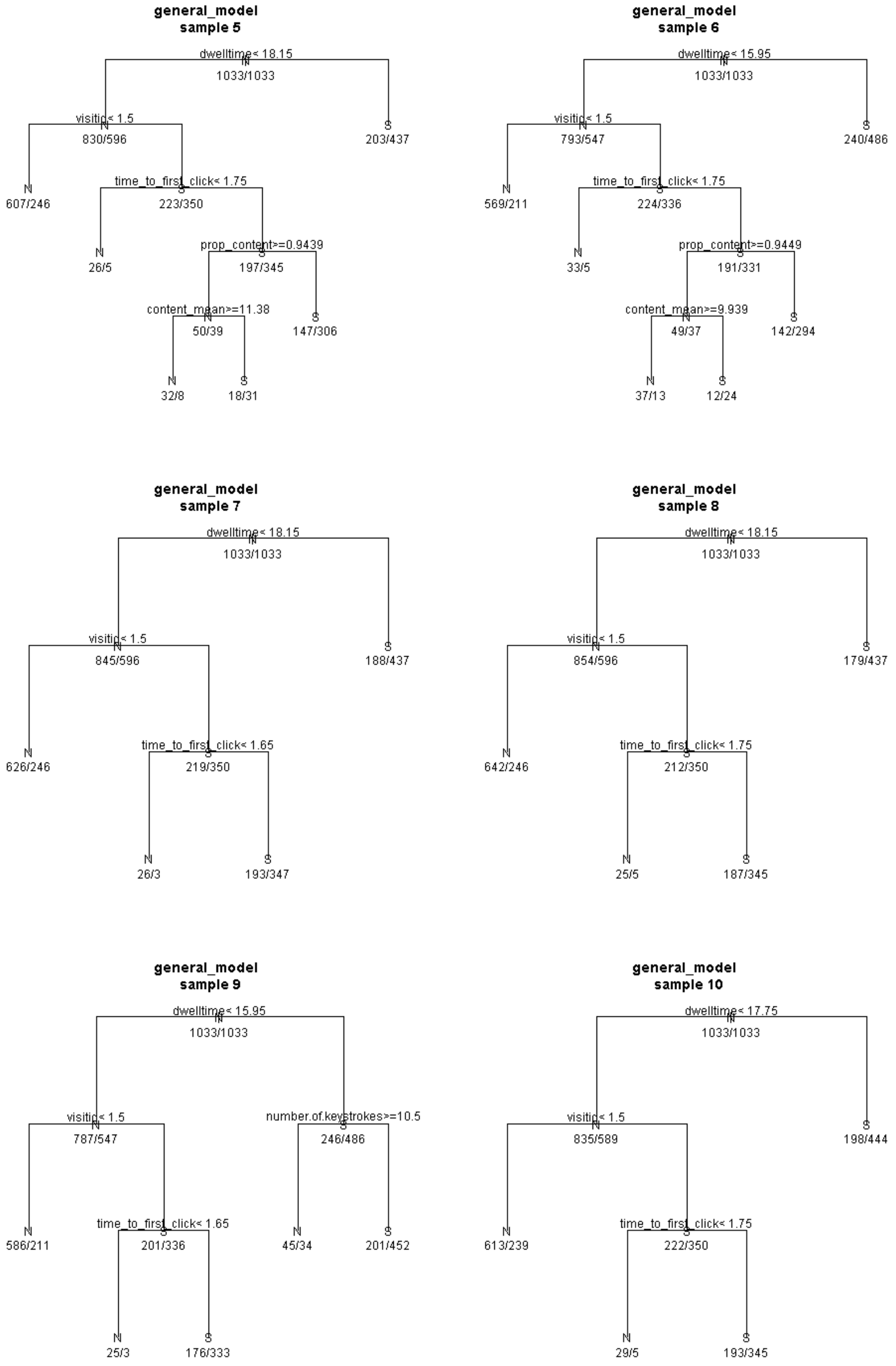
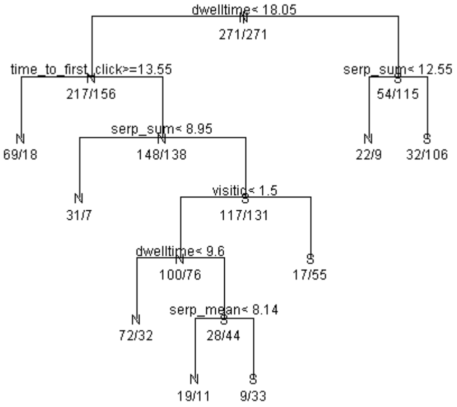


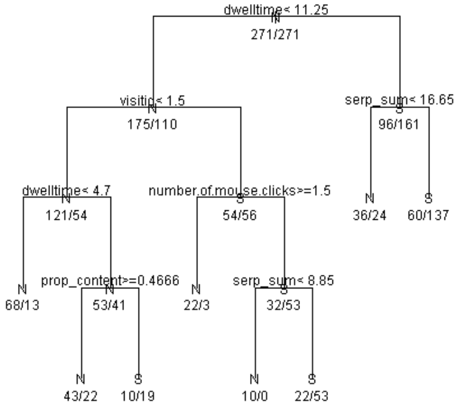
Figure 22. Predictive modeling on ten samples using recursive partitioning for the general model

b. For BIC task

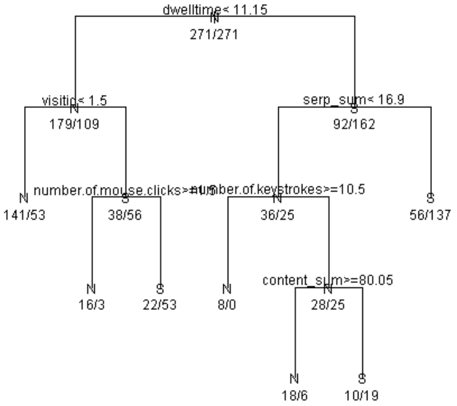
specific_model_BIC
sample 1



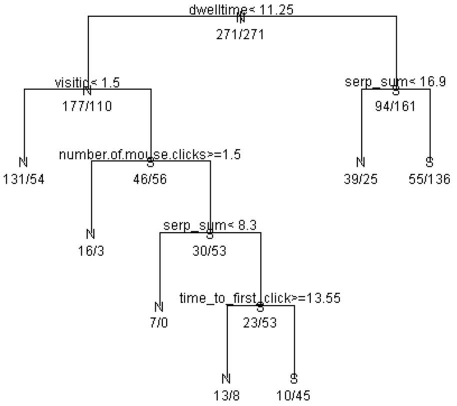
specific_model_BIC
sample 2



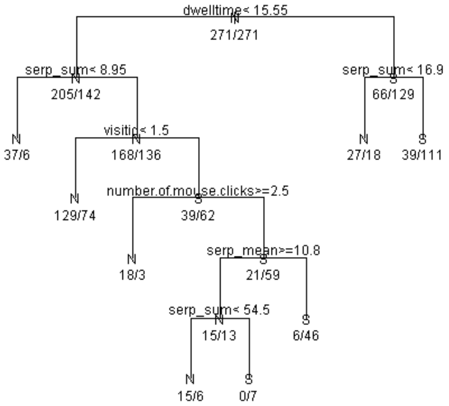
specific_model_BIC
sample 3



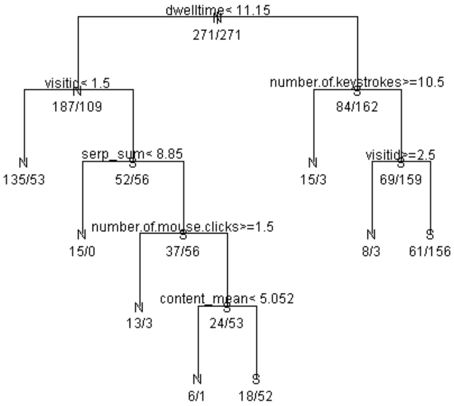
specific_model_BIC
sample 4



specific_model_BIC
sample 5



specific_model_BIC
sample 6



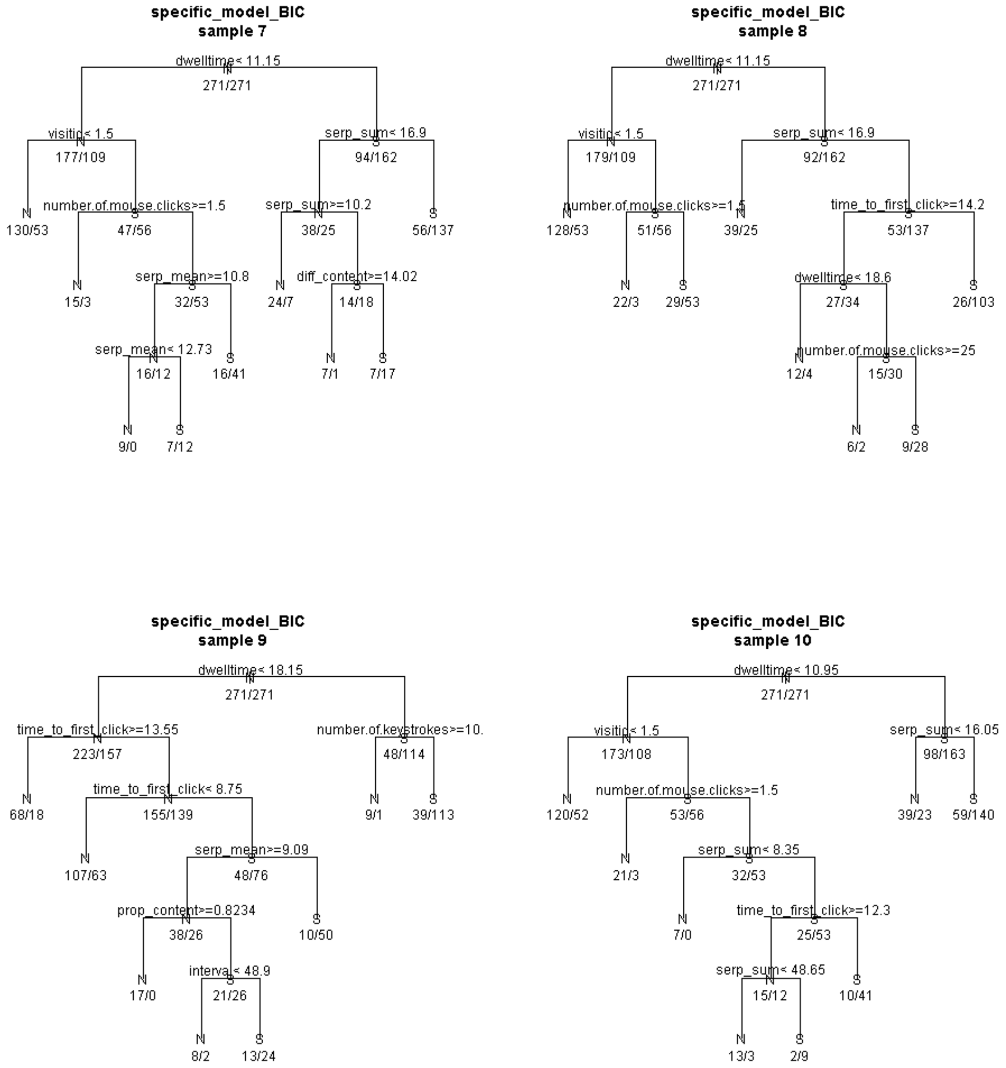
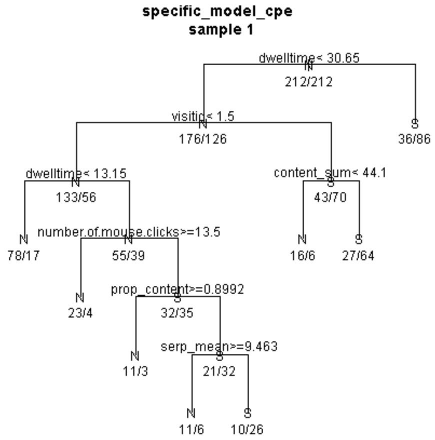
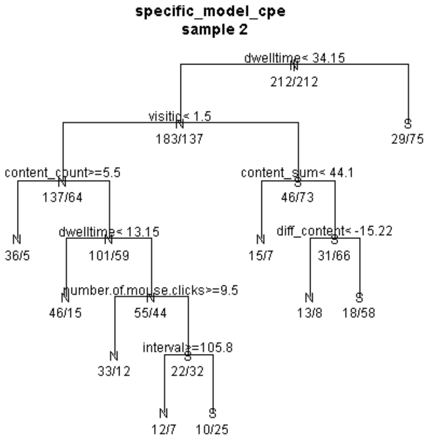
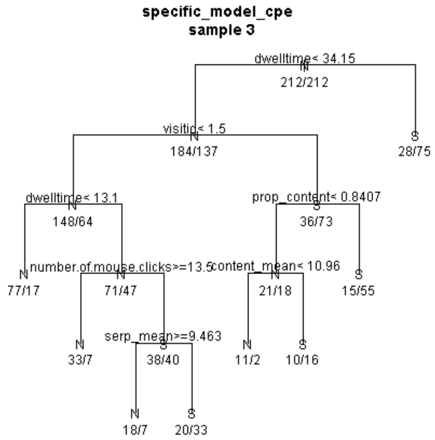


