

**UNDERSTANDING, MODELING AND
SUPPORTING CROSS-DEVICE WEB SEARCH**

by

Shuguang Han

B.S., Wuhan University, China, 2008

M.S., Wuhan University, China, 2010

Submitted to the Graduate Faculty of
the School of Information Sciences in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH
SCHOOL OF INFORMATION SCIENCES

This dissertation was presented

by

Shuguang Han

It was defended on

January 21, 2018

and approved by

Daqing He, Ph.D., Professor, University of Pittsburgh

Peter Brusilovsky, Ph.D., Professor, University of Pittsburgh

Yu-Ru Lin, Ph.D., Assistant Professor, University of Pittsburgh

Eugene Agichtein, Ph.D., Associate Professor, Emory University

Dissertation Director: Daqing He, Ph.D., Professor, University of Pittsburgh

UNDERSTANDING, MODELING AND SUPPORTING CROSS-DEVICE WEB SEARCH

Shuguang Han, PhD

University of Pittsburgh, 2018

Recent studies have witnessed an increasing popularity of cross-device web search, in which users resume their previously-started search tasks from one device to later sessions on another. This novel search mode brings new user behaviors such as cross-device information transfer; however, they are rarely studied in recent research. Existing studies on this topic mainly focused on automatic cross-device search task extraction and/or task continuation prediction; whereas it lacks sufficient understanding of user behaviors and ways of supporting cross-device search tasks. Building an automated search support system requires proper models that can quantify user behaviors in the whole cross-device search process. This motivates me to focus on understanding, modeling and supporting cross-device search processes in this dissertation.

To understand the cross-device search process, I examine the main cross-device search topics, the major triggers, the information transfer approaches, and users behavioral patterns within each device and across multiple devices. These are obtained through an on-line survey and a lab-controlled user study with fine-grained user behavior logs. Then, I work on two quantitative models to automatically capture users' behavioral patterns. Both models assume that user behaviors are driven by hidden factors, and the identified behavioral patterns are either the hidden factors or a reflection of hidden factors. Following prior studies, I consider two types of hidden factors — search tactic (e.g., the tactic of information re-finding/finding would drive to click/skip previously-accessed documents) and user knowledge (e.g., knowing the knowledge within a document would drive users to skip the

document). Finally, to create a real-world cross-device search support use case, I design two supporting functions: one to assist information re-finding and the other to support information finding. The effectiveness of different support functions are further examined through both off-line and on-line experiments.

The dissertation has several contributions. First, this is the first comprehensive investigation of cross-device web search behaviors. Second, two novel computational models are proposed to automatically quantify cross-device search processes, which are rarely studied in existing researches. Third, I identify two important cross-device search support tasks and implement effective algorithms to support both of them, which can be beneficial for future studies on this topic.

Keywords: Cross-device Web Search; Cross-Device Search Process; Cross-device Search Support; Search Process Modeling; Information Re-finding; Information Exploration; Search Tactics; Search as Knowledge Learning

TABLE OF CONTENTS

PREFACE	xiv
1.0 INTRODUCTION	1
1.1 Problem Statement	2
1.2 Research Framework	3
1.3 Research Questions	6
1.3.1 Understanding Cross-Device Web Search	7
1.3.2 Modeling Cross-Device Web Search	7
1.3.3 Supporting Cross-Device Web Search	8
1.4 Scope Definition	8
1.4.1 Search Conditions	9
1.4.2 Search Users	10
1.4.3 Search Tasks	10
1.5 Overview of the Chapter Structure	11
1.6 Terminologies	12
2.0 RELATED WORK	15
2.1 Understanding Cross-Device & Cross-Session Web Searches	16
2.1.1 Causes of Cross-Device & Cross-Session Web Searches	16
2.1.2 Task Topics of Cross-Device & Cross-Session Web Search	18
2.1.3 Cross-Device & Cross-Session Web Search Processes	19
2.1.3.1 Information Transferring Across Multiple Sessions	20
2.1.3.2 Search Content Change Across Multiple Sessions	21
2.1.4 Summary	22

2.2	Modeling Cross-device & Cross-session Web Search Processes	23
2.2.1	Qualitative Information Search Process Models	23
2.2.2	Behavior-based Search Process Modeling	24
2.2.3	Content-based Search Process Modeling	25
2.2.4	Summary	27
2.3	Supporting Cross-Device & Cross-Session Web Searches	27
2.3.1	Automatic Cross-Device & Cross-Session Search Task Identification	28
2.3.2	Cross-Device & Cross-Session Search Support Tools	29
2.3.3	Cross-Device & Cross-Session Search Support Algorithms	30
2.3.3.1	Utilizing Behavior History for Search Support	31
2.3.3.2	Employing Fine-grained Search Interactions	32
2.3.3.3	Searching for Novelty and Redundancy	33
2.3.4	Summary	34
3.0	METHODOLOGY	35
3.1	Methodology Outline	36
3.2	Surveying Cross-Device Search Experience	37
3.3	Experiment Design for User Study I	38
3.3.1	Search Conditions	38
3.3.2	Task Design	39
3.3.3	Research System	39
3.3.4	Relevance Judgment	41
3.3.5	User Study Procedure	43
3.3.6	User Study Data	44
3.3.7	Summary	45
3.4	Simulation Experiment Design	45
3.5	On-line Evaluation Experiment Design (User Study II)	47
3.5.1	Search Tasks & Conditions	48
3.5.2	User Study Procedure	48
3.5.2.1	Overall Procedure	48
3.5.2.2	Pre-task and Post-task Questionnaire	50

3.5.2.3	In-task and Post-task Relevance Judgments	50
3.5.3	Evaluation of Search Performance	52
3.5.4	Summary	53
4.0	UNDERSTANDING CROSS-DEVICE WEB SEARCH	54
4.1	Cross-device Search Topics, Triggers and Information Transferring	54
4.2	Understanding Search Behavior Changes	56
4.2.1	Behavioral Patterns in Cross-Device Web Search	56
4.2.2	Cross-Device Information Re-finding Behaviors	58
4.2.2.1	Usefulness of Information in S1.	58
4.2.2.2	Cross-Session Information Re-finding	59
4.2.2.3	Temporal Analysis of Information Re-finding	60
4.2.2.4	Summary.	62
4.2.3	Understanding Search Strategy through Post-task Interviews	63
4.3	Understanding Search Content Change	64
4.3.1	Metrics Related to Search Content Change	64
4.3.2	Temporal Changes of Search Content Metrics	66
4.3.2.1	Overall Temporal Patterns.	67
4.3.2.2	Comparisons between M-D and D-D.	68
4.3.2.3	Forgetting Effects.	69
4.3.2.4	Direct Evidence from User Study II.	69
4.4	Answers to RQ1	70
5.0	MODELING CROSS-DEVICE WEB SEARCH PROCESS	73
5.1	Behavior-based Search Process Modeling	74
5.1.1	Modeling Search Process with HMM	74
5.1.2	Model Setup for HMM	76
5.1.3	Explaining HMM Outputs	78
5.1.4	Validating HMM Outputs	81
5.1.4.1	A Qualitative Analysis.	81
5.1.4.2	A Quantitative Analysis.	82
5.1.4.3	Summary.	85

5.2	Content-based Search Process Modeling	85
5.2.1	Modeling Search Process with Knowledge Learning	87
5.2.2	Understanding and Validating the KSP Model	90
5.2.2.1	Understanding User Knowledge in KSP.	92
5.2.2.2	Validating Information Repetition Hypothesis.	93
5.2.2.3	Validating Information Exploration Hypothesis.	94
5.2.3	Summary	96
5.3	Answer to RQ2	98
6.0	SUPPORTING CROSS-DEVICE WEB SEARCH	100
6.1	Designing Cross-device Search Support	100
6.1.1	Designing Cross-Device Search Support Functions	100
6.1.1.1	Function I: Supporting Information Re-finding.	101
6.1.1.2	Function II: Supporting Information Finding.	101
6.1.2	Building a Cross-Device Search Support System	102
6.1.3	Re-ranking Click History to Support Re-finding	105
6.1.4	Re-ranking Search Engine Result Pages	106
6.1.4.1	Detecting Search States.	107
6.1.4.2	SERP Re-ranking based on the KSP Model.	107
6.1.4.3	Connection with Existing Studies.	109
6.1.5	Evaluating Cross-Device Search Support	109
6.2	Off-line Evaluation for Cross-device Search Support	111
6.2.1	Off-line Evaluation for the Information Re-finding Support	111
6.2.1.1	Model Setup.	111
6.2.1.2	Experimental Results.	112
6.2.2	Off-line Evaluation for the Information Finding Support	113
6.2.2.1	Experiment Setup	113
6.2.2.2	Experimental Results	114
6.3	On-line Evaluation for Cross-Device Search Support	115
6.3.1	Experimental System and Baseline System	115
6.3.2	Data Collection	116

6.3.2.1	Overall Description	116
6.3.2.2	Clicks for Recall and Exploration	117
6.3.2.3	Participant-reported Document Utility Measures	118
6.3.3	User Perception on the Baseline and Experimental Systems	120
6.3.4	User Behaviors on Baseline and Experimental Systems	122
6.3.4.1	Information Re-finding Behaviors	122
6.3.4.2	Information Finding Behaviors	125
6.3.4.3	Summary and Discussion	126
6.4	Answers to RQ 3	128
7.0	DISCUSSION, CONCLUSION AND FUTURE WORK	129
7.1	Conclusion & Contributions	129
7.1.1	Conclusions	129
7.1.2	Contributions	130
7.2	Discussion of Results	132
7.2.1	Re-finding and Finding Behaviors in Cross-device Web Search	132
7.2.2	Knowledge Growth during Searching	133
7.2.3	Modeling Search Process with Hidden Variables	135
7.2.4	Cross-Device Search VS. Same-Device Cross-Session Search	136
7.3	Limitation	137
7.3.1	Representativeness of User Study	137
7.3.2	Over-simplified Model Assumptions	138
7.3.3	Increased User Load with Complex Interface	139
7.4	Future Work	139
7.4.1	Generalizing to Real-world Search Tasks	139
7.4.2	Improving Search Process Models	140
	APPENDIX A. MECHANICAL TURK QUESTIONNAIRE	142
	APPENDIX B. QUESTIONNAIRES FOR USER STUDIES	147
B.1	Entry Questionnaire	147
B.2	Training Task	148
B.3	Post-task Questionnaires for User Study I	149

B.3.1 Post-task Questionnaire after the 1st Session	149
B.3.2 Post-task Questionnaire after the 2nd Session	149
B.4 Pre-task and Post-task Questionnaires for User Study II	150
B.4.1 Pre-task Questionnaire before the 1st Session	150
B.4.2 Post-task Questionnaire after the 1st Session	150
B.4.3 Pre-task Questionnaire before the 2nd Session	150
B.4.4 Post-task Questionnaire after the 2nd Session	151
BIBLIOGRAPHY	152