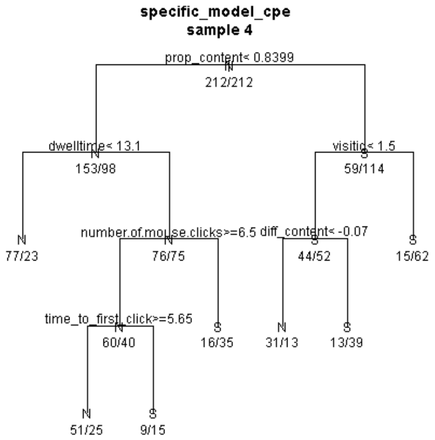
Figure 23. Predictive modeling of BIC on ten samples using recursive partitioning

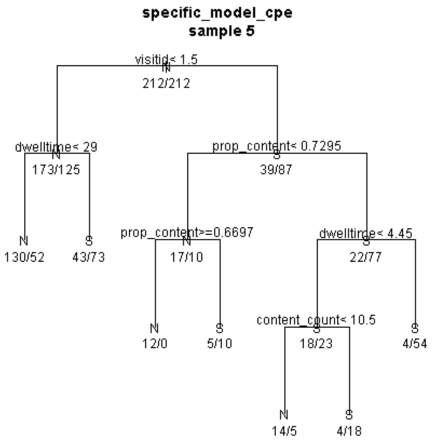
c. For CPE task

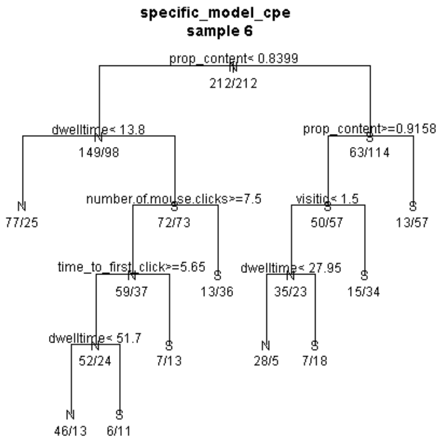

specific_model_cpe
sample 1


specific_model_cpe
sample 2


specific_model_cpe
sample 3


specific_model_cpe
sample 4


specific_model_cpe
sample 5


specific_model_cpe
sample 6

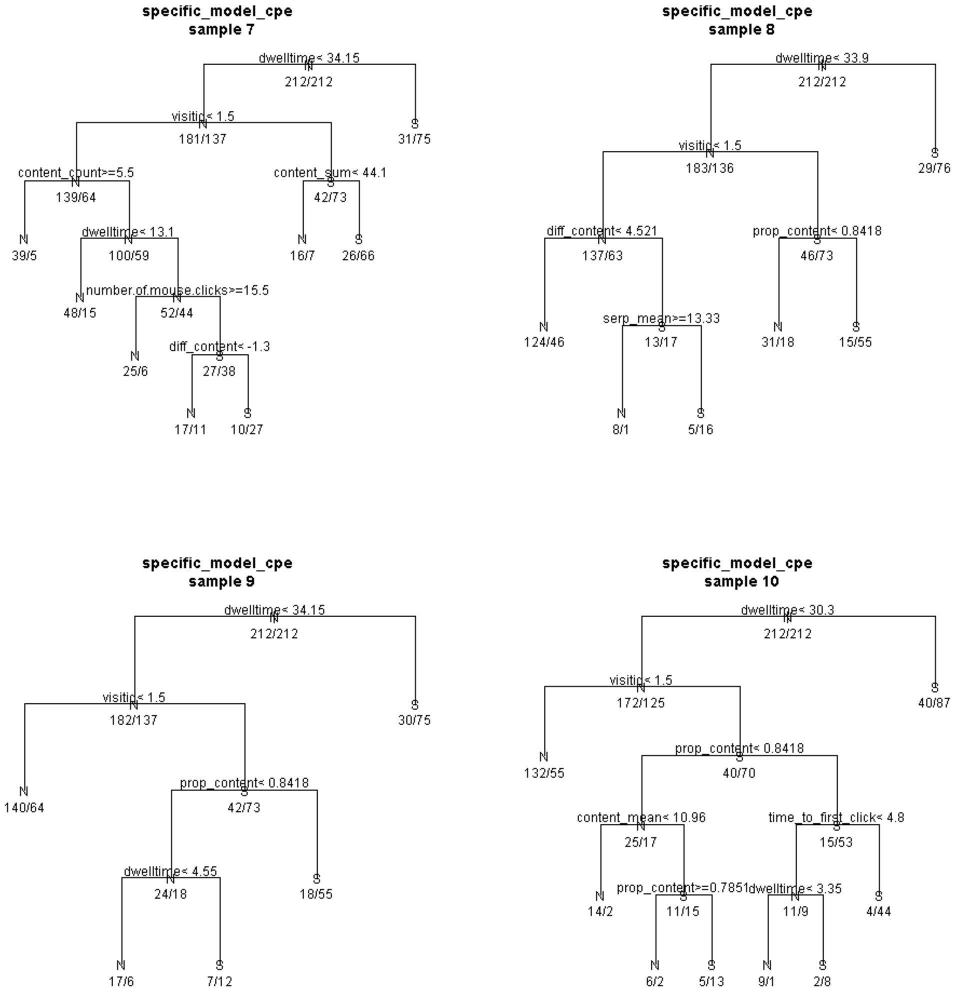
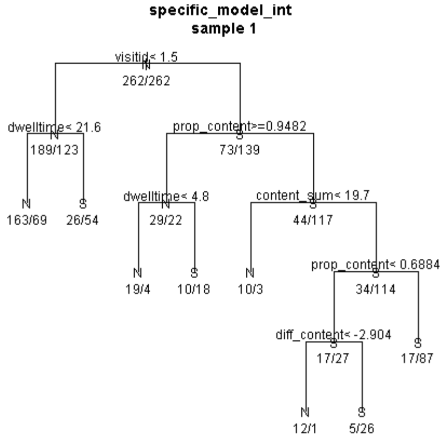
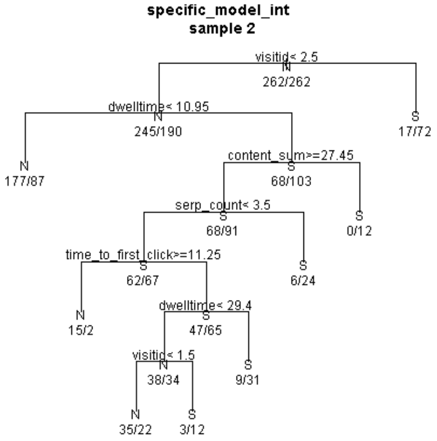
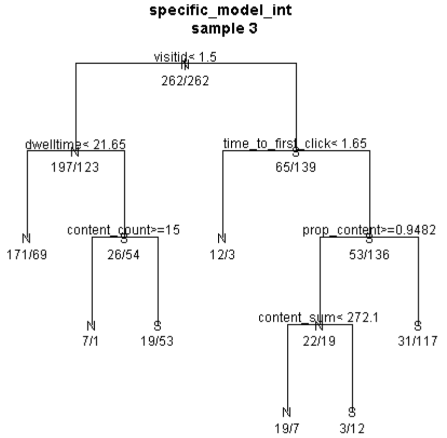


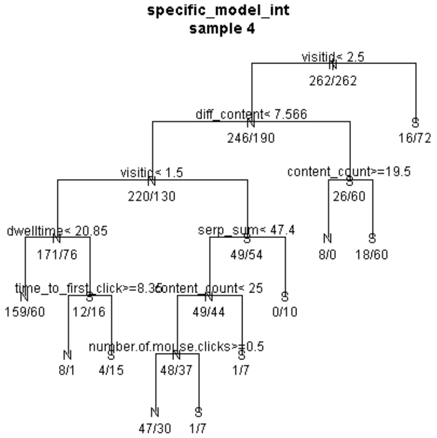
Figure 24. Predictive modeling of CPE on ten samples using recursive partitioning

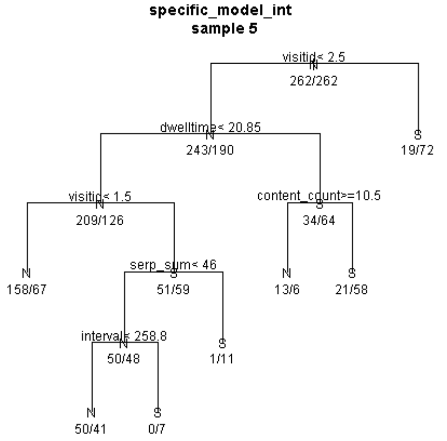
d. For INT task

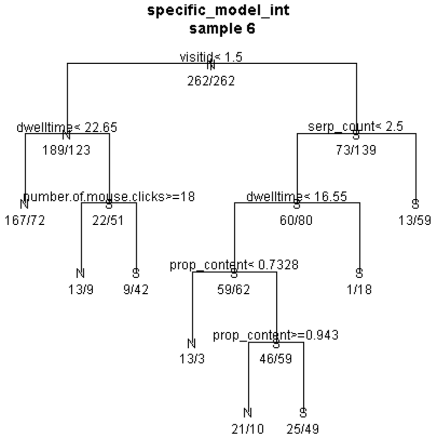


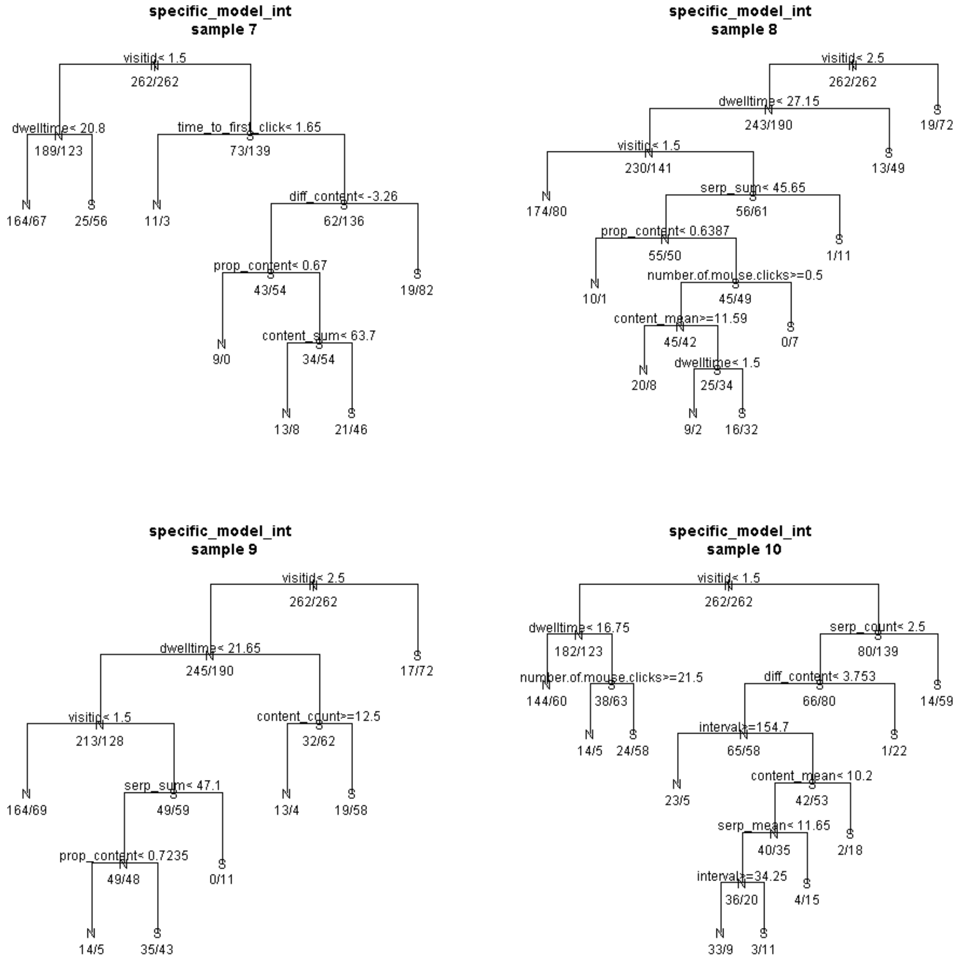












specific_model_int
sample 7

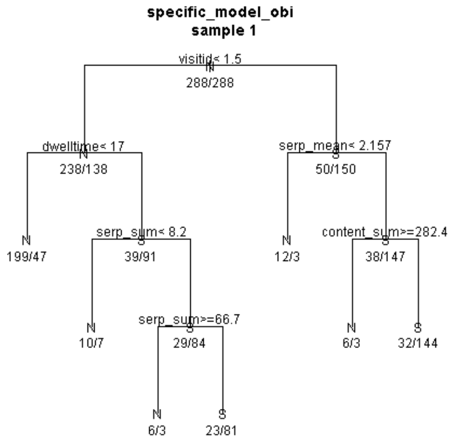
specific_model_int
sample 8

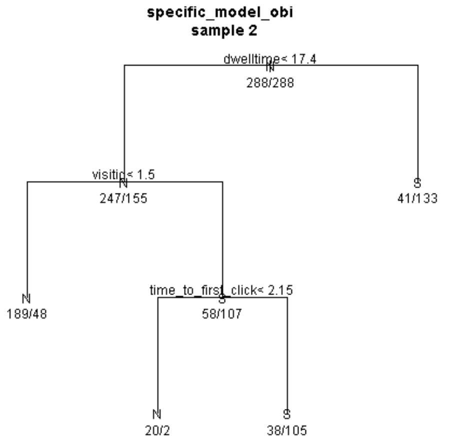
specific_model_int
sample 9

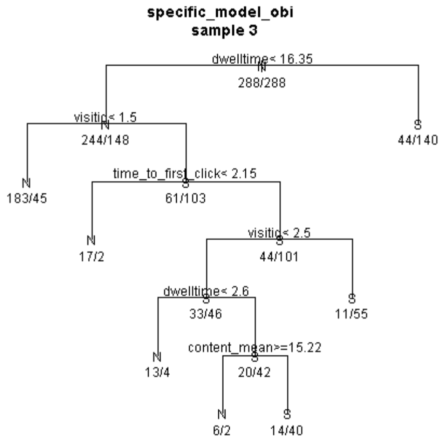
specific_model_int
sample 10

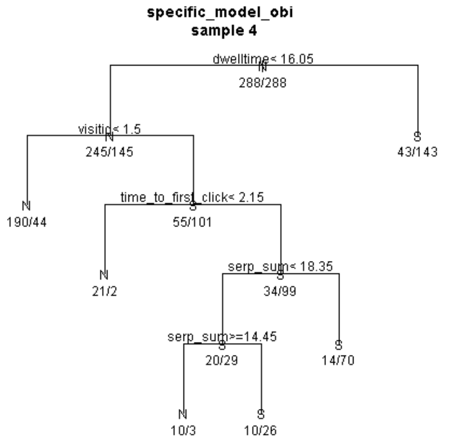
Figure 25. Predictive modeling of INT on ten samples using recursive partitioning

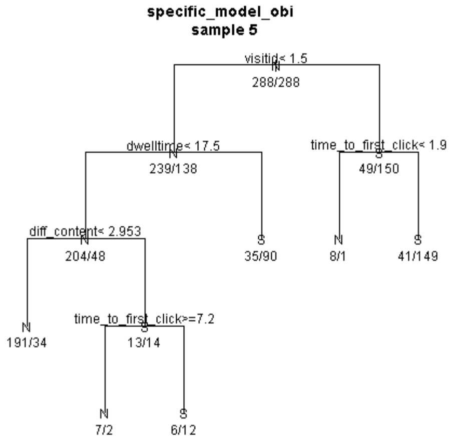
e. For OBI task

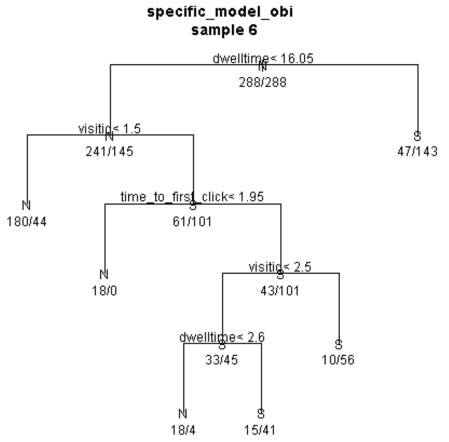












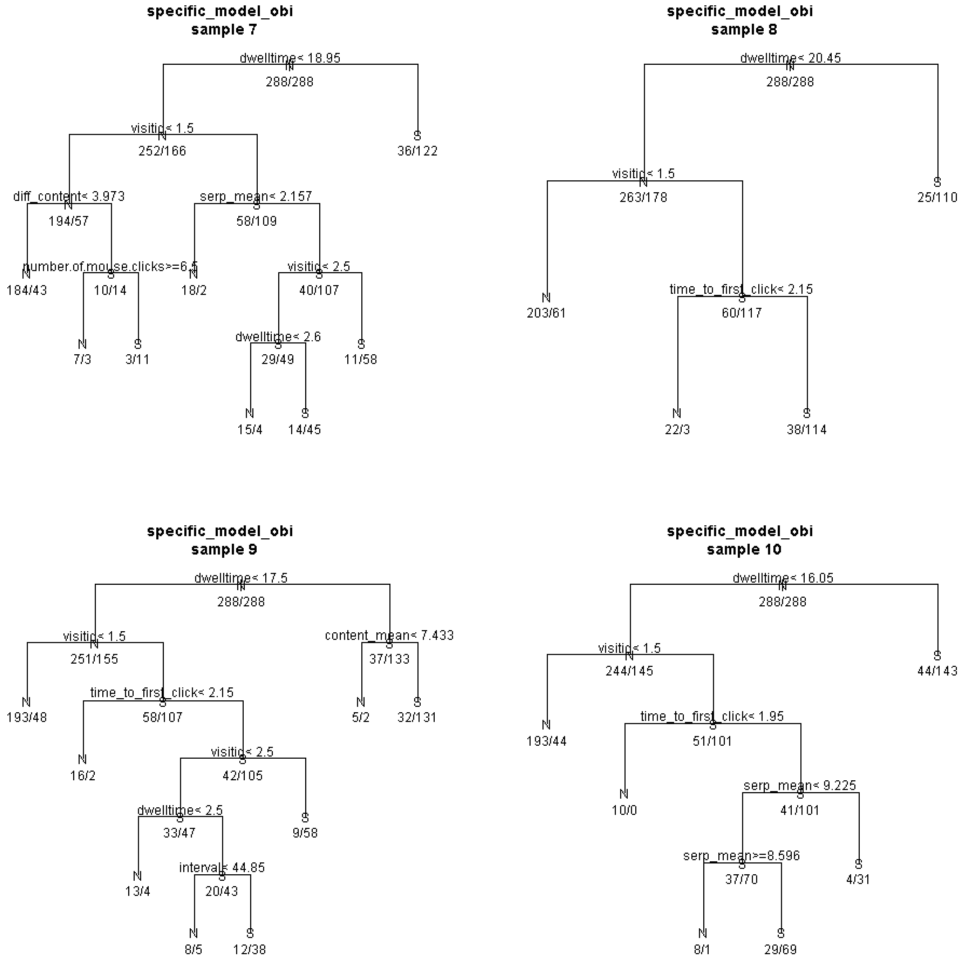


Figure 26. Predictive modeling of OBI on ten samples using recursive partitioning

C. Selected predictors in random samples in RQ1

a. For the general model

Table 61. Selected predictors in sample 1 (general model)

Variables selected	Beta	e^B	P-value
dwelltime	0.05	1.05	<.001
visit_id	0.60	1.82	<.001
number.of.keystrokes	-0.04	0.96	<.001

Table 62. Selected predictors in sample 2 (general model)

Variables selected	Beta	e^B	P-value
dwelltime	0.04	1.04	<.001
visit_id	0.53	1.70	<.001
number of mouse clicks	-0.01	0.99	0.02
number.of.keystrokes	-0.02	0.98	0.04
serp_sum	0.01	1.01	<.05
time_to_first_click	-0.02	0.98	<.05

Table 63. Selected predictors in sample 3 (general model)

Variables selected	Beta	e^B	P-value
dwelltime	0.05	1.05	<.001
visit_id	0.39	1.48	<.001
number.of.keystrokes	-0.04	0.96	<.001

Table 64. Selected predictors in sample 4 (general model)

Variables selected	Beta	e^B	P-value
dwelltime	0.05	1.05	<.001
visit_id	0.55	1.73	<.001
number.of.mouse.clicks	-0.01	0.99	<.05
number.of.keystrokes	-0.03	0.97	<.005

Table 65. Selected predictors in sample 5 (general model)

Variables selected	Beta	e^B	P-value
dwelltime	0.05	1.05	<.001
visit_id	0.53	1.70	<.001
number.of.mouse.clicks	-0.01	0.99	<.05
number.of.keystrokes	-0.05	0.95	<.001
content_sum	-0.003	1.00	<.05

Table 66. Selected predictors in sample 6 (general model)

Variables selected	Beta	e^B	P-value
dwelltime	0.05	1.05	<.001
visit_id	0.45	1.57	<.001
number.of.mouse.clicks	-0.01	0.99	<.05
number.of.keystrokes	-0.03	0.97	<.01
serp_sum	0.01	1.01	<.05

Table 67. Selected predictors in sample 7 (general model)