LIST OF TABLES

1	A detailed description of six cross-session web search tasks.	40
2	Mean (S.D.) of several descriptive measures in cross-device search.	58
3	Mean (S.D.) of several information re-finding measures in cross-device search.	60
4	Search interactions adopted in this dissertation.	77
5	Spearman correlation between topical relevance and several other measures.	93
6	Mean (S.D.) of several measurements under different conditions.	94
7	Mean (S.D.) of several experimental metrics over different search conditions.	97
8	Repetition percentage for top three SERP results in different conditions.	102
9	Features to predict repeated click for any document c_i	106
10	P@3 performances of different approaches under different conditions.	113
11	Average ranking positions of clicked documents under different approaches.	114
12	Mean (S.D.) of several metrics in the first/second sessions for user Study II.	117
13	Mean (and S.D.) of participants' responses.	118
14	Pearson correlations among four relevance and novelty measures.	118
15	Mean (S.D.) of several measures under different document access purposes.	120
16	Mean (S.D.) of user perceptions (range: 1 - 7) on several measures.	120
17	Mean (S.D.) of user perceptions on information re-finding support.	121
18	Mean (and S.D.) of information re-finding measures in the continued sessions.	124
19	Mean (S.D.) of information exploration metrics for system effectiveness.	127

LIST OF FIGURES

1	Chapter organization for this dissertation.	12
2	An overall framework for my dissertation methodology.	36
3	The mobile (right) and desktop (left) search interfaces for Jing.	42
4	One example user experiment procedure for User Study I.	44
5	Tasks and search process models studied in the simulation experiment.	46
6	An illustration of the on-line evaluation experiment design.	47
7	Task and condition orders of User Study II.	49
8	One example user experiment procedure for User Study II.	50
9	The contextual relevance judgment interface for User Study II.	51
10	Cross-device search topic distribution based on Mechanic Turk responses.	55
11	Information reuse percentages (S.E.) for queries/query terms and clicks.	62
12	Temporal evolution of several content metrics in cross-session searches.	66
13	Knowledge growth pattern summarized from user study results.	67
14	Mean (S.E.) of participant-reported knowledge in four search stages.	70
15	Hidden variable model for search behavior (content) modeling.	74
16	An illustration of HMM-based search process model.	75
17	Experiment data (search behavior) and HMM parameter analysis.	78
18	HMM results for the continued session of M-D.	79
19	HMM results for the continued session of D-D.	80
20	Temporal distribution of different search strategies in continued sessions.	82
21	HMM results for the first session of M-D search.	83
22	Performances of predicting repeated clicks in M-D for different approaches.	85

23	HMM results for the first session of D-D search.	86
24	Performances of predicting repeated clicks in D-D for different approaches. . .	86
25	An illustration of the Knowledge-based Search Process (KSP) model.	87
26	The expected probability of document click $P(C)$ over user knowledge.	91
27	True click probability distribution (left), and its related factors (right).	95
28	An illustration of the prototype system for cross-device search support.	103
29	An illustration of search result reminder.	104
30	An illustration of the cross-device search support evaluation framework.	111
31	Temporal distribution of repeated/recall click percentage over time.	124
32	Knowledge growth and saturation pattern identified in our user studies.	134

PREFACE

I still remember the first summer I arrived in the States. It was my first time take flight and travel to a new country. I was so passionate about almost everything. I talked to people, learned language, took courses and participated in several great research projects. Thanks myself for being so passionate at the very beginning of my PhD study, which paved the way for my whole six-year doctorate study.

I was born in a poor family. For a child from rural area in China, living a good life might even be hard, let alone getting a doctoral degree from abroad. It is far beyond a dream. I never know how much belief my parents had in order to support me finish more than 20-years education. Thanks my parents for helping me around this long journey.

I am deeply grateful to my advisor, Professor Daqing He, for his kindness and help during the past six years. If it is not him, I might not even be able to come to the States and start my great journey to research. In the past six years, I have learned a lot from him. I still remember the weekly discussion on research projects, I always received many insightful suggestions. I also want to express my gratitude to my committee members, Professor Peter Brusilovsky, Professor Yuru Lin and Professor Eugene Agichtein, for their insightful suggestions and valuable guidance through my dissertation study. Thanks for Professor Rosta Farzan for her offering of one year's research assistantship, from which I have learned a lot about statistics.

Also thanks my Google and Yahoo! internship hosts David Huynh, Peng Dai, Praveen Paritosh, Dmitry Lagun, Vidhya Navalpakkam, Xing Yi, Zhen Yue and Alyssa Glass. Of course, the awesome internship host Dr. Sharon Hsiao from Arizona State University.

Finally, big thanks to my colleagues and lab mates from University of Pittsburgh. It is all of them who made my life at Pitt enjoyable and memorable. Thanks for Zhen Yue

(she helps too much and I never know how to pay it back), Jiepu Jiang, Wei Jeng, Evgeny Karandaev, Sharon Hsiao, Denis Parra, Yiling Lin, Xidao Wen, Rui Meng and Yu Chi. Particularly thanks for my big brothers Sung-Min Kim (now, I play better bowling than him) and Shih-Yi James Chien (now, I play better basketball than him).

Shuguang Han, 2017, Sunnyvale, CA

1.0 INTRODUCTION

With a wide adoption of mobile devices (e.g., tablets, smart phones, game consoles and etc.) in web search, modern search engines have started to research users' mobile search behaviors and investigate effective ways of supporting such behaviors [53, 73, 105, 127, 153, 142]. Mobile devices bring both convenience and challenges to web search. On the one hand, mobile devices break the time and location barriers — people can search for information at any time and place where their information needs are triggered. In the meantime, mobile devices can easily access people's search contexts such as the location information, which can be further applied to personalize users' mobile search results [135]. On the other hand, mobile devices, particularly the smartphones have many physical constraints — small screen and keyboard sizes, constrained mobile data usage, and limited cellphone battery. All of these limitations impede the continuity of users' web search processes on mobile devices. As a result, modern search engines have started to observe several novel search forms such as cross-session mobile web search [142] and cross-device web search [55, 105].

Specifically, through analyzing Microsoft Bing's web search logs, Montanez et al. [105], in the year of 2014, found that more than 5% of searchers are now multi-device users, and they are highly-active users and take account for about 16% of all search queries. Using similar search engine logs from Microsoft Bing in 2012, Wang et al. [142] found that 15% of device switches accompanied with task continuation. A recent study in 2014 by Facebook [41] also revealed the same phenomenon that we are entering a multi-device world, and around 40% of on-line adults had the experience of starting a task on one device but completing it on another. Those numbers are likely to grow significantly in recent years because of the increasing trend of mobile device adoption.

The emerging of cross-device web search has intrigued several recent studies to inves-

tigate the general cross-device web search patterns (e.g., topical, temporal and geo-spatial search patterns) [105, 142], the ways of automatic detecting web search tasks that will be continued in the near future [2, 142], and the methods to extract cross-session web search tasks from web-scale search query logs [139]. All of the above studies present some initial understanding of cross-device web search; however, there remains very few analysis regarding the complicated search behaviors such as cross-device information transfer in this new search condition. This is the main research topic for my dissertation.

1.1 PROBLEM STATEMENT

The complexity of a cross-device web search lies on the fact that it involves multiple search sessions, multiple search devices and interruptions. As a result, traditional research findings based on single query, single session and single device cannot truly reflect user behaviors in this new search scenario since people may develop different strategies.

Conventional information retrieval research mostly focused on satisfying an individual's single information need (i.e., *ad-hoc information retrieval*) [98], whereas recent studies have realized the importance of modeling and supporting information requests at the whole-session level [46, 68, 75, 95, 113]. Despite the development of related techniques in the past decade, the understanding and supporting of long-term, single-session searches remain a challenging topic, let alone when considering multiple search sessions. A promising direction of supporting a multi-session search, as studied in Luo et al. [95] for single-session search, is to investigate users' search strategies and differentiate search results based on different search strategies. This is one of my research topics in this dissertation.

Besides, when searching on multiple types of devices, people's search behaviors often change significantly due to the differences between mobile and desktop [127, 153] — comparing to search on desktop computers, mobile devices are less convenient for typing search queries and reading through webpage content. In addition, a relatively easy access of location context also enables search engines to incorporate such information for ranking better search results [103, 135]. All of these factors have led to the differences of search topic

distributions and search behavior patterns among different devices [127]. Because of the involvement of multiple devices, I expect to observe a different set of cross-device web search patterns comparing to the same-device cross-session search. To understand such difference, multiple search conditions are included and compared in our analysis.

As for the search interruption, researchers found that it can cause people to forget the task content they meant to complete [106], which further drives users to re-access the information they encountered in prior sessions for task resumption [1, 134, 137]. Indeed, previous studies on cross-device web uses (though some of them are not specifically targeting on web search) [36, 74, 109] identified that information sharing and synchronization across different devices are important user behaviors; however, due to the lack of sufficient understanding of cross-device web search behaviors, it is unclear what information should be transferred, when is the most appropriate time for information synchronization and how to effectively transfer such historical search information across multiple devices. Furthermore, even with a clear understanding of users' search processes, designing effective interfaces and algorithms to support cross-device search processes remain big challenges. This dissertation provided an extensive analysis for all of the above-mentioned problems.

Overall, the purpose of this dissertation is twofold. First, I will provide a more comprehensive understanding, through both qualitative and quantitative analyses, of cross-device web search processes, particularly how different factors (devices and interruptions) affect user behaviors. Second, I plan to design and develop proper quantitative search process models so that they can be incorporated for better cross-device search support. The following sections provide more details for each purpose.

1.2 RESEARCH FRAMEWORK

With regarding to the inadequate understanding of this new search condition, my dissertation started with inquiring the major cross-device web search topics, the important search strategies users adopted to transfer information across different devices, and the main triggers for users to perform a cross-device web search. This was obtained through distributing

an on-line questionnaire, asking respondents to report their cross-device search experiences. However, a simple survey cannot provide accurate data for investigating users' search behaviors across different stages, which motivated me to perform a lab-controlled user study with complete behavior logging functions. Although modern search engines recorded large-scale search transaction logs, an accurate extraction of cross-session web search tasks requires significant post-processing effort [72, 139]. More importantly, search log data is usually of high privacy, making it difficult to access necessary search content. Therefore, we follow prior studies on similar topics [50, 55, 90] and obtain search data through user studies. However, I acknowledge that a lab study might introduce artificial factors, since the search tasks are pre-assigned rather than originated from users.

With the collected search behaviors, I firstly investigated users' cross-device search behaviors in each separate device and the behavior changes in multiple devices. Among them, I primarily focused on examining how people behave in the continued search devices (e.g., how people synchronize information across devices, and how people advance their previously-started search progresses) since it is the most unique characteristic comparing to other search conditions. In addition to the behavior changes across the whole search process, prior studies [38, 90, 144] found that the content complexity for different search behaviors also changes (e.g., content of a query or a clicked document) — with searching goes on, people tend to retrieve more difficult and exploratory search content. Therefore, I further examined such pattern in my study. Overall, the first goal of this dissertation is to provide a comprehensive **understanding** of cross-device web search processes.