Variables selected	Beta	e^B	P-value
dwelltime	0.05	1.05	<.001
visit_id	0.51	1.67	<.001
number.of.mouse.clicks	-0.02	0.98	<.01
number.of.keystrokes	-0.03	0.97	<.01

Table 68. Selected predictors in sample 8 (general model)

Variables selected	Beta	e^B	P-value
dwelltime	0.05	1.05	<.001
visit_id	0.51	1.67	<.001
number.of.keystrokes	-0.04	0.96	<.001
content_sum	-0.003	1.00	<.05

Table 69. Selected predictors in sample 9 (general model)

Variables selected	Beta	e^B	P-value
dwelltime	0.05	1.05	<.001
visit_id	0.44	1.55	<.001
number.of.mouse.clicks	-0.01	0.99	<.05
number.of.keystrokes	-0.05	0.95	<.001
content_sum	-0.002	1.00	<.05

Table 70. Selected predictors in sample 10 (general model)

Variables selected	Beta	e^B	P-value
dwelltime	0.05	1.05	<.001
visit_id	0.47	1.60	<.001
number.of.keystrokes	-0.04	0.96	<.001
serp_sum	-0.01	0.99	<.05
time_to_first_click	-0.02	0.98	<.05

b. For BIC task

Table 71. Selected predictors for sample 1 (BIC)

Variables selected	Beta	e^B	P-value
dwelltime	0.06	1.06	<.001
number.of.keystrokes	-0.09	0.91	<.001

Table 72. Selected predictors for sample 2 (BIC)

Variables selected	Beta	e^B	P-value
dwelltime	0.05	1.05	<.001
number.of.keystrokes	-0.07	0.93	<.05
content_count	0.10	1.11	<.05

Table 73. Selected predictors for sample 3 (BIC)

Variables selected	Beta	e^B	P-value
dwelltime	0.06	1.06	<.001
number.of.keystrokes	-0.08	0.92	<.001
content_count	0.10	1.11	<.05

Table 74. Selected predictors for sample 4 (BIC)

Variables selected	Beta	e^B	P-value
dwelltime	0.06	1.06	<.001
number.of.keystrokes	-0.09	0.91	<.001
content_count	0.15	1.16	<.001
content_sum	-0.008	0.99	<.01

Table 75. Selected predictors for sample 5 (BIC)

Variables selected	Beta	e^B	P-value
dwelltime	0.05	1.05	<.001
number.of.keystrokes	-0.07	0.93	<.001
content_count	0.09	1.09	<0.05
content_sum	-0.006	0.99	<.05

Table 76. Selected predictors for sample 6 (BIC)

Variables selected	Beta	e^B	P-value
dwelltime	0.05	1.05	<.001
number.of.keystrokes	-0.07	0.93	<.01
content_count	0.11	1.12	<.01
content_sum	-0.007	0.99	<.05

Table 77. Selected predictors for sample 7 (BIC)

Variables selected	Beta	e^B	P-value
dwelltime	0.04	1.04	<.001
number.of.keystrokes	-0.07	0.93	<.01
content_count	0.13	1.14	<.01
content_sum	-0.007	0.99	<.05

Table 78. Selected predictors for sample 8 (BIC)

Variables selected	Beta	e^B	P-value
dwelltime	0.05	1.05	<.001
number.of.keystrokes	-0.06	0.94	<.05
content_count	0.11	1.12	<.01
content_sum	-0.006	0.99	<.05
time_to_first_click	-0.04	0.96	<.05

Table 79. Selected predictors for sample 9 (BIC)

Variables selected	Beta	e^B	P-value
dwelltime	0.05	1.05	<.001
number.of.keystrokes	-0.09	0.91	<.01
content_mean	0.05	1.05	<.05
content_count	0.18	1.20	<.001
content_sum	-0.01	0.99	<.001
prop_content	-1.62	0.20	<.05
time_to_first_click	-0.04	0.96	<.05

Table 80. Selected predictors for sample 10 (BIC)

Variables selected	Beta	e^B	P-value
dwelltime	0.05	1.05	<.001
number.of.mouse.clicks	-0.03	0.97	<.05
number.of.keystrokes	-0.07	0.93	<.01
content_mean	0.06	1.06	<.01
content_count	0.12	1.13	<.01
content_sum	-0.006	0.99	<.05
time_to_first_click	-0.04	0.96	<.05

c. For CPE task

Table 81. Selected predictors for sample 1 (CPE)

Variables selected	B	e^B	P-value
dwelltime	0.05	1.05	<.001
visit_id	0.76	2.14	<.001
number.of.mouse.clicks	-0.03	0.97	<.01

Table 82. Selected predictors for sample 2 (CPE)

Variables selected	B	e^B	P-value
dwelltime	0.05	1.05	<.001
visit_id	0.79	2.20	<.001
number.of.mouse.clicks	-0.02	0.98	<.05

Table 83. Selected predictors for sample 3 (CPE)

Variables selected	B	e^B	P-value
dwelltime	0.05	1.05	<.001
visit_id	0.79	2.20	<.001
number.of.mouse.clicks	-0.02	0.98	<.05

Table 84. Selected predictors for sample 4 (CPE)

Variables selected	B	e^B	P-value
dwelltime	0.04	1.04	<.001
visit_id	0.62	1.86	<.001
number.of.mouse.clicks	-0.02	0.98	<.01

Table 85. Selected predictors for sample 5 (CPE)

Variables selected	B	e^B	P-value
dwelltime	0.04	1.04	<.001
visit_id	0.68	1.97	<.001
number.of.mouse.clicks	-0.02	0.98	<.05

Table 86. Selected predictors for sample 6 (CPE)

Variables selected	B	e^B	P-value
dwelltime	0.04	1.04	<.001
visit_id	0.60	1.82	<.001
number.of.mouse.clicks	-0.03	0.97	<.01

Table 87. Selected predictors for sample 7 (CPE)

Variables selected	B	e^B	P-value
dwelltime	0.05	1.05	<.001
visit_id	0.75	2.12	<.001
number.of.mouse.clicks	-0.03	0.97	<.01
content_mean	-0.03	0.97	<.05

Table 88. Selected predictors for sample 8 (CPE)

Variables selected	B	e^B	P-value
dwelltime	0.06	1.06	<.001
visit_id	0.62	1.86	<.001
number.of.mouse.clicks	-0.03	0.97	<.01

Table 89. Selected predictors for sample 9 (CPE)

Variables selected	B	e^B	P-value
dwelltime	0.05	1.05	<.001
visit_id	0.72	2.05	<.001

number.of.mouse.clicks	-0.02	0.98	<.05

Table 90. Selected predictors for sample 10 (CPE)

Variables selected	B	e^B	P-value
dwelltime	0.05	1.05	<.001
visit_id	0.67	1.95	<.001
number.of.mouse.clicks	-0.03	0.97	<.01
content_mean	-0.03	0.97	<.05

d. For INT task

Table 91. Selected predictors for sample 1 (INT)

Variables selected	B	e^B	P-value
dwelltime	0.06	1.06	<.001
visit_id	0.77	2.16	<.001
number.of.mouse.clicks	-0.03	0.97	<.05
number.of.keystrokes	-0.05	0.95	<.05
serp_mean	-0.10	0.90	<.01
serp_sum	0.05	1.05	<.01

Table 92. Selected predictors for sample 2 (INT)

Variables selected	B	e^B	P-value
dwelltime	0.05	1.05	<.001
visit_id	0.88	2.41	<.001
number.of.mouse.clicks	-0.03	0.97	<.05
serp_mean	-0.10	0.90	<.01
serp_sum	0.04	1.04	<.05

Table 93. Selected predictors for sample 3 (INT)

Variables selected	B	e^B	P-value
dwelltime	0.06	1.06	<.001
visit_id	0.93	2.53	<.001
number.of.mouse.clicks	-0.04	0.96	<.01
content_sum	-0.006	0.99	<.01
serp_mean	-0.13	0.88	<.001
serp_count	-0.30	0.74	<.05
serp_sum	0.06	1.06	<.001

Table 94. Selected predictors for sample 4 (INT)

Variables selected	B	e^B	P-value
dwelltime	0.06	1.06	<.001
visit_id	0.89	2.44	<.001
number.of.mouse.clicks	-0.05	0.95	<.01
serp_count	-0.28	0.76	<.05
serp_sum	0.06	1.06	<.001

Table 95. Selected predictors for sample 5 (INT)

Variables selected	B	e^B	P-value
dwelltime	0.05	1.05	<.001
visit_id	0.68	1.97	<.001
number.of.mouse.clicks	-0.04	0.96	<.01
content_count	0.06	1.06	<.05
content_sum	-0.006	0.99	<.01
serp_mean	-0.13	0.88	<.001
serp_count	-0.26	0.77	<.05
serp_sum	0.07	1.07	<.001

Table 96. Selected predictors for sample 6 (INT)

Variables selected	B	e^B	P-value
dwelltime	0.07	1.07	<.001
visit_id	0.74	2.10	<.001
number.of.mouse.clicks	-0.05	0.95	<.001
serp_mean	-0.15	0.86	<.001
serp_count	-0.29	0.75	<.05
serp_sum	0.07	1.07	<.001

Table 97. Selected predictors for sample 7 (INT)

Variables selected	B	e^B	P-value
dwelltime	0.07	1.07	<.001
visit_id	0.68	1.97	<.001
number.of.mouse.clicks	-0.04	0.96	<.01
serp_sum	0.05	1.05	<.01

Table 98. Selected predictors for sample 8 (INT)

Variables selected	B	e^B	P-value
dwelltime	0.06	1.06	<.001
visit_id	0.82	2.27	<.001
number.of.mouse.clicks	-0.04	0.96	<.05
content_mean	0.05	1.05	<.05
content_count	0.08	1.08	<.01
content_sum	-0.005	1.00	<.05
serp_mean	-0.15	0.86	<.001
serp_count	-0.37	0.69	<.01
serp_sum	0.06	1.06	<.01

Table 99. Selected predictors for sample 9 (INT)

Variables selected	B	e^B	P-value
dwelltime	0.04	1.04	<.001
visit_id	0.87	2.39	<.001
number.of.mouse.clicks	-0.05	0.95	<.001
content_sum	-0.004	1.00	<.05
serp_mean	-0.13	0.88	<.01
serp_count	-0.27	0.76	<.05
serp_sum	0.06	1.06	<.01

Table 100. Selected predictors for sample 10 (INT)

Variables selected	B	e^B	P-value
dwelltime	0.06	1.06	<.001
visit_id	0.72	2.05	<.001
number.of.mouse.clicks	-0.05	0.95	<.001
serp_mean	-0.13	0.88	<.001
serp_count	-0.22	0.80	<.05
serp_sum	0.06	1.06	<.001

e. For OBI task

Table 101. Selected predictors for sample 1 (OBI)

Variables selected	B	e^B	P-value
dwelltime	0.07	1.07	<.001
visit_id	0.86	2.36	<.001

Table 102. Selected predictors for sample 2 (OBI)

Variables selected	B	e^B	P-value
dwelltime	0.07	1.07	<.001
visit_id	0.74	2.10	<.001

Table 103. Selected predictors for sample 3 (OBI)

Variables selected	B	e^B	P-value
dwelltime	0.08	1.08	<.001
visit_id	0.75	2.12	<.001
content_count	-0.13	0.88	<.01

Table 104. Selected predictors for sample 4 (OBI)

Variables selected	B	e^B	P-value
dwelltime	0.09	1.09	<.001
visit_id	0.84	2.32	<.001
content_count	-0.11	0.90	<.05
serp_mean	0.15	1.16	<.01
serp_count	0.21	1.23	<.05