The purpose of understanding cross-device web search behaviors is to provide an appropriate search support. In this dissertation, I mainly focused on **supporting** web search for the continued session because it is the most unique difference from single-session search and the corresponding search behaviors are more complex. The main objective of providing such support is to facilitate people to better discover *useful* documents. Here, the document usefulness is highly contextual-dependent — a previously-accessed relevant document might be useful at the time when a user attempts to recall the forgotten information, whereas it is not useful if she has already obtained sufficient knowledge and wanted to explore more novel information. This example shows the complexity of designing an effective search sup-

port algorithm, which drives me to further study the automatic ways of identifying searchers' hidden intentions (e.g., recalling old information or exploring new information) or knowledge levels. These outputs were then incorporated into my search support algorithm for properly differentiating the returned search results.

To understand the effectiveness of a search support algorithm, I performed two different types of experiments — a *simulation experiment* and an *on-line experiment*. The simulation experiment treated users' future search behaviors (queries or clicks) as ground-truth so that the algorithms with better prediction performances of future behaviors are preferred. The on-line experiment examined search performance through a live comparison of user performances between the search systems with and without a search support algorithm, and I expect that the one with the search support algorithm should perform better.

Implementing the above search support algorithm requires a quantitative **modeling** of the whole cross-device search process. Since query and click are the two most important search behaviors, such model should effectively capture and forecast query and click patterns. Following prior studies on similar topics [8, 10, 54, 155], my dissertation explored two search process modeling approaches, both assumed that user behaviors are driven by latent factors.

- The first approach models users' search processes with both observed search actions and hidden search tactics. It assumes that user behaviors are driven by a hidden search tactic. A search tactic refers to a high-level plan a user designed for facilitating her search processes [8]. For example, a user, in a continued search session, may sometimes plan to restore their task contexts (because of forgetting) so that she is likely to re-access previously-accessed information. At the other time, she may want to explore in breadth; thus, search a general topic and click several relevant or marginal relevant documents to cover multiple topics. To automatically discover these hidden search tactics, I adopted the unsupervised approach proposed in [54, 155], and further adapted it into my search condition. Their approach employed the standard Hidden Markov Model (HMM) for automatically bridging search tactics (modeled as the hidden variables in HMM) with observed search actions. It assumes a Markov process on search tactics and all search behaviors are generated by them. One drawback for this approach is that it only considers different types of search behaviors while ignoring the search content (e.g., the content of

a search query or a clicked document) associated with each behavior. This motivates me to consider the second approach.

- The second approach models a user’s search course as the knowledge learning process. Learned from Belkin’s ASK (*Anomalous State of Knowledge*) model [9] model, this approach views information seeking as a process of consuming search content and gaining new knowledge to answer the user’s information need. If there remains a mismatch between user knowledge and the needed knowledge for answering the information need, the user will keep searching till gaining the remaining knowledge. As a result, a user’s knowledge state is the main hidden factor that drives the adoption of certain search behaviors (e.g., clicking a document, issuing a query). However, the ways of modeling user knowledge and the mechanisms of driving user behaviors based on user knowledge remain a challenging topic, which were extensively explored in my dissertation.

To summarize, this dissertation focused on the **understanding, modeling** and **supporting** of cross-device web searches. The understanding of cross-device web search emphasized on analyzing the temporal change of users’ search behaviors and search content. Then, I investigated the ways of modeling and supporting cross-device web search processes. Here, the modeling and supporting primarily focused on the continued session because search behaviors in this stage are usually more difficult and may require more effort to understand people’s search contexts and search stages [40, 55].

1.3 RESEARCH QUESTIONS

Based on the above discussions of research problems and research framework, this dissertation specifically studied a set of research questions from the following three aspects. Each aspect can be further divided into several small sub-questions. The detailed descriptions of these research questions were presented in the above Section 1.2; here, I provide a brief summary of all of them.

1.3.1 Understanding Cross-Device Web Search

My first research question (**RQ1**) attempted to acquire a deep understanding about how people search in a cross-device web search process, how do they search on each separate device and across multiple devices? This can be further divided into the following three sub-questions.

- RQ 1.1: what are the major cross-device web search topics, how do people transfer information across devices, and what are the common triggers of cross-device web searches?
- RQ 1.2: what search behaviors (and search tactics) do users adopt in different cross-device web search stages, and how do they differ from those in same-device cross-session search tasks?
- RQ 1.3: what content do users search and explore in different cross-device web search stages, and what are the differences between cross-device and same-device cross-session web searches?

1.3.2 Modeling Cross-Device Web Search

As mentioned in Section 1.2, to better quantify users' cross-device search processes, I explored two different search process modeling approaches — a behavior-based modeling approach that captures users' search tactics and search behaviors within one unified mechanism, and a content-based modeling approach that models the change of search content based on the knowledge learning mechanism. The quantitative modeling outputs from both of these two search process models will be applied for supporting a cross-device web search process. As a result, the second research question (**RQ2**) tried to understand the effectiveness of behavior-based and content-based search process models.

- RQ 2.1 (behavior-based search process modeling): can a search process model be built to automatically identify users' cross-device search tactics? What are the differences between cross-device web search and same-device cross-session web search when modeling their search processes?

- RQ 2.2 (content-based search process modeling): can a search process model be built to automatically quantify searchers' knowledge change (search content change)? What are the differences between cross-device web search and same-device cross-session web search when modeling their search processes?

1.3.3 Supporting Cross-Device Web Search

The ultimate goal of understanding and modeling cross-device web search processes is to provide an appropriate cross-device search support, which is the third research question (**RQ3**) studied in this dissertation. RQ3 can be further decomposed into the following three sub-questions. Note that my dissertation only focused on the support of search queries in the **second session** and data from the first session was used to feed data for building search support algorithms. RQ 3.1 - RQ 3.2 focused on the off-line development of cross-device search support algorithms, while RQ 3.3 examined the efficiency, effectiveness and user perceptions of search support algorithms through on-line experiments.

- RQ 3.1: Can the behavior-based search process modeling approach be used to support cross-device web searches?
- RQ 3.2: Can the content-based search process modeling approach be used to support cross-device web searches?
- RQ 3.3: Can cross-device search support algorithms based on different search process models be effectively applied for supporting real-world search tasks? How do real-world users perceive the usefulness of different search support functions?

1.4 SCOPE DEFINITION

With the continuing development of modern search engines, web search can happen at any time, any place and with any form. Although I only focused on one specific search format (i.e., the cross-device web search), there remains full of variabilities — the search can happen across two, three or even more devices, and it can also happen for any kind of search tasks.

Therefore, I limited the generalizability of my conclusions from the following three aspects, including search conditions, search users and search tasks.

1.4.1 Search Conditions

Cross-device web search can happen across more than two sessions and/or two devices; however, I only take into account the case of two search devices and two search sessions based on the following two considerations. First, including more than two devices will significantly increase the difficulty of data acquisition since there lacks a publicly-available dataset; therefore, user study would be the only way for acquiring proper data collections. Conducting a user study with multiple devices and multiple sessions can easily increase task load and cause user fatigue. Second, two-device setting is the most commonly-adopted search condition in prior studies [105, 142]. Additionally, two-device setting can be viewed as the basic unit of any cross-device search condition, and thus the understanding of two-device setting can benefit any other condition. I do acknowledge that it is a very interesting topic to study cross-device web search with three or more devices, and I think the methodology from this dissertation can be easily repeated to a new search condition. However, a further validation is needed to reach a more confirmed conclusion.

My dissertation chooses desktop and smartphones for studying the cross-device web search. This is based on the following reasons: (1) desktop and smartphones are the two most popular web search devices; (2) desktop and smartphones represent two typical types of search devices - one with relatively large screen and keyboard sizes and the other with small screen and keyboard, one with low mobility the other is easy to be used at any place. Other mobile devices such as game console and tablet [105] may be of less representativeness. Mobile search queries on game consoles may be of significant difference from the search queries on other devices. The sizes of tablets are in between desktop computers and smartphones, in which users' search behaviors might lie between desktop and smartphones. However, I do believe that other devices such as tablet and game consoles are also interesting and should be properly studied in the future.

1.4.2 Search Users

The target users of my study are individual searchers with general cross-device web search experiences. However, due to the difficulty of obtaining demographic information of these users, I adopt the convenient sampling approach by recruiting native-English speakers and young generation (most of them are in the age of 20 - 30) with a background of college education. On the other hand, this sample selection approach also takes into account user familiarity with smartphones — younger generation tends to be more comfortable to use smartphones for typing, searching and browsing. In addition, the user study design in my dissertation is also consistent with several of previous studies which focused on similar or the same search topics [50, 55, 155]. I expect differences of user behaviors when generalizing the conclusions to non-native speakers or younger/older generation, which requires additional user studies and analysis.

1.4.3 Search Tasks

Studying cross-device web search with user studies requires designing proper search tasks. This dissertation focuses on the exploratory search tasks that involve long search sessions and multiple search queries. There are three main motivations for us to study this type of search tasks. First, cross-device web search tasks are often more complicated than simple navigational or fact-based information needs, which makes it appropriate to study the exploratory tasks in this scenario. Second, since exploratory search tasks usually involve relatively long-time interactions, they provide us with opportunities to accurately model users' search contexts using abundant search histories and user interactions. Third, modern search engines have provided reasonable support for individual search query, but it is still a challenge to support exploratory information needs with multiple search queries [100].

Lab-controlled user studies often received criticisms because of their adopted tasks are hard to be generalized. One potential remedy is to design representative search tasks. To handle this generalizability problem, my dissertation developed search tasks covering several of the most popular cross-device web search topics, and the task content was directly employed from user reports. Although with the above efforts to remove task selection bias,

each adopted search task also has its own specificity and score limitation. As a result, the conclusions reached in this dissertation might not be appropriate for other types of search topics that do not share any common characteristic with my designed search tasks.

1.5 OVERVIEW OF THE CHAPTER STRUCTURE

Figure 1 provides an illustration of the overall chapter organization for my dissertation. Chapter 2 enumerated a list of related studies of this topic, including the research of analyzing, modeling and supporting cross-device web searches. Due to the relative novelty of the cross-device web search, recent research on this topic has just started. Therefore, when surveying related studies, I also included the research work for cross-session web searches, which are not necessarily to be searched with different devices.

Chapter 3 presented a detailed description of the methodology that was used for designing a cross-device user experiment, obtaining data collection, examining the effectiveness of different search process models and experimenting the efficiency and effectiveness of cross-device search support algorithms.