Table 105. Selected predictors for sample 5 (OBI)

Variables selected	B	e^B	P-value
dwelltime	0.10	1.11	<.001
visit_id	1.10	3.00	<.001

Table 106. Selected predictors for sample 6 (OBI)

Variables selected	B	e^B	P-value
dwelltime	0.08	1.08	<.001
visit_id	1.01	2.75	<.001
content_count	-0.15	0.86	<.01
serp_count	0.22	1.25	<.05

Table 107. Selected predictors for sample 7 (OBI)

Variables selected	B	e^B	P-value
dwelltime	0.07	1.07	<.001
visit_id	0.89	2.44	<.001
content_count	-0.11	0.90	<.05

Table 108. Selected predictors for sample 8 (OBI)

Variables selected	B	e^B	P-value
dwelltime	0.06	1.06	<.001
visit_id	0.60	1.82	<.001

Table 109. Selected predictors for sample 9 (OBI)

Variables selected	B	e^B	P-value
dwelltime	0.08	1.08	<.001
visit_id	0.93	2.53	<.001

Table 110. Selected predictors for sample 10 (OBI)

Variables selected	B	e^B	P-value
dwelltime	0.09	1.09	<.001
visit_id	0.97	2.64	<.001
content_count	-0.14	0.87	<.01
content_sum	0.007	1.01	<.05

D. Descriptive analysis results on whole-session level measures in RQ2

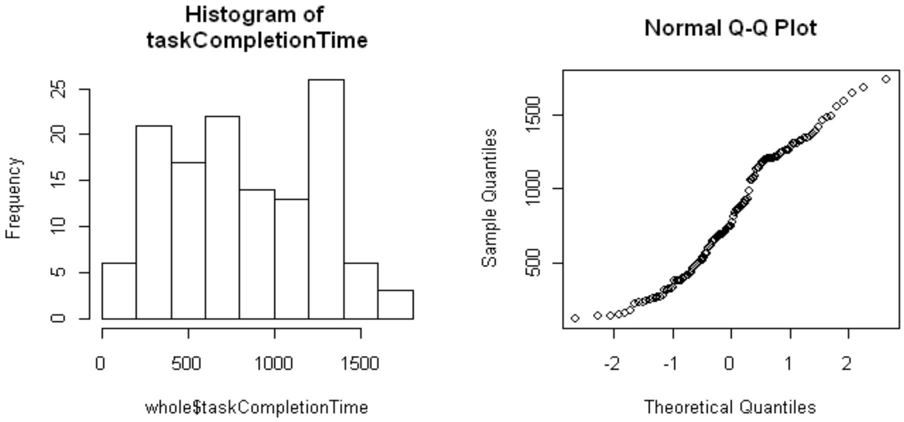


Figure 27. The histogram and the normal Q-Q Plot of task completion time

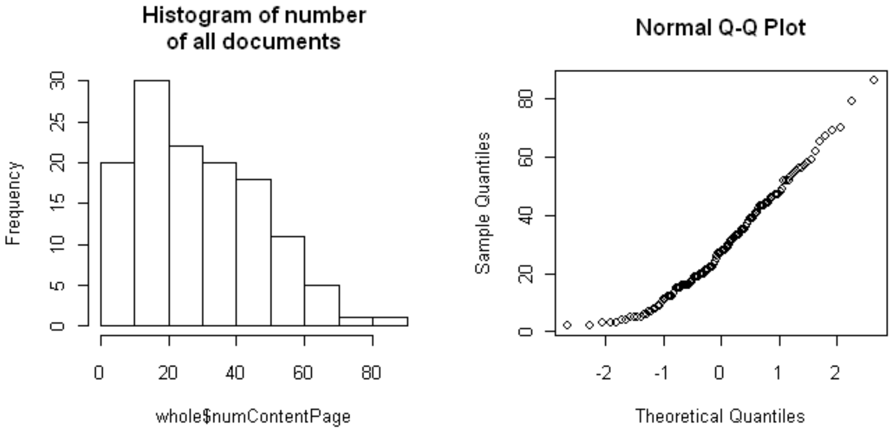


Figure 28. The histogram and the normal Q-Q Plot of number of all documents

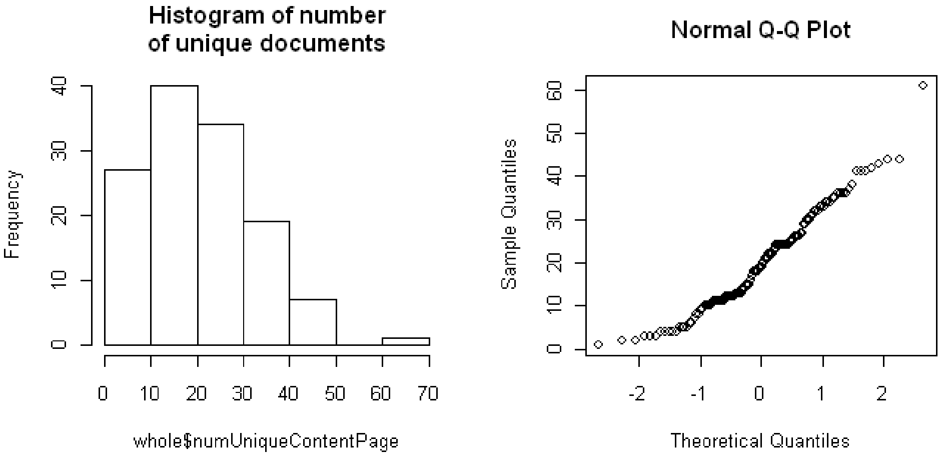


Figure 29. The histogram and the normal Q-Q Plot of Numbers of unique documents

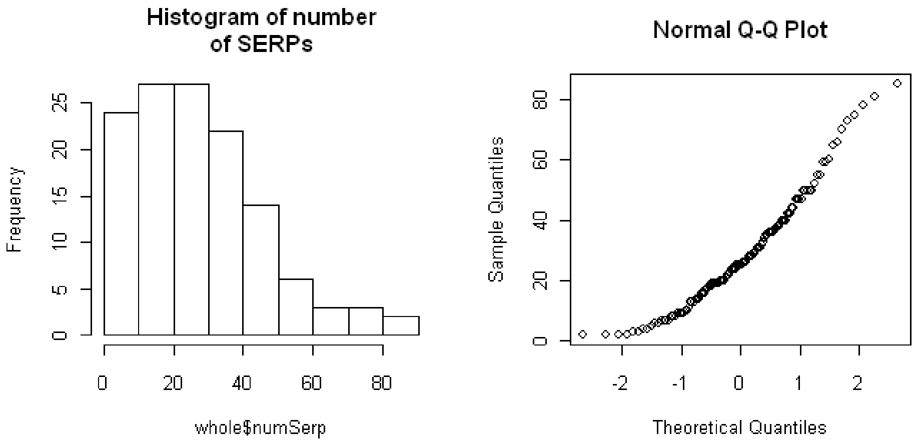


Figure 30. The histogram and the normal Q-Q Plot of Number of SERPs

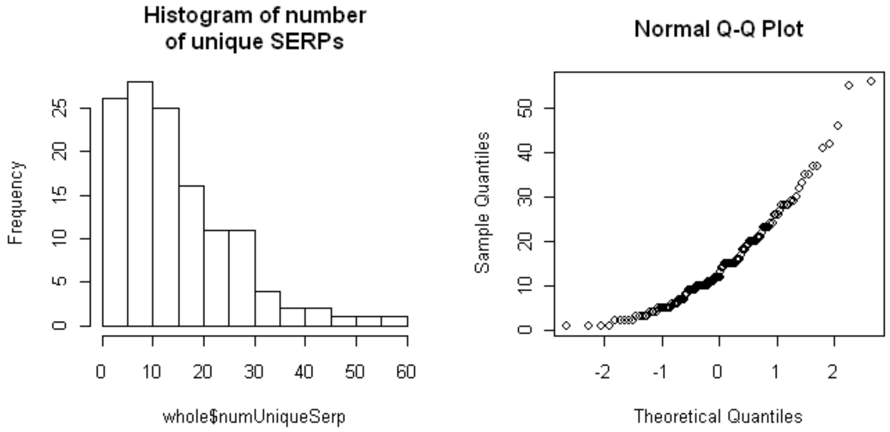


Figure 31. The histogram and the normal Q-Q Plot of Number of unique SERPs

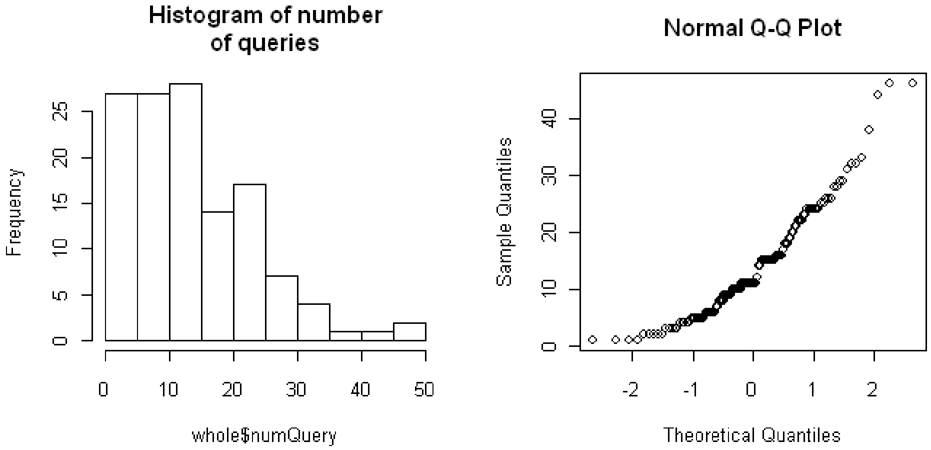


Figure 32. The histogram and the normal Q-Q Plot of Number of queries

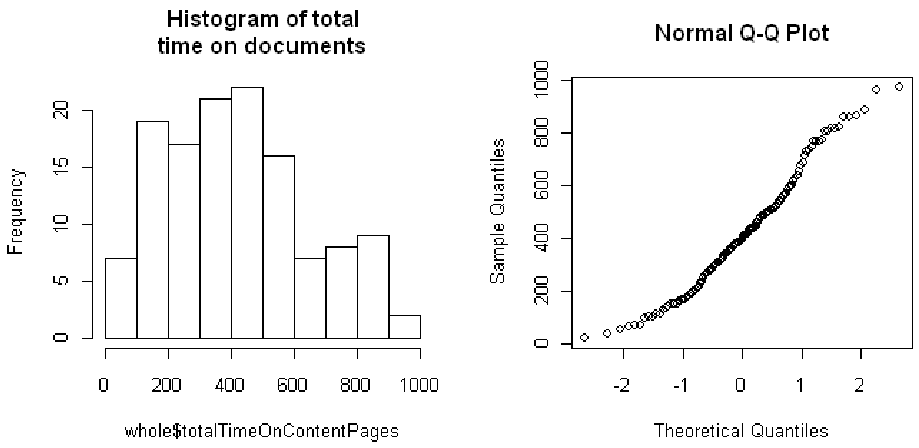


Figure 33. The histogram and the normal Q-Q Plot of Total time spent on documents (seconds)