Building on top of the data collections obtained from Chapter 3, the follow-up three chapters (i.e., Chapter 4, Chapter 5 and Chapter 6) focused on each of the three research questions (understanding, modeling and supporting cross-device web search process) as mentioned in Section 1.3. Specifically, Chapter 4 provided an extensive analysis of search behavior and search content changes across different stages of a cross-device search process. Based on this understanding, the next Chapter 5 targeted to build two automated and quantitative search process models for capturing important behavioral and content search patterns. Then, Chapter 6 experimented the methods of applying quantitative search process models for supporting cross-device web search with both off-line (simulation) and on-line experiments.

Finally, Chapter 7 provided a detailed summarization of my findings in the above chapters. Then, an extensive and comprehensive discussion of the obtained experimental results and their potential implications are offered. At last, I envisioned a set of future research directions to advance the research on the cross-device web search.

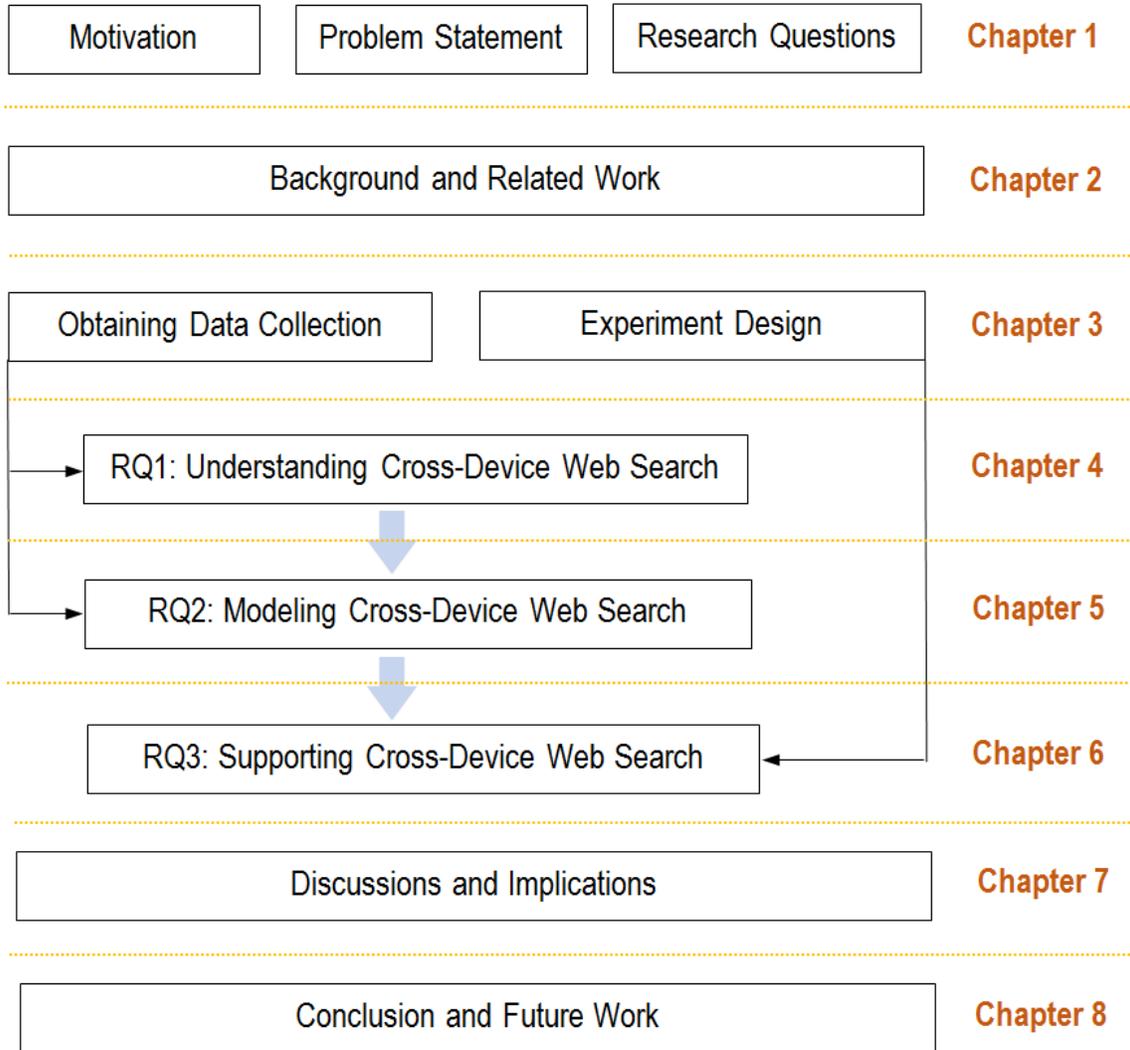


Figure 1: Chapter organization for this dissertation.

1.6 TERMINOLOGIES

The following terminologies are frequently used in my dissertation. Here provides formal definitions for all of them.

Bayesian Knowledge Tracing (BKT) is a widely-adopted algorithm in the computer education domain for modeling and tracing students' mastery of knowledge through continuous learning and practicing. BKT represents students' knowledge with a set of binary

variable with each denoting one skill. Each binary variable is associated with one probability indicating the likelihood of a student masters the corresponding skill. The most common implementation of BKT utilized the hidden variable in Hidden Markov Model to represent students' mastery of knowledge. A more detailed description of this model can be found in Corbett and Anderson [32] and Yudelson et al. [154].

Click-through Documents refer to the documents users clicked from search engine result pages or landing pages.

Context-sensitive Information Retrieval models have been widely adopted in recent researches for better capturing users' search contexts and further locating relevant documents. Search contexts can refer to many factors such as location, time, search histories and etc. [111]. My dissertation follows the line of research work that infers users' search contexts through their historical queries and click histories [55, 68, 124]. Using this setting is based on the consideration of data availability.

Cross-Device Web Search is one type of cross-session web search where the search devices on different sessions also vary. In this dissertation, I only consider two types of devices - desktop computer (D) and smart phone (M). The mobile-to-desktop search is referred as M-D and the desktop-to-desktop search is referred as D-D.

Exploratory Search refers to a search type where users spend long time for exploring the result space, learning new knowledge and investigating tasks faced [100].

Cross-Session Web Search refers to a search scenario where a user searches for the same tasks but could not complete in one single search session so that the tasks spans two or more search sessions. This dissertation only studies the two-session cross-session web search.

Hidden Markov Model (HMM) is a machine learning algorithm that often applies for modeling a temporal event. HMM assumes a Markov process on the hidden states (meaning that the current hidden state only depends on one previous hidden state while has no connections to other prior states) and each temporal event is only driven by the current hidden state. HMM can model temporal events that go beyond the simple observational level; therefore, it has been used for interpreting users' tactics in web search [125, 155].

Dwell Time is referred as the total time a user spends on a click-through document. Long dwell time with more than 30 seconds on a document often indicates the document is

relevant and the corresponding user is satisfied with that document [56, 53].

MTI is an abbreviation for Mobile Touch Interaction. This dissertation takes into account several commonly-used touch-based mobile interactions such as tap, double tap, drag up, drag down, drag left, drag right, pinch in and pinch out.

KSP is an abbreviation for the Knowledge learning based Search Process Model I proposed in Section 5.2. KSP model captures users' search content change in the continued sessions through quantified user knowledge and the learning curve. The underlying rationale of the KSP model is learned from both the BKT model in computer education domain (as mentioned above) and the ASK model in information retrieval community [9].

SERP is an abbreviation of Search Engine Result Page, which usually corresponds to a list of documents in the decreasing order of relevance.

A **Search Tactic** is a move a user made to further his search process [8]. For example, a user may sometimes explore information in breadth so that he/she shall search a general topic and click several relevant or marginal relevant documents to cover multiple topics; while the user may prefer to search deeply for a small topic at other time.

2.0 RELATED WORK

With the primary focuses of understanding, modeling and supporting cross-device web searches, relevant studies of this dissertation are divided into following three aspects. Since the study of cross-device web search has just started, when summarizing related research, I also included the studies of other similar search conditions such as cross-session web search, mobile-based web search, long-time single session search, and so on.

- The understanding of cross-device and cross-session web searches, which mainly includes topics about what causes cross-session and cross-device web searches (§2.1.1), what are the major search topics (§2.1.2), how do people search on cross-session and cross-device web searches, and particularly, how do people transfer information across multiple search sessions and/or devices (§2.1.3).
- The modeling of cross-device and cross-session web search processes, including surveying widely-adopted theoretical information seeking models (§2.2.1), the modeling of both search behaviors (§2.2.2) and their associated search content (§2.2.3). Here, the search conditions other than cross-device and cross-session searches are also considered since the proposed techniques can be easily applied in my study.
- The supporting of cross-device and cross-session web searches, in which I surveyed several of existing tools (§2.3.2) and their supporting techniques (§2.3.1). However, these tools only focused on providing partial search support (e.g., information transfer) at the interface level, whereas an algorithmic-level support for cross-device web search is missing. Considering related algorithms developed in other similar search conditions can also be beneficial for my study, I surveyed the algorithmic support for exploratory search tasks in Section 2.3.3.

2.1 UNDERSTANDING CROSS-DEVICE & CROSS-SESSION WEB SEARCHES

Modern search engines have started to shift their focuses from supporting simple information needs to facilitating complex and exploratory web search tasks [100, 113, 139]. A complex information need usually requires a long-time search exploration, in which users may search in one continuous session or may span multiple sessions. The latter is called a cross-session web search in literature [2, 81]. Particularly, when the search devices in different sessions are also different, it is referred as a cross-device web search [105, 142]. Cross-device and cross-session web searches are also studied as multiple-device and multiple-session web searches in literature [16, 89, 97, 129]. This dissertation uses these terminologies exchangeably. Existing research of understanding cross-session and cross-device web search processes can be categorized into the following three aspects.

2.1.1 Causes of Cross-Device & Cross-Session Web Searches

Based on a diary study, Sellen et al. [121] found that 40% of people’s *information gathering* tasks cannot be completed in one single session, which mainly due to **Interruption**. Indeed, people get interrupted frequently in their daily lives. Mark et al. [101] showed that normal people are interrupted in every 11 minutes and it often requires long time to recover from an interruption [65] — a process called task resumption [119] or multitask continuum [120]. Interruptions happened at different time can have different effects. Iqbal et al. [65] found that if an interruption happened at the time when people switched between two sub-tasks, it has less negative impact on task accomplishment. To minimize task restoration efforts and to better support task resumption, Teevan et al. [136] proposed to break down a large task into several small easy-to-achieve subtasks. However, the implementation of her idea is difficult because of the difficulty of extracting sub-tasks and the complexity of accurate predictions of the start and end of a sub-task.

Interruption might be caused by many factors, among which *Physical Constraint* (e.g., time limit and device constraint such as power limit and device usability) is one of the most

important factors. For example, a student may search for assignment information during a course, and the class is off before she can obtain necessary information. Therefore, she may continue searching for the same information once arriving home. In another scenario, a user is searching on her smartphone but the cellphone is power-off. Later on, she may continue her search once the phone is charged. Both of them are typical examples regarding how physical constraints can cause cross-session web searches. As a result, despite the flexibility the mobile devices bring, it still has significant limitations — mobile devices can only provide restricted input with small screen sizes and keyboard sizes, and have the issue of being expensive on mobile data usage. Once encountering complex web search tasks, either of these factors will increase the probability of halting a task and resuming it in later stages, i.e., performing a cross-device web search [55, 105, 142].

Besides the physical constraints, *Social Interruption* is another important factor. The development of modern techniques have provided normal people with various of instant messaging tools and social media applications for facilitating their communication, but these tools also create disruptions [34], which can break the continuous of an undertaking task. For example, an incoming Facebook chat message or an incoming phone call can easily interrupt users' search processes. *Personal Interruption* is the third important factor that causes cross-session and cross-device web uses. For instance, a user may intentionally switch to an unrelated task such as examining Twitter and Facebook before completing the current task [78]. The intentional interruptions are highly connected with people's cognitive states, and are particularly hard to predict. It is worth noting that interruptions are not always bad things. Dai et al. [35] and Rzeszotarski et al. [116] found that inserting micro-breaks for lengthy (more than hours) crowdsourcing tasks can reduce human fatigue effects and improve crowdsourcing quality. However, this type of search tasks is usually quite rare, which is thus not taken into account in this dissertation.

In addition to interruption, **Forgetting** is another reason that leads to cross-session and cross-device web searches. Human have cognitive limits — they cannot memorize every piece of information they explored in a lengthy search process, particularly when the search tasks are difficult and exploratory. According to Morris et al. [106], many users (63.2%) have to adopt tools to assist their memorization for task resumption. Without a proper

support, people often need to re-issue the same or slightly different search queries [134] to re-access previous explored information [137], which inevitably leads to a cross-session or cross-device web search. Note that interruption and forgetting are not mutually exclusive. Human forgetting always exists in any type of search conditions, and it is particularly obvious when people experience a long-time search interruption [7].

2.1.2 Task Topics of Cross-Device & Cross-Session Web Search

Not all search tasks are equally like to happen in a cross-session web search. Simple tasks such as navigational information needs that can be easily accomplished with one or two search queries usually do not span across different search sessions.

To understand the main search topics for cross-session web searches, MacKay and Watters [97] conducted two studies — a diary study asking participants to fill out a diary every time before they started a search task, and a field study requiring participants to install a customized web browser to record their search interactions. Based on the data obtained from both two studies (22 participants for the diary study and 24 participants for the field study), MacKay and Watters identified that the top three information needs are school work (28.51%), general topic (28.09%) and research (10.64%), which are not surprising because of the major participants are students, school faculties and staffs. After mapping these information needs into another categorization scheme, they found that the top three task types are information gathering task (53.27%), fact-finding task (27.80%) and transactional task (10.75%). The information gathering task refers to those tasks that need to collect a large amount of information in order to compare, choose or decide about something. The dominating percentage of this type of task indicates the complexity of cross-session and cross-device web search tasks.

Based on a large-scale analysis of search logs from Microsoft Bing, Agichtein et al. [2] analyzed the connections between task characteristics (e.g., task intent, task complexity, task topical categories) and task continuation. They found that the “information maintenance” (keep monitoring of a running topic — such as tracking the news of airplane crash) tasks have the largest probability of continuing, followed by the “transactional” and “communication”

while the “fact-finding” and “information gathering” task have the least probabilities of continuing. These results are contradictory from MacKay and Watters [97]. This might come from the following two reasons. First, these two studies reported different statistics. MacKay and Watters [97] provided task type percentages given the task is a cross-session search task while Agichtein et al. [2] computed the probability regarding that given a task type, how likely this task will be continued to search in the future (some of them may not). Second, user samples in these two studies are far from each other. MacKay and Watters [97] only surveyed 46 participants and most of them are school workers or staffs, whereas Agichtein et al. [2] sampled 1,000 users from Microsoft Bing’s search logs.

In terms of the task topical category, Agichtein et al. [2] identified that search tasks related to adult, kids and teens, news and games are more likely to be searched in cross-session web search conditions while sports, science and health tasks are the least likely cross-session topics. Since this study only took into account the cross-session web searches that happened within the same device, a later study by Wang et al. [142] further investigated the task topical distribution for cross-device web search. They found that *Image* and *Navigational* search tasks are more likely to be resumed, partially because of the overall task popularity. On the other hand, *Books*, *Celebrity* and *Music* are the high-likely resumed search topics without overall task popularity.

2.1.3 Cross-Device & Cross-Session Web Search Processes

The above two sections did not examine the studies that looked into user behaviors during the whole cross-device or cross-session search processes, which is the focus of this section. Although prior literature has provided a fruitful research outputs regarding to the understanding of exploratory web search processes [82, 99, 100, 146, 145] and long-time session search processes [46, 66, 68, 69, 70, 95, 75], they seldom considered the search scenarios with multiple search sessions or devices. The most unique feature of a cross-session web search is that the later session is the continuous of the previous one, in which users may have already mastered partial knowledge so that they do not need to search from scratch in the continued search session [55]. Observing the importance of information transfer across multiple

search sessions/devices, among the limited number of studies on this topic, most focused on analyzing the information transfer behaviors and strategies.

2.1.3.1 Information Transferring Across Multiple Sessions Through a university-based user study, Teevan et al. [134] found that people often forgot their original search queries even within a short interruption time like one hour. Without a proper support for transferring historical search information, users in a cross-session search process often need to issue duplicated queries or revisit the same webpages [55, 106, 137] in order to restore their previous task progresses. This was studied as information re-finding in literature [133, 137]. Teevan et al. [134] found that 40% of Yahoo!'s overall search queries are re-finding queries. Another study [45] reported a even higher percentage. Although cross-session task resumption often accompanies with information re-finding, the re-finding behaviors happen not only across multiple search sessions but also within a single session [21, 69, 117, 134]. Recent studies have investigated the differences between cross-session and within-session information re-finding [108, 137]. Tyler and Teevan [137] discovered that the within-session re-finding tends to be re-evaluating search results while the cross-session re-finding is for task resumption. In addition, they found that the cross-session re-finding is more likely to happen at the beginning of a search session than at the end. A similar result, based on a lab-controlled user study with several properly-designed cross-device web search tasks, was reported in a recent research study conducted by Han et al. [55].

In addition to issuing duplicated queries and revisiting historical webpages, when resuming search tasks, people also frequently employed assistant tools such as bookmarking, and emails to store information temporarily for later uses [55]. To further understand people's information transferring strategies, Morris et al. [106] distributed an on-line survey asking respondents to fill their information transferring strategies across multiple search sessions. They found that 55% of the respondents took the active storage & active retrieval strategies (writing notes - 30%, bookmarking - 27.8%, saving pages - 5.3%, email to oneself - 2.4% and etc.) and 8.2% of them took the passive storage & active retrieval strategies (using browser history - 7.6% or query history - 1.2%). However, there remains a large amount of users (49%) took the passive storage & passive retrieval strategies (36% of the respondents

try to memorize the information, 14% left the browser open for later use and 9% employed other strategies). Similarly, using the same questionnaire, Kane et al. [74] surveyed 175 IT employees about their information transferring strategies in cross-device web search tasks. They found that people frequently transfer information across different devices but usually adopted manual methods such as emailing to themselves or directly writing a note (65.1%). Note that there is an important difference about the reported percentages between Morris et al. [106] and Kane et al. [74] — the strategy of “email to oneself” has only 2.4% in Morris et al.’s study whereas it has the largest amount (65.1%) in Kane et al.’s study. This might be either because of the difference of the study scope (i.e., Kane et al. [74] only focused on cross-device web searches while Morris et al. [106] did not take specific consideration of search devices) or due to the differences of participants (Kane et al. [74] only surveyed IT employees while Morris et al. [106] studied a wider range of users including software developers, managers, attorneys, students and etc.). Overall, both studies suggested the importance of developing proper assistants for better supporting of information transfer.

2.1.3.2 Search Content Change Across Multiple Sessions Real-world cross-session and cross-device search processes, particularly in the continued sessions, are way more complex than only information re-finding. With a consistent searching and learning, people may behave differently across multiple search sessions. Recent studies indeed observed that users’ search content, including both query and click, changed significantly during the whole cross-session search courses [5, 38, 147]. Eickhoff et al. [38] found that with search goes on, users tend to issue more complex search queries and explore more diverse content. Many researchers attribute such change to the growth of users’ knowledge and domain expertise [90, 144, 147, 158]. White et al. [144] found that with different levels of domain knowledge (e.g., domain experts and novices), people tend to exhibit different search behaviors. Wildemuth [147] analyzed how users’ search strategies are related to domain expertise. She found that people tend to adopt different search strategies with the growth of their knowledge levels. Zhang et al. [158] found that several implicit behavioral measures such as average length for search queries and average SERP (Search Engine Result Page) ranking positions of the clicked documents can be used to predict users’ overall domain knowledge. It is worth

noting that knowledge change is observed not only in a long-term search process (e.g., over an academic semester [147] and cross multiple search sessions [90]) but also in a short-term within-session search process [38].

2.1.4 Summary

The above literature review provides a comprehensive survey about the *causes*, *task topics* and *search processes* related to cross-device & cross-session web searches. They are used as guidelines to design my later user experiments in order to simulate real-world search tasks. In designing user experiments, I simulate a multi-task environment so that searching for one task can be thought as the interruption of another. In addition, search tasks are developed based on popular cross-device and cross-session search topics [97]. When analyzing the search process, both information transferring approaches and search content change patterns are analyzed in corresponding to the major subsection topics surveyed in this section.

Although having similar research topics as previous studies, my dissertation also significantly differs from them in the following aspects. First, the majority of previous researches either focused on investigating search behaviors within a single session or targeted on analyzing search processes for the same-device cross-session web searches, whereas my primary focus is on cross-device search behaviors. Second, existing studies only provided a coarse-grained understanding of the information transferring behaviors. With additional goals on automatically modeling and supporting search behaviors in the continued sessions, a more fine-grained comprehension of cross-device search is needed. Specifically, I will analyze how people search on each separate session (device), and across multiple devices, the difference between searching in the starting and the continued sessions, the search tactics users adopted, the type of search content users applied in different search stages, and etc. Here, I expect to identify important and fine-grained behavioral patterns to guide the automatic modeling of cross-device and cross-session web search processes.

2.2 MODELING CROSS-DEVICE & CROSS-SESSION WEB SEARCH PROCESSES

Because of the relative novelty of cross-session and cross-device web searches, so far, there are very few related studies working on modeling such search processes. As a result, this section summarized relevant studies of search process modeling from a more generalized topic that includes all types of long-session [11, 46, 68, 87, 95] and exploratory web search tasks [100, 145]. Different from the conventional information retrieval research that only concentrates on fulfilling individual search queries, exploratory and long-time session search tasks usually require users issuing multiple queries and performing long-time explorations. The complexity of these tasks often trigger users to adopt different search strategies, making it difficult to truly understand and properly model the change of users' information needs during the whole search courses [70]. The following sections listed a few of existing search process models, which I categorized into three groups.