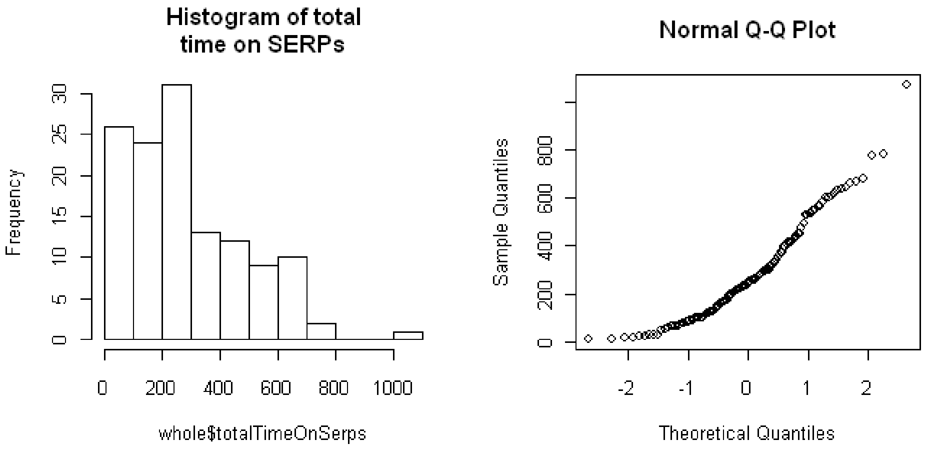


Figure 34. The histogram and the normal Q-Q Plot of Total time spent on SERPs (seconds)

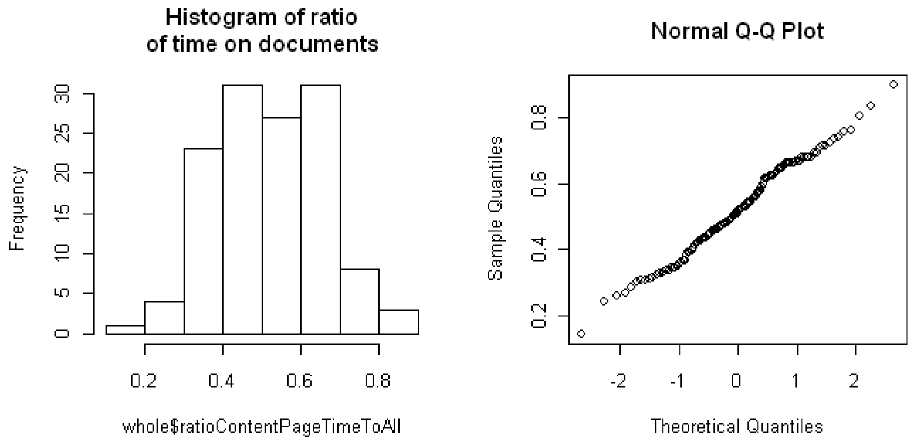


Figure 35. The histogram and the normal Q-Q Plot of Ratio of document time to all

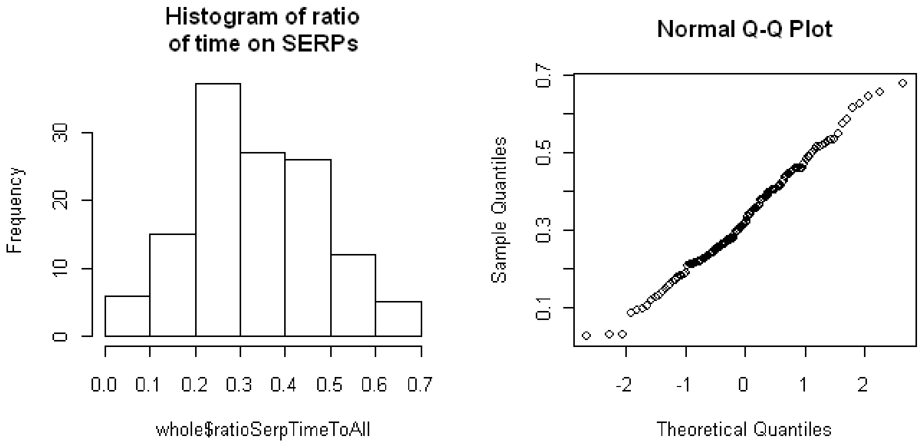


Figure 36. The histogram and the normal Q-Q Plot of Ratio of SERP time to all

E. *Descriptive analysis results on within-session level measures in RQ2*

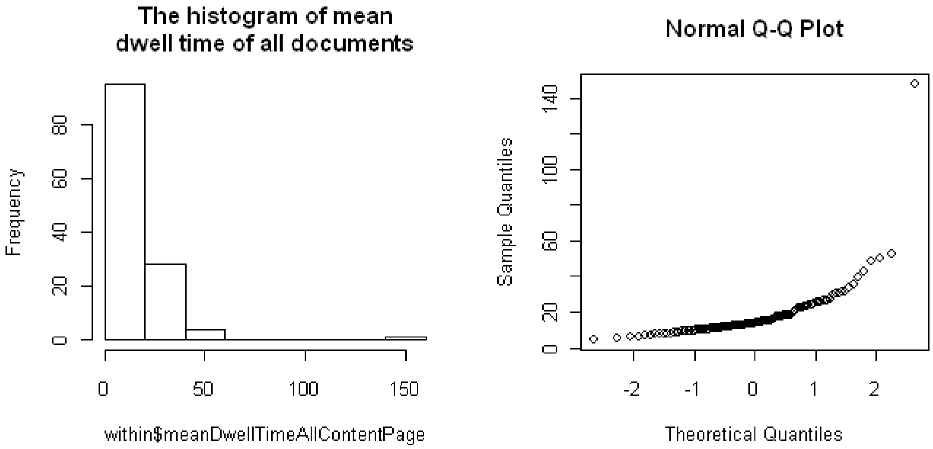


Figure 37. The histogram and the normal Q-Q Plot of Mean dwell time of all documents

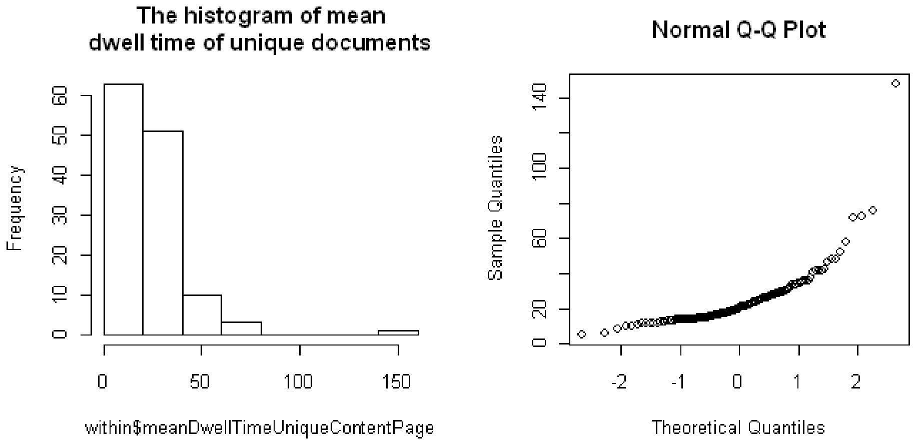


Figure 38. The histogram and the normal Q-Q Plot of Mean dwell time of unique documents

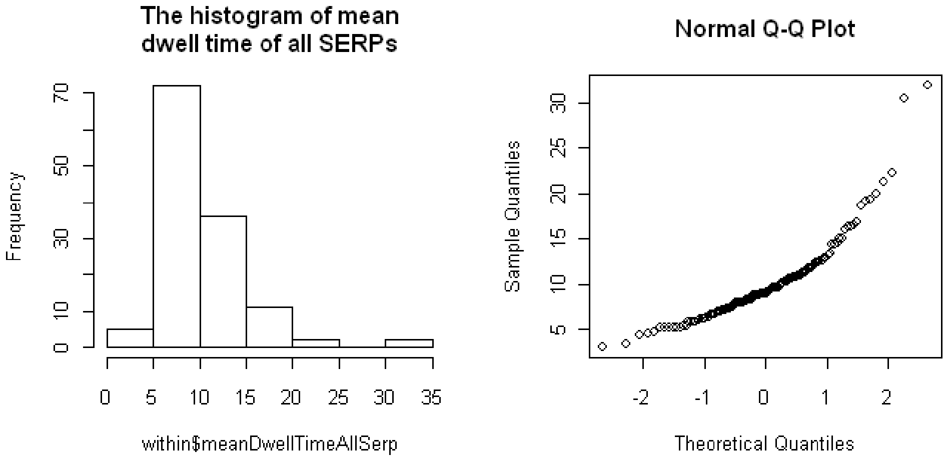


Figure 39. The histogram and the normal Q-Q Plot of Mean dwell time of all SERPs

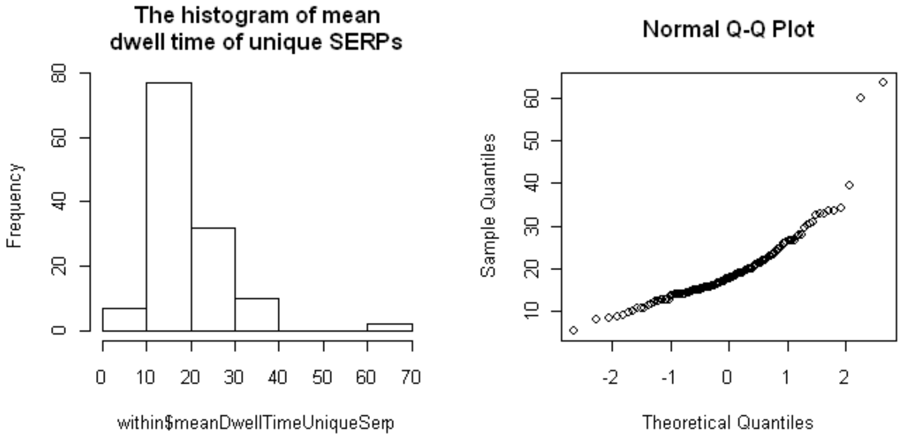


Figure 40. The histogram and the normal Q-Q Plot of Mean dwell time of unique SERPs

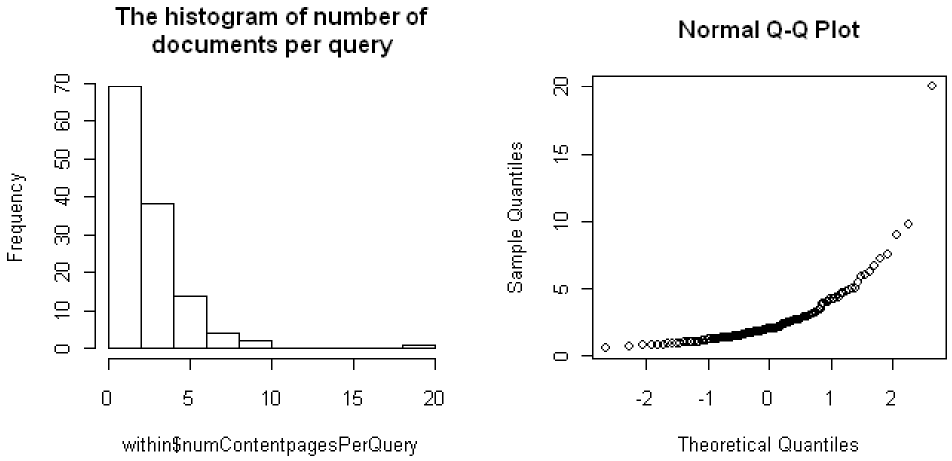


Figure 41. The histogram and the normal Q-Q Plot of Number of documents per query