2.2.1 Qualitative Information Search Process Models

Besides the development of automated search result ranking algorithms [98, 58], information retrieval researchers also proposed a set of theoretical (with qualitative constructs) models for characterizing human's information seeking processes [9, 39, 82, 99]. In 1991, Kuhlthau [82] designed six stages, including initiation, selection, exploration, formulation, collection and search closure, to characterize normal users' information seeking processes. Marchionini [99] proposed an information seeking process model with eight stages (recognize accept, define problem, select source, formulate query, execute query, examine results, extract information and reflect/stop), and different stages can transit among each other with different probabilities. Belkin, in 1980, proposed ASK (*Anomalous State of Knowledge*) model [9] to capture users' search content and knowledge changes during their search processes. ASK model views information retrieval as a process of gaining new knowledge to resolve the anomaly between users' current knowledge level and the knowledge required for the undertaking search tasks. As long as there remains the knowledge mismatch, users will keep searching and exploring

for gaining the needed knowledge.

These theoretical information seeking process (ISP) models provided explicit descriptions about how people frame their information needs and carry out their search tasks. All of them were widely cited by later researches of information seeking processes in library and information science. They were also extended to multiple variants when studying information seeking processes in other domains [59, 122, 155], and used as design guidelines when building information retrieval systems [83]. However, search stages defined in the above three ISP models are subjective and are often manually crafted by domain experts, which makes it difficult to develop machine learning algorithms to automatically map search stages with specific search behaviors. As a result, these ISP models are even harder to be incorporated with search support algorithms for facilitating cross-device search processes.

2.2.2 Behavior-based Search Process Modeling

To overcome the difficulty of mapping search stages with recorded user behaviors, researchers took an alternative data-driven approach and attempted to automatically infer search stages from system-recorded user behaviors. For instance, based on the recorded user behaviors, Holscher and Strube [60] examined and compared different search stages and search strategies between Internet experts and newbies, Chen and Cooper [25] identified several of usage patterns in a web-based library catalog with a stochastic model. However, these studies simply treated the aggregation of several system-logged actions as a search stage or a search strategy without further considering the underlying search intentions (i.e., search tactics or search strategies as defined in [8]). Recent study by Xie and Joo [150] raised the importance of using search tactic as a means to examine search processes, and they manually categorized 13 different search tactics, including learning, exploring, examining search results, modifying search items, monitoring search progresses, preserving relevant information and etc. Again, since most of these search tactics are manually defined and categorized by domain experts, it is highly unscalable, and cannot be generalized to other tasks or domains.

To fix this gap, Han et al. [54] and Yue et al. [155, 156] proposed to employing the Hidden Markov Model (HMM) for automatic search tactic identification, in which search tactics were

modeled as hidden states and search behaviors (e.g., query, click and etc.) were generated by these search tactics. Since HMM is a well-established unsupervised machine learning approach, it can be easily applied in any large-scale dataset. Note that the HMM-based search process model is similar to the Probability Ranking Principle (PRP) in interactive information retrieval developed by Fuhr [43]. The PRP model assumed that users move between situations (a situation can be thought as a hidden state in HMM) in a search process. In each situation, users have a list of choices, and each choice (e.g., a user search action) has its cost, acceptance probability (transition probability in HMM), and benefit. Except the modeling of cost and benefits, HMM is consistent with this theoretical framework. It is also worth noting that the HMM-based approach has also been successfully applied for inferring users' search contexts [20] and predicting future user behaviors [140] in many web-scale search logs. Meanwhile, HMM was also applied for analyzing and interpreting various of sequential behavioral data such as the eye-tracking data from web search tasks [125] and the supervisory control experiment data from human robot interaction system [15, 27]. These studies all provide positive evidences to demonstrate the validity of utilizing HMM for modeling sequential web search behaviors.

However, there are two open challenges for HMM-based search process modeling — the identified hidden states are usually hard to interpret [54, 155], and it is unclear that how can the identified search tactics be effectively applied for search support. The first challenge was referred as the *interpretability* issue in a recent study by He et al. [57], which further drove the authors to study efficient and human-supervised (and thus highly interpretable) ways of mapping observed user actions with search tactics. Though potentially solve the interpretable issue, it also introduces human effort which would affect the algorithm scalability. Regarding to the second challenge, a set of off-line and on-line experiments in §5.1 were designed to experiment the most effective ways of HMM-based cross-device search support.

2.2.3 Content-based Search Process Modeling

Behavior-based search process models usually did not make a fine-grained analysis of the search content, which is another important and indispensable component in a search process.

As mentioned, prior studies [5, 38, 147] have identified that user-generated search content, including the issued queries and the clicked documents, changes significantly with search goes on. That is mainly due to users' knowledge learning during search [114]. However, there is only a limited number of studies that worked on modeling content change and knowledge learning during a long-time human search process [9, 10, 46, 95].

Belkin's ASK model [9, 10] is one such model. It views information retrieval as a knowledge learning process which attempts to resolve the mismatch between the learned user knowledge and required task knowledge. However, ASK only focused on providing qualitative constructs for describing a long-session search process, making it difficult to be incorporated with any search support algorithm. One major reason for being unable to produce quantitative outputs might be related to the difficulty of quantifying user knowledge. Recent studies by Guan et al. [46] and Luo et al. [95] took an alternative approach by directly capturing the change of query terms and returned search results instead of explicitly modeling user knowledge. Specifically, they proposed to use Markov Decision Process (MDP) to model the long-session complex search processes, which both assumed two agents in a web search process — a user agent whose main actions are changing queries (adding terms, removing terms and keeping terms) and a search engine agent that responds to users' query changes by adjusting search techniques and query term weights. Although having the ability of bridging multiple search actions (query, click, examining SERP) into one unified framework, MDP-based search process models did not provide a fine-grained analysis for the content of clicked documents, and failed to explicitly capture users' reading behaviors and knowledge learning processes. As a result, these two models are not directly linked to various of prior studies on knowledge-based information seeking process.

To resolve the problems encountered in both ASK and MDP-models, in this dissertation, I attempted to build a quantitative knowledge learning based search process model. Particularly, my proposed model is built on top of Belkin's ASK model with further consideration of representing and tracing users' knowledge change. To properly quantify user knowledge and its learning process, I borrowed the theories and practices developed in psychology and education domains [32, 44, 152] for modeling knowledge acquisition and growing processes. To be specific, I adopted a widely-used student knowledge modeling approach called *Bayesian*

knowledge tracing (BKT) model for capturing students' knowledge learning process [32]. Follow-up studies [110, 154] have also attempted to advance this model through personalizing different learning parameters on different students. A desirable feature of the BKT is that it explicitly modeled human's knowledge states and traced the change of these states. This can be easily integrated with Belkin's ASK model [9] for information searching.

2.2.4 Summary

This section surveyed a list of widely-cited qualitative and quantitative search process models for long-session exploratory search tasks. The qualitative models focused on building holistic views for human information seeking processes; however, these models often ignore the implementation details and are hard to find corresponding application scenarios. The quantitative models adopted data-driven approaches to automatically discover, simulate and learn users' search patterns. These models can decompose a long search process into several computation steps so that they can be integrated with automated search support algorithms.

The search process models developed in my dissertation were built on top of several past successes; however, they also differ from prior studies in two aspects. First, almost all of the existing search process models were designed and explained in the context of single-session search, whereas my study particularly emphasized on the modeling of continued sessions in a cross-session and/or cross-device search process. Previous studies [55, 137] have already recognized the complexity of user behaviors in this stage so that I expect to discover a different search process model. Second, one of my developed search process models (see the KSP model in Section 5.2) took into account both qualitative constructs and computation steps, which attempted to bridge the gap between qualitative and quantitative models.

2.3 SUPPORTING CROSS-DEVICE & CROSS-SESSION WEB SEARCHES

The ultimate goal of understanding and modeling cross-device and cross-session web searches is to support user search. Related studies can be summarized from the following aspects.

2.3.1 Automatic Cross-Device & Cross-Session Search Task Identification

A comprehensive study of cross-device and cross-session search support should start with correctly identifying a different set of cross-session search queries that belong to the same task. Although there has been a number of studies [42, 53, 72, 86, 94] focusing on extracting tasks from large-scale search engine transaction logs, they mostly focused on task extraction within one single search session (30 minutes inactivity as timeout cutoff for defining a session). Related studies on identifying cross-session and cross-device web search tasks are rare.

To the best of our knowledge, the first comprehensive research came from Kotov et al. [81] in 2011. They examined two different cross-session task identification problems. First, the same-task query identification problem: given a query, identify all of the prior related queries (belong to the same task) issued by the given user. Second, the task continuation prediction problem: given a multiple-query search task, predict whether the user will continue searching for this task in the future. Both tasks were treated as classification problems and were examined with datasets extracted from Microsoft Bing’s search logs. Agichtein et al. [2] extended Kotov’s work by considering four additional feature groups in classification: search topic features, user engagement features (e.g., satisfaction and dissatisfaction of the search engine results), search history features and topic/query repetition features (e.g., how likely a query/topic will be reoccur in the dataset). The follow-up experiments demonstrated the usefulness of taking into account these additional features. Besides identifying cross-session web search tasks, Wang et al. [142] and Montanez et al. [105] are interested in predicting task continuation for cross-device web search. They found that geo-spatial and temporal features are particularly important in this setting and should be included.

The above-mentioned two task identification problems [81] took different views (backward view or forward view) in a cross-session web search. An effective cross-session support should take advantages of both views instead of applying them separately. Therefore, later on, Wang et al. [139] went beyond these two defined problems, and developed a semi-supervised query clustering model based on latent structural SVM to extract cross-session search tasks in one unified algorithm.

Although the above studies have provided several automatic task extraction algorithms

for high-accuracy identification of cross-session search tasks [2, 81, 139, 142, 105], it remains unclear that how the extracted tasks can be applied for better modeling of the cross-sessions search processes, and more importantly, how the extracted tasks can be used to develop better cross-session search support algorithms and/or interfaces. Identifying cross-session search tasks is only the first step. Incorporating such information into different search algorithms and support user search process are the ultimate goals, which is discussed in the below two sections.

2.3.2 Cross-Device & Cross-Session Search Support Tools

One important characteristic of cross-device and cross-session web searches is that the later sessions are the continuous of the previous sessions. In a continued search session, people usually need to restore their previous task completion progresses and problem states [6, 119]. When encountering highly complex search tasks, people may not be able to memorize everything they searched and explored. A proper tool that can assist human’s memorization and task resumption can be useful. The simplest design of this tool is to display users’ browsing and searching histories so that it enables people to quickly re-find or recall the information they examined.

Session Heights [67] and Auto Web [88] are two such toolkits to assist users’ page re-visitation experiences. Instead of providing a list of previously-visited web pages using page URLs or summarized snippets, Session Heights [67] provided the thumbnail of each visited web page and displayed them in a chronological order. In this way, it can help people to quickly skim through the page content or even page display styles. However, when prior sessions involve too many web pages, such display might be inefficient since each web page occupies too much space. Auto Web [88] provided a simple solution by categorizing users’ search histories into different groups based on document content and only displayed few representative web pages in each group. Then, the automatic selection of representative web pages became another important issue since mis-categorizing a useful web page would take even more user effort for result re-finding.

Besides the information re-visitation behaviors, Morris et al. [106] found that people also

frequently adopted the note-taking and result bookmarking functions for task resumption. Therefore, they developed SearchBar, a system that provided both note-taking and content bookmarking functions. In addition, the system also enabled users to preserve and rate the importance of a web page. Meanwhile, SearchBar also stored and displayed users' rich interaction histories including issued queries, clicked web pages and query-click relations. SearchPad [12] is a similar toolkit that provided alike functions. Though these systems could potentially assist task resumption for complex web search tasks, the usability of these systems has become an important issue — introducing too many user interactions may potentially confuse users and decrease the likelihood of using some functions.

The above-mentioned systems search support systems mostly worked on desktop-to-desktop search condition, whereas the information transfer and task restoration across multiple devices were not studied. In recent years, with more and more observations about task migration from desktop to mobile devices, Chang and Li [23] developed Deepshot, a tool that can transfer tasks from desktop to mobile, through taking photos from a desktop website using phone camera. It facilitated task transfer but may only work for simple tasks – a cross-session web search task usually involves multiple queries and web pages, which is not easy to be captured through a single taken photo.

2.3.3 Cross-Device & Cross-Session Search Support Algorithms

Despite the increasing attention of cross-session and cross-device web searches, research on developing automatic cross-session search support algorithms remains very few. Here, an algorithmic search support refers to developing an algorithm that can facilitate users better locate useful documents, i.e., ranking useful documents at top positions so that users can quickly access of these documents. In a cross-device web search, a later search session is the continuation of the prior one; thus, it is necessary to transfer useful search history information across different sessions. Although not targeting on cross-device web search, previous studies on context-sensitive information retrieval [124], personalized web search [135], and session search [46, 68, 95] could be helpful to transfer historical search information and develop appropriate cross-device search support algorithms.

2.3.3.1 Utilizing Behavior History for Search Support This section provides detailed descriptions for the above-mentioned search support algorithms, and particularly on how they use historical search information. Specifically, Shen et al. [124] proposed, based on the framework of model-based feedback [157], a context-sensitive information retrieval model using implicit feedbacks (including both queries and clicks) for search support. However, this model only utilized the short-term search history, which motivated a later study to focus on examining the effectiveness of long-term search history information [132]. The short-term and long-term search histories were also adopted in many personalized web search algorithms [143, 128]. Bennett et al. [11] compared the differences between these two types of search histories and explored methods to leverage both of them. The TREC session search tasks targeted on supporting the search at the session level rather than at a single query level. The algorithm proposed by Jiang et al. [68] achieved the best performance in TREC session search track 2011. The model was built upon the context-sensitive retrieval model by Shen et al. [124]. The session search algorithm proposed by Guan et al. [46] further captured both of the query changes with adjacent queries and its relationships with previously-clicked documents. Later on, Luo et al. [95] developed a more complex Markov Decision Process (MDP) based model that includes query changes, clicks and search engine results into one unified framework.

When utilizing search histories for search support, the historical search queries and the historical clicked documents are two commonly-adopted resources. Studies [68, 124, 132] found that both of them are important relevant feedback information but the latter is more effective. Previous studies developed two different ways of utilizing click-through documents. The first method is to employ the full text of a click-through document [118]. However, there might contain too much noise such as ads or copyright information in the full content of a web page. In addition, a web page may contain diverse topics while people may only concentrate on part of them. Therefore, a proper algorithm that can identify relevant information at the sub-document level is necessary.

The second method went deeper into the sub-document level to locate relevant content chunks. For example, Liu and Croft [91] estimated relevant content based on the best-matched passage in a document. Lv and Zhai [96] proposed a Positional Relevance Model

(PRM) to weight the words in a document that are closer to the occurrence positions of the query words. Shen et al. [124] employed the document snippet from SERPs as the representation of a click-through document. All of these approaches were demonstrated, through extensive experiments in corresponding studies, to be more effective than simply applying the full-text of the click-through documents.

2.3.3.2 Employing Fine-grained Search Interactions The above-mentioned algorithms did not take into account the fine-grained behavioral information recorded in search engines, which were proved to be effective in many information retrieval tasks [47, 48, 85, 71]. Buscher et al. [19] found that the display time of a passage (inferred from the scrolling behaviors on desktop computers) is an important implicit feedback for extracting sub-document relevant information. Such information can be further applied as fine-grained relevance feedback to improve search ranking. Lagun et al. [85] discovered that the result display time is aligned with users' eye-gaze attention. Thus, the display time can be properly leveraged to measure user satisfaction on the results that do not require user clicks. Speaking of the eye-gaze attention, Buscher et al. [18] experimented the ways of extracting sub-document relevant information from users' eye movements. They found that, as expected, such information can help identify more accurate relevant information. However, this approach involves tremendous external resources and thus may be inapplicable in live search systems.

It is worth noting that most of the above studies were performed on desktop computers, whereas certain user behaviors might be unavailable or hard to be collected from mobile. For example, there are no mouse-based scrolling behaviors on mobile; however, mobile devices are often support with touch interactions. The Mobile Touch Interactions (MTI) have been demonstrated to be useful in various search tasks [48, 55, 50]. Guo et al. [48] found that MTIs are important indicators of document relevance and user satisfaction. They identified that users' inactive time has a positive correlation with document relevance, whereas gesture speed has a negative correlation. Han et al. [52] demonstrated that the page content corresponding with slow-speed touch gestures and long inactive time are good indicators of user interest. However, it is still challenging to determine the corresponding content for each MTI because it depends on both the information layout and the users behavioral habits. Biedert et al.

[13] found that, in general, most touch movements lie in the middle of users reading zones. Therefore, Han et al. [55] assumed that users read the content closest to their touch positions, which narrowed down the inference of relevant information at the sub-document level. More specifically, Han et al. [55] developed a similar cross-device search support algorithm as Buscher et al. [19], in which they treated the dragging up/down movements as the scrolling behaviors on desktop computers and the inactive time as the display time. MTIs were used to accurately infer users' read content than using the full text. Experimental results demonstrated the effectiveness of MTI-based relevance feedback at the sub-document level.

However, mobile behaviors on clicked documents may not provide sufficient search history information to infer a robust relevance feedback model. This is because the limited screen and keyboard sizes of mobile devices prevent users from fully exploring click-through documents. As an alternative, researchers found that the interactions on the SERP can provide additional information. For example, Huang et al. [61, 62] found that mouse cursor activities on SERPs align with users' eye movements, and those activities can be further used to infer document relevance. Han et al. [55] also demonstrated the effectiveness of utilizing MTIs on the SERPs rather than the ones on clicked documents.

2.3.3.3 Searching for Novelty and Redundancy The above-mentioned studies only focused on returning users with the most relevant documents. This search paradigm worked well in the case of ad-hoc information retrieval but might be inappropriate for a long-session web search. According to several of recent studies [29, 68, 69, 70, 159], users' judgments of document usefulness changed with time goes on — a relevant document might be less useful if the searcher has obtained most of the knowledge regarding the document, whereas a marginal relevant but novel document could be useful if the user's purpose is to explore new information. Since the major goal of my dissertation is to support cross-device web search at the later stage, this becomes an extremely important issue. The detection of document novelty and document redundancy could be relatively easy based on several of existing approaches [4, 159]; however, assessing the time when a user needs to access redundant documents or novel documents is difficult since it is related to human search strategies, the obtained knowledge and search skills [69, 70, 151]. In the case of cross-session and cross-device web

searches, this becomes a more complex problem since users may forget information after interruption [55, 137]. My dissertation also attempted to provide an initial understanding about the timing when users wanted to access redundant and/or novel documents through lab-controlled user experiment.

2.3.4 Summary

This section provided a summary of existing cross-session and cross-device search support techniques, including the automatic extraction of search tasks from large-scale search transaction logs, both the interface-level and the algorithm-level support of search tasks. Instead of algorithmically extracting search tasks from behavior logs, my dissertation collects such data from user studies for the following reasons. First, there are no publicly-available search log datasets because of privacy issues. Second, existing cross-session task extraction algorithms might not be easily generalizable to cross-device search scenarios because of device difference. In addition, the model might not produce a high extraction accuracy so that applying such model for future analysis would be problematic.

Following the previous studies, both interface-level and algorithm-level search support functions are included in my dissertation; however, the design and implementation rationale of my search support functions also largely differ from existing studies in terms of: 1) rather than directly displaying search history with the chronological order as did in prior studies [88, 67], my approach provides an intelligent ranking of search history based on the likelihood of being revisited; 2) when ranking or re-ranking search results, my search support system not only considers document relevance (as did in the majority of prior studies [18, 19, 50, 55, 68, 124, 132]), but also takes into account novelty based on users' current knowledge states.

3.0 METHODOLOGY

To answer the above-mentioned three research questions, this chapter describes the methodology to achieve these answers. RQ1 and RQ2 aim to examine cross-device web search behaviors and further understand the underlying mechanisms that drive these behaviors. This requires a complete set of user behavior logs. However, publicly-available search logs either lack the functionality of recording complex, fine-grained user behaviors of our interests or are having high privacy concerns. As a result, we collect user behaviors through user studies using Jing¹, a self-built research system with properly-designed user behavior logging functions. In addition, user studies also require developing proper web search tasks, recruiting a reasonable amount of participants, designing appropriate experimental procedures, and etc. All of them are discussed in this chapter.

RQ3 attempts to develop a cross-device search support algorithm. To examine the algorithm effectiveness, the most intuitive way is to perform on-line experiments with A/B testing. However, such evaluation has a high cost so that a viable approach is to train algorithms in off-line, and then select the best one for on-line evaluation. My dissertation also adopts this two-step approach with the first step focusing on comparing multiple search support algorithms, and the second step on the on-line evaluation. To enable an off-line comparisons of multiple search algorithms in the first step, I follow the Cranfield experiment paradigm [30] to build a test collection with a set of queries and users' relevance judgments. This is obtained through lab-controlled user studies and post-task questionnaire asking for assessing document relevance. Then, a set of search support algorithms and their parameters will be trained and examined based on this test collection, and the fine-tuned ones will be applied for on-line search support with another user study.

¹<http://crystal.exp.sis.pitt.edu/Jing/>

3.1 METHODOLOGY OUTLINE

There are four types of experiments in my dissertation, each targeting to answer one or two aspects of my research questions. An overall framework of my dissertation methodology can be illustrated in Figure 2.