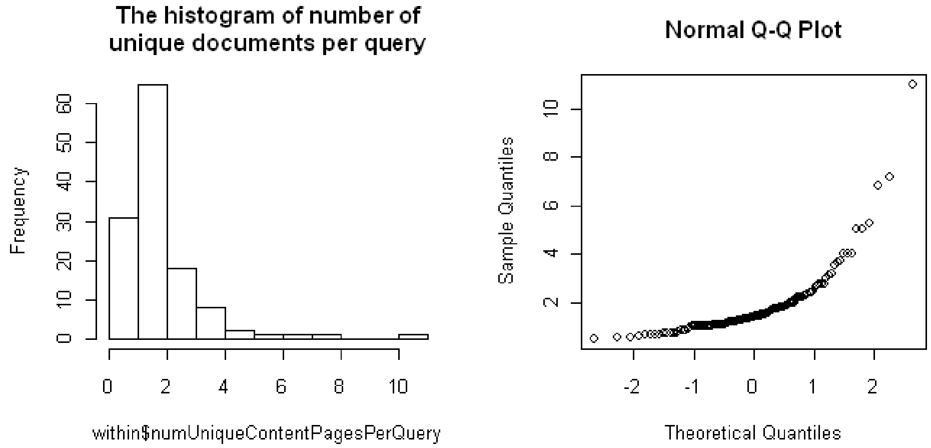


Figure 42. The histogram and the normal Q-Q Plot of Number of unique documents per query

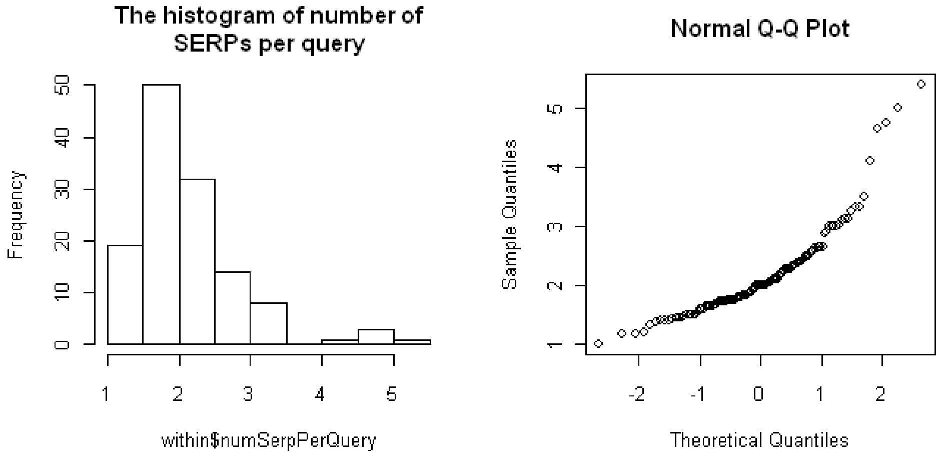


Figure 43. The histogram and the normal Q-Q Plot of Number of SERPs per query

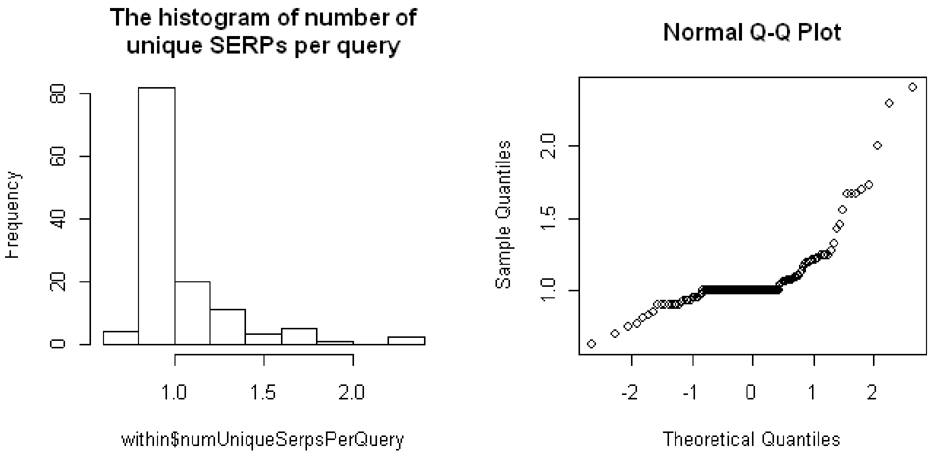


Figure 44. The histogram and the normal Q-Q Plot of Number of unique SERPs per query

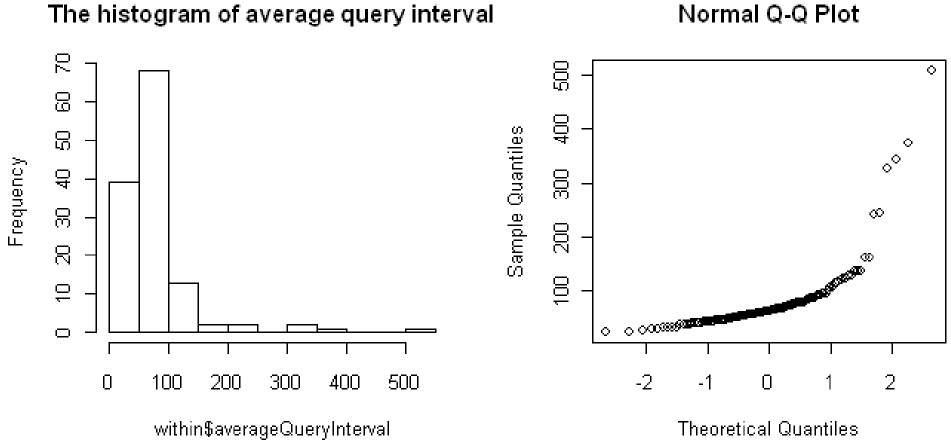


Figure 45. The histogram and the normal Q-Q Plot of Average query interval

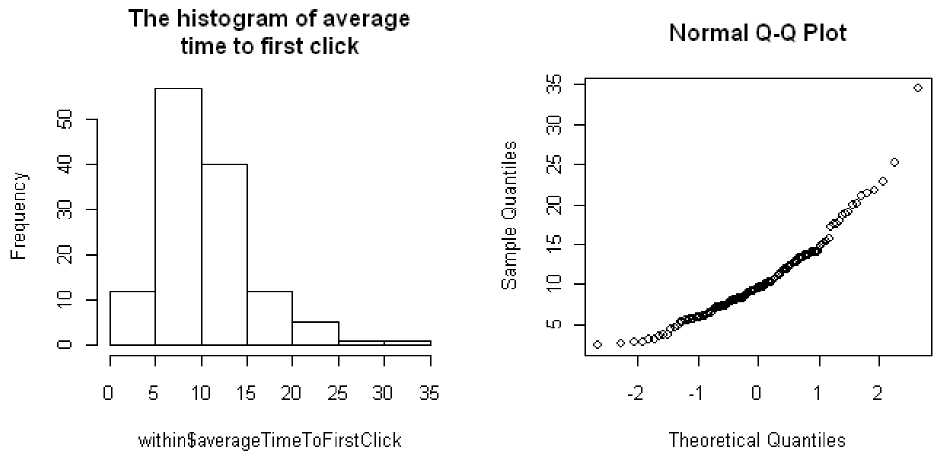


Figure 46. The histogram and the normal Q-Q Plot of Average Time To First Click

References

Agichtein, E., Brill, E., Dumais, S., & Ragno, R. (2006). Learning user interaction models for predicting web search result preferences. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 3-10). Seattle, Washington, USA.

Aula, A., Khan, R. M., & Guan, Z. (2010). How does search behavior change as search becomes more difficult?. In *Proceedings of the 28th international Conference on Human Factors in Computing Systems* (pp. 35-44). Atlanta, Georgia, USA.

Bai, J., Nie, J., Cao, G., & Bouchard, H. (2007). Using query contexts in information retrieval. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, The Netherlands. 15-22.s

Beeferman, D. & Berger, A. (2000). Agglomerative clustering of a search engine query log. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '00). ACM, New York, NY, USA, 407-416. DOI=10.1145/347090.347176 http://doi.acm.org/10.1145/347090.347176

Belkin, N. J., Carballo, J. P., Cool, C., Lin, S., Park, S. Y., Rieh, S. Y., et al. (1998). Rutgers' TREC-6 interactive track experience. *Proceedings of the Sixth Text REtrieval Conference*, 597-610.

Bierig, R., Cole., M.J., & Gwizdka, J. (2009). A user centered experiment and logging framework for interactive information retrieval. In N.J. Belkin, R. Bierig, G. Buscher, L. van Elst, J. Gwizdka, J. Jose, et al. (Eds), CEUR Workshop Proceedings: 512. *Proceedings of the SIGIR 2009 Workshop on Understanding the User: Logging and interpreting user interactions in information search and retrieval*, UIIR'2009 (pp. 8-11). Aachen, Germany: CEUR Workshop Proceedings.

Borlund, P. (2003). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. Information Research, 8(3), paper no. 152. Retrieved from http://informationr.net/ir/8-3/paper152.html.

Breiman, L. (1984). *Classification and Regression Trees*. Boca Raton: Chapman & Hall/CRC.

Broder, A. (2002). A taxonomy of Web search. *SIGIR Forum*, 36(2).

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

Campbell I & van Rijsbergen CJ (1996) The ostensive model of developing information. In: Ingwersen P and Pors NO, Eds. Information Science: Integration in Perspective, *Proceedings of the CoLIS-2 conference*, Copenhagen, pp. 251-268.

Chang, Y.-S., He, K.-Y., Yu, S., & Lu, W.-H. (2006). Identifying User Goals from Web Search Results. In *Proceedings of the 2006 IEEE/WIC/ACM international Conference on Web intelligence* (pp. 1038-1041). Web Intelligence. IEEE Computer Society, Washington, D.C.

Chen, S.Y. & Ford, N. (1998). Modeling user navigation behaviours in a hypermedia based learning system: An individual differences approach. Knowledge Organization, 25(3), 67-78.

Chirita, P., Firan, C. S., & Nejdl, W. (2006). Summarizing local context to personalize global web search. Proceedings of the 15th ACM International Conference on Information and Knowledge Management, Arlington, Virginia, USA. 287-296.

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

Claypool, M., Le, P., Wased, M., & Brown, D. (2001). Implicit interest indicators. In *Proceedings of the 6th International Conference on Intelligent User Interfaces* (pp. 33-40). Santa Fe, New Mexico, United States.

Cole, M., Gwizdka, J, **Liu, C.**, Belkin, N. J. (2011). Dynamic Assessment of Information Acquisition Effort during Interactive Search. In: *Proceedings of the Annual Conference of the American Society for Information Science & Technology (ASIS&T) 2011 (10p.).* New Orleans, LA, October 9-13, 2011.

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. Position paper presented at *the 3rd Workshop on Human-Computer Interaction and Information Retrieval (HCIR) 2009,* October 23, 2009, Washington DC.

Cool, C., & Spink, A. (2002). Issues of context in information retrieval (IR): an introduction to the special issue. *Information Processing and Management*, *38*(5), 605–611.

Dou, Z., Song, R., & Wen, J. (2007). A large-scale evaluation and analysis of personalized search strategies. *Proceedings of the 16th International Conference on World Wide Web,* Banff, Alberta, Canada. 581-590.s

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

Ford, N. & Chen, S.Y. (2000). Individual differences, hypermedia navigation, and learning: an empirical study. *J. Educ. Multimedia Hypermedia* 9(4), 281-311.

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

Goecks, J., & Shavlik, J. (2000). Learning users' interests by unobtrusively observing their normal behavior. *Proceedings of the 5th International Conference on Intelligent User Interfaces,* New Orleans, Louisiana, United States. 129-132.

Goldstein, K.M., & Blackman, S. (1978). Cognitive style: Five approaches and relevant research. New York: John Wiley.

Guo, Q., and Agichtein, E. (2012). Beyond Dwell Time: Estimating Document Relevance from Cursor Movements and other Post-click Searcher Behavior. In *Proceedings of the 21st* *international conference on World Wide Web (WWW '12)*. ACM, New York, NY, USA, 569-578.

Hassan, A., Jones, R., & Klinkner, K. L. (2010). Beyond DCG: User behavior as a predictor of a successful search. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (pp. 221-230). New York, New York, USA.

Hosmer, D. W., Lemeshow, S., John Wiley & Sons., & Wiley InterScience (Online service). (2000). *Applied logistic regression*. New York: Wiley.

Huang, C., Chien, L., & Oyang, Y. (2003). Relevant term suggestion in interactive web search based on contextual information in query session logs. *J.Am.Soc.Inf.Sci.Technol., 54*(7), 638-649.

Huang, J., White, R.W., Buscher, G., Wang, K. (2012). Improving Searcher Models Using Mouse Cursor Activity *SIGIR 2012*, pp. 195-204.

Joachims, T. (2002). Optimizing Search Engines Using Clickthrough Data. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 133–142), Edmonton, Alberta, Canada.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., & Gay, G. (2007). Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans.Inf.Syst., 25*(2)

Kanoulas, E., Carterette, B., Clough P., and Sanderson, M. (2011) In *The Twentieth Text REtrieval Conference Proceedings (TREC 2011),* National Institute of Standards and Technology, 2012.

Kellar, M., Watters, C., Duffy, J., & Shepherd, M. (2004). Effect of task on time spent reading as an implicit measure of interest. *Proceedings of the American Society for Information Science and Technology, 41*(1), 168-175. doi: 10.1002/meet.1450410119.

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

Kelly, D. (2005). Implicit feedback: Using behavior to infer relevance. In A. Spink and C. Cole (Eds.) *New Directions in Cognitive Information Retrieval* (pp.169-186). Netherlands: Springer Publishing.

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* (pp. 377-384). Sheffield, UK.

Kelly, D. & Cool, C. (2002). The effects of topic familiarity on information search behavior. In *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries* (JCDL '02). ACM, New York, NY, USA, 74-75. DOI=10.1145/544220.544232 http://doi.acm.org/10.1145/544220.544232

Kelly, D., Gyllstrom, K., & Bailey, E. W. (2009). A comparison of query and term suggestion features for interactive searching. *Proceedings of the 32nd International ACM SIGIR*

Conference on Research and Development in Information Retrieval, Boston, MA, USA. 371-378.

Kelly, D., & Teevan, J. (2003). Implicit feedback for inferring user preference: A bibliography. *SIGIR Forum, 37*(2), 18-28.

Kim, J. (2009). Describing and predicting information-seeking behavior on the web. *Journal of the American Society for Information Science and Technology, 60*(4), 679-693.

Kim, K.-S. (2001). Information seeking on the Web: Effects of user and task variables. *Library & Information Science Research,* 23, 233–255.

Konstan et al., (1997).Applying collaborative filtering to usenet news. Communications of the ACM. (40), 3, 77-87.

Lee, U., Liu, Z., & Cho, J. 2005. Automatic identification of user goals in Web search. In *Proceedings of the 14th international Conference on World Wide Web* (pp. 391-400) Chiba, Japan.

Lewis, D.D. (1995). Evaluating and optimizing autonomous text classification systems. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '95),* Edward A. Fox, Peter Ingwersen, and Raya Fidel (Eds.). ACM, New York, NY, USA, 246-254.

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

Li, Y. & Belkin, N.J. (2008). A faceted approach to conceptualizing tasks in information seeking. *Inf. Process. Manage*. 44, 6 (November 2008), 1822-1837. DOI=10.1016/j.ipm.2008.07.005 http://dx.doi.org/10.1016/j.ipm.2008.07.005

Liu, C., Gwizdka, J., & Belkin, N.J. (2010). Analysis of Query Reformulation Types on Different Search Tasks. *A poster presented at i-conference 2010*, Urbana-Champaign, IL, February 3-6, 2010. Retrieved from http://hdl.handle.net/2142/15049.

Liu, C., Gwizdka, J., & Liu, J. (2010). Helping identify when users find useful documents: Examination of query reformulation intervals. The *proceedings of the3rd Information Interaction in Context Symposium (IIiX'2010)*.

Liu, C., Gwizdka, J., Liu, J., Xu, T., & Belkin, NJ. (2010). Analysis and Evaluation of Query Reformulations in Different Task Types. In Proceedings of ASIST 2010.Pittsburgh, PA. October 22-27, 2010.

Liu, J. & Belkin, N.J. (2010). Personalizing information retrieval for multi-session tasks: The roles of task stage and task type. *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR '10)*. Geneva, Switzerland, July 19-23, 2010.

Liu J., Cole M., Liu C., Bierig R., Gwizdka J., Belkin N.J, Zhang J, & Zhang. X. **(**2010). Search behaviors in different task types. *Proceedings of ACM-IEEE Computer Society Joint Conference on Digital Libraries (JCDL) 2010*. Goldcoast, Australia, June 21-25, 2010.

Liu, J., Gwizdka, J., Liu, C., & Belkin, N.J. (2010). Predicting task difficulty for different task types. To appear in *Proceedings of ASIST 2010.*

Liu, F., Yu, C., & Meng, W. (2002). Personalized web search by mapping user queries to categories. *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, McLean, Virginia, USA. 558-565.

Loper, E. and Bird, S. (2002). NLTK: The Natural Language Toolkit. Proceedings of the ACL02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistic, Volume 1, July, 2002 Association for Computational Linguistics.

MacMullin, S. D., & Taylor, R. S. (1984). Problem dimensions and information traits. *The Information Society*, 3, 91–111.

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.

Melucci, M., & White, R. W. (2007). Utilizing a geometry of context for enhanced implicit feedback. *CIKM '07: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, Lisbon, Portugal. 273-282.

Morita, M. & Shinoda, Y. (1994). Information Filtering Based on User Behavior Analysis and Best MatchText Retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 272–281), Dublin, Ireland.

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

Nichols, D. (1997). Implicit Rating and Filtering. In *Proceedings of the 5th DELOS Workshop on Filtering and Collaborative Filtering* (pp. 31—36). Budapest, Hungary.

Radlinski, F. & Joachims, T. (2005). Query Chains: Learning to rank from implicit feedback. In *Proceedings of the 11th Conference on Knowledge Discovery and Data mining* (pp. 239-248). Chicago, Illinois, USA.

Rocchio,J.J. (1971). Relevance feedback in information retrieval. In *Salton, G. The SMART Retrieval System: Experiements in Automatic Document Processing* (pp. 313-323). Upper Saddle River, NJ:Prentice-Hall Inc.

Oard, D.W., & Kim, J. (1998) Implicit Feedback for Recommender Systems. In *AAAI Workshop on Recommender Systems*, Madison, WI: 81-83. http://www.glue.umd.edu/~oard/research.html

Oard, D. W. & Kim, J. (2001). Modeling Information Content Using Observable Behavior. In *Proceedings of the 64th Annual Meeting of the American Society for Information Science and Technology* (pp. 38–45). Washington, D.C., USA.

Rose, D. E. & Levinson, D. (2004). Understanding user goals in web search. In *Proceedings of the 13th international Conference on World Wide Web* (pp. 13-19) New York, NY, USA.

Shapira, B., Taieb-Maimon, M., & Moskowitz, A. (2006). Study of the usefulness of known and new implicit indicators and their optimal combination for accurate inference of users interests. In *Proceedings of the 2006 ACM Symposium on Applied Computing* (Dijon, France, April 23 - 27, 2006). SAC '06. ACM, New York, NY, 1118-1119. DOI= http://doi.acm.org/10.1145/1141277.1141542

Shen, X., Tan, B., & Zhai, C. (2005). Context-sensitive information retrieval using implicit feedback. *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Salvador, Brazil. 43-50.

Tan, B., Shen, X., & Zhai, C. (2006). Mining long-term search history to improve search accuracy. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* Philadelphia, PA, USA. 718-723.

Teevan, J., Dumais, S. T., & Horvitz, E. (2005). Personalizing search via automated analysis of interests and activities. *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Salvador, Brazil. 449-456.

Teevan, J., Dumais, S. T., & Horvitz, E. (2010). Potential for personalization. *ACM Trans.Comput.-Hum.Interact., 17*(1), 4:1-4:31.

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 the 31st Annual ACM Conference on Research and Development in Information Retrieval* (pp. 163-170), Singapore, Singapore.

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.

Toms, E., O'Brien, H., Mackenzie, T., Jordan, C., Freund, L., Toze, S., et al. (2008). Task effects on interactive search: The query factor. In N. Fuhr, J. Kamps, M. Lalmas & A. Trotman (Eds.), *Focused access to XML documents* (pp. 359-372) Springer Berlin / Heidelberg.

Tyler, S. K., and Teevan, J. (2010). Large scale query log analysis of re-finding. In *Proceedings of the third ACM international conference on Web search and data mining (WSDM '10).* ACM, New York, NY, USA, 191-200.

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

Rocchio,J.J. (1971). Relevance feedback in information retrieval. In Salton, G. *The SMART Retrieval System: Experiments in Automatic Document Processing* (pp. 313-323). Upper Saddle River, NJ: Prentice-Hall Inc.

Wang, P., Hawk, W. B., & Tenopir, C. (2000). Users' interaction with world wide web resources: An exploratory study using a holistic approach. *Inf.Process.Manage., 36*(2), 229-251.

Weller, H.G., Repman, J., & Rooze, G.E. (1994). The relationship of learning, behavior, and cognitive styles in hypermedia-based instruction: Implications for design of HBI. Computers in the Schools, 10(3/4), 401-420.

White, R. W., Dumais, S. T., & Teevan, J. (2009). Characterizing the influence of domain expertise on web search behavior. *Proceedings of the Second ACM International Conference on Web Search and Data Mining,* Barcelona, Spain. 132-141.

White, R. W. & Kelly, D. (2006). A study on the effects of personalization and task information on implicit feedback performance. In *Proceedings of the 15th ACM international conference on Information and knowledge management* (pp. 297-306). Arlington, Virginia, USA.

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

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

Yang, X., Xiang, P., & Shi, Y. (2009). Finding User's interest blocks using significant implicit evidence for web browsing on small screen devices. *World Wide Web, 12*(2), 213-234.

VITA

Chang Liu

Education

2007-2012	Rutgers University, New Brunswick, NJ Ph.D. in Library and Information Science at School of Communication and Information Minor: Graduate Certificate in Cognitive Science, Rutgers Center for Cognitive Science
2005-2007	Peking University, Beijing, China. Master in Information Science at Department of Information Management
2001-2005	Peking University, Beijing, China. B.A. in Information Science at Department of Information Management B.E. in Economics at China Center for Economic Research.

Journal papers published during Ph.D. studies

Li, Y., & **Liu, C.** (2012). Information Science in the USA: A Review for the Full Papers Presented in the Annual Meeting of ASIST 2011. *Journal of the China Society for Scientific and Technical Information*. 2012, 31 (5), 452-469. (in Chinese)

Cole, M. J., Gwizdka, J., **Liu, C.**, Bierig, R., Belkin, N. J., & Zhang, X. (2011). Task and User Effects on Reading Patterns in Information Search. *Interacting with Computers*, 23(4), 346 - 362.

Qu, P., **Liu, C.**, & Lai, M. (2010). Chinese College Students' Web Querying Behaviors: A Case Study of Peking University. *Chinese Journal of Library and Information Science*, 3(4): 23-36.

Liu, C., Qu, P., & Li, L. (2009).A review of main topics in information behavior research. *Library and Information Service*, 53(2): 24-28. (in Chinese)