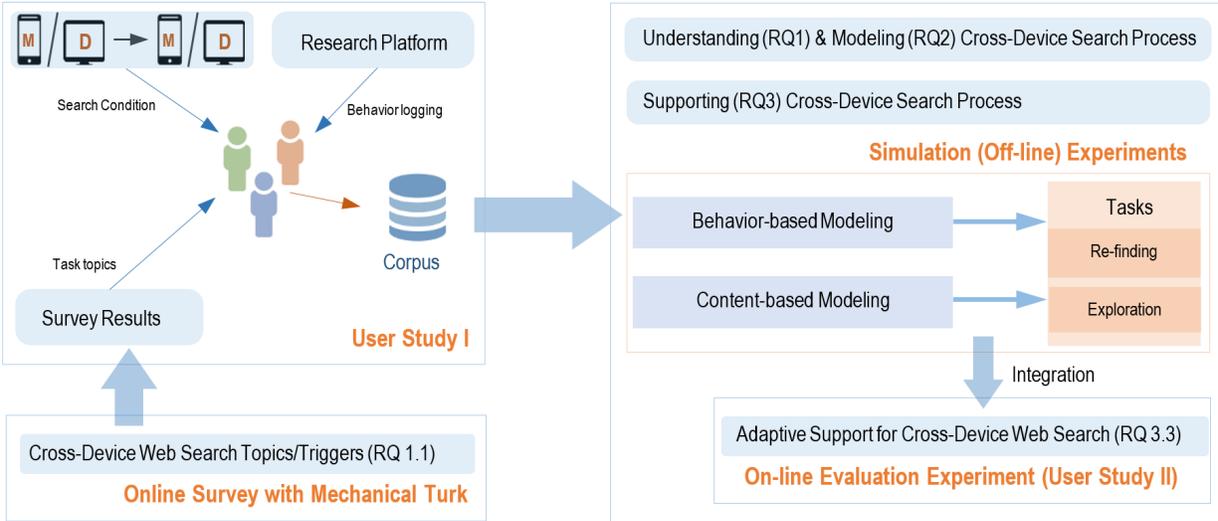


Figure 2: An overall framework for my dissertation methodology.

The **on-line mechanical turk survey** tries to answer RQ 1.1 through asking respondents to recall their cross-device web search experiences. The obtained survey results can provide an overall understanding about the cross-device search patterns. In addition, the collected search topics are the basis for the future design of cross-device search tasks in my later user studies. To answer the rest research questions of RQ1 and RQ2, I need a data collection with recorded search behaviors from both cross-device and same-device cross-session search conditions. This is gathered through a well-designed user study (named as **User Study I**). With the collected datasets, I then analyze user behaviors and search content changes, and further simulate such change with different quantitative search process models. The obtained data collection is also used to study the cross-device web search support algorithms through a set of **simulation experiments**, in which the data collection is divided

into different groups (time-based cutoff) to simulate cross-device search support experiments — a better support algorithm should better predict users’ future behaviors. Finally, I integrate the above developed search support algorithms into one unified framework, and further examine its performance with real-world users through a **summative on-line evaluation experiment** (named as **User Study II**). The details of each experiment, particularly the simulation (off-line) experiment and the on-line evaluation experiment will be explained in more details in the following several sections.

3.2 SURVEYING CROSS-DEVICE SEARCH EXPERIENCE

In order to obtain a comprehensive understanding of the cross-device search process, I would like to know what are the main search topics, how do people transfer information across multiple devices, what triggers people to perform a cross-device web search, and etc. This is collected through online questionnaire. The main method of the online survey is to ask the participants to recall their recent cross-device search experiences, in which they started a search task on their mobile devices but the search was not completed due to various reasons such as interruptions, complexity in search or continuous interest. Later, they finished the search of that topic on a desktop device. The questions in the survey concentrate on the information needs, the triggers to their information needs, the methods for preserving and transferring information from mobile to desktop, and other details of their information needs. I designed 12 questions including ten multi-choice questions and two short answer questions. The question about cross-device web search topics contains a list of information need categories, which were designed based on previous studies on cross-device web search [81, 142] with slight modifications. The participants can select all the topics that they have past experience. The short answer questions asked the participants to describe one of their cross-device search experiences in details by answering a short question with five sub-questions. The participants needed to describe their search needs, triggers of the information need, search strategies/queries used on mobile devices, method of preserving information and search strategies/queries used in the continued session on desktops.

The survey was conducted online through Amazon Mechanical Turk² in October 2013. A copy of the 12 survey questions are provided in Appendix A. For quality assurance, I designed a simple screening question (i.e., question 1 in Appendix A) at the beginning of the survey, asking respondents’ cross-device web search experience. Only if they answered yes, the rest 11 questions will be displayed; otherwise, the user’s IP address were blocked from answering the follow-up questions. In addition, I also manually examined all the responses to double check the validity of the survey results. Each turker who completed the survey with reasonable quality was compensated with one dollar. For the ones who were blocked from answering my questions did not receive any compensation.

3.3 EXPERIMENT DESIGN FOR USER STUDY I

To understand and model cross-device web search processes, I need to obtain one such data collection at first. This is achieved through a lab-controlled user study (**User Study I**), which is described in the following sections.

3.3.1 Search Conditions

I include both cross-device and same-device cross-session search conditions for better understand and compare user behaviors in different search scenarios. The same-device cross-session search is used as a baseline. This dissertation also considers two types of search devices — mobile devices (denoting as **M**) such as smartphones, and desktop devices (denoting as **D**). As a result, a complete investigation of cross-device web search behaviors should consider four conditions: Mobile-to-Desktop (**M-D**), Desktop-to-Mobile (**D-M**), Mobile-to-Mobile (**M-M**), and Desktop-to-Desktop (**D-D**). The first two belong to the cross-device search condition and the last two are the same-device cross-session searches. This dissertation only takes into account two search conditions — **M-D** and **D-D** for three reasons. First, M-D search is an important cross-device search type, which acquired relatively few understanding

²<http://www.mturk.com/>

from previous studies [55, 142]. Second, including all four search conditions will significantly increase the difficulty of conducting a user study since a participant might feel too tired with all conditions. Third, people may conduct fewer user behaviors on mobile due to the difficulty of using such device; therefore, an accurate inference of user intentions is more challenging and needed for M-D.

3.3.2 Task Design

A user study requires designing a set of cross-device web search tasks to resemble people’s real-world information needs, and also simulate the triggers of these tasks. The on-line survey identified that the searching for product, people, video and news are four major topics of information needs on cross-device web search. Due to the difficulty of developing experimental systems supporting video search (because of the difficulty of recording user behaviors during video playing), I select product, news and people as three types of user study. Specifically, I design six different tasks under three categories — two news search tasks, two product search tasks and two people search tasks. The detailed descriptions of each task can be found in Table 1. I acknowledge that these tasks are assigned rather than initiated by the participants themselves; thus including artificial factors. However, user study is still a valid and common approach when studying interactive information retrieval systems [14]. In addition, all six search tasks were designed to resemble the intrinsic diversity task [113] which are identified as an important and common type of complex web search tasks.

3.3.3 Research System

User studies in this dissertation are performed using Jing³, a cross-device web search system designed and implemented by myself. The system provides search support functions for cross-device web search tasks, and equips with a full functionality of user behavior tracking and analysis. Jing returns search results from Google mobile if the users are searching on mobile devices, or otherwise from general Google search. Screen captures of Jing on desktop and mobile devices are shown in Figure 3. After typing a query, Jing will automatically send

³<http://crystal.exp.sis.pitt.edu/Jing/>

Table 1: A detailed description of six cross-session web search tasks.

Product Search Task 1 (PD1)	Product Search Task 2 (PD2)
<p>You were in an Apple store and saw the latest iPad mini, this triggers you to find more information about different tablets. Example needs include what are the available tablets besides iPad? How many difference types and sizes are there for tablets? What are the differences among iPad, android-based tablets, windows tablets and other tablets? What are the good and bad reviews for them?</p>	<p>You heard that your friends were talking about interchangeable-lens cameras. You are interested in finding more information. How is it different from the point-and-shoot camera and the single-lens reflex (SLR) camera? What are the manufacturers for interchangeable lens cameras? What are good and bad reviews such cameras?</p>
People Search Task 1 (PE1)	People Search Task 2 (PE2)
<p>You read a novel The Mysteries of Pittsburgh on a bus. You are interested in finding more information about the author Michael Chabon and his works. What is the background of his first novel? What are the reviews of the books he have written? Are there any films based on his books and the reviews? Who are the people that impact his life? What are his most recent activities?</p>	<p>In an economic class, your teacher introducing an American economist Herbert A. Simon. You are interested in finding more information about him and his works. What are the important contributions of him? In what fields did he contribute? In what schools did he teach? What courses did he teach? Who are the people that impact his life? Who are his collaborators and what kind of work did they collaborate on?</p>
News Search Task 1 (NE1)	News Search Task 2 (NE2)
<p>You were involved in a class discussion about Americas automotive industry. You are interested in finding out news in the past few years about the car company GM. What are the news that related to the employment, earnings and stock of GM? What is the news related to the decline of the automotive industry? What is the news that related to the recovery of GM from the economic recession? (The news are not necessary to be recent).</p>	<p>You were involved in a class discussion about America’s hi-tech internet industry. You are interested in finding out news in the past few years about Facebook. What are the news that related to the employment, earnings and stock of Facebook? What are the news that related to the emergency and quick-expansion of the social media industry? What is the news that related to the down side of the company? (The news is not necessary to be the latest).</p>

a search request, along with the search device used, to Google search API and display the corresponding Google search results. Note that the returned search results and search result displays are different when using different search devices. There are two major functions for Jing. First, Jing records all of user-generated search behaviors (query and click) and their associated content. Each search behavior is recorded with a timestamp. Second, to better support users’ information preservation behaviors during the complex web search tasks, I provide a function for users to save relevant web pages on both devices. Mobile users can trigger a pop-up saving confirmation window by holding their fingers on the page. Similarly, desktop users can trigger the saving function through holding the left button of their mouse on the page. Jing provides a workspace to store and display the saved web pages (with titles and URLs) for each search task. A user can click on the links of saved pages to revisit.

3.3.4 Relevance Judgment

To understand the usefulness of each document, I also collect participants’ relevance judgments each time after they complete their search tasks, i.e., the time they finished up the second search session [30]. In user study I, I only focus on the topical relevance of a document. Specifically, after completing each task in my user study, a participant is asked to rate the overall relevance of a document on a 5-point Likert scale. Since one document might be saved by multiple participants with different relevance scores, I adopt a pooling-based approach when computing the final document relevance. The simplest way is to average relevance scores from multiple users. However, the average might be inaccurate if one document is only rated by a small number of participants. Therefore, we adopt a Bayesian smoothing approach [51] to eliminate this bias. The method can be formalized as Equation 3.1, in which the smoothed average relevance score $\hat{r}'(d)$ of document d is computed based on the interpolation between document average score $\hat{r}(d)$ and the overall average C (average score over all documents) as shown in Formula 3.1. v denotes the number of participants who saved the document d .

$$\hat{r}'(d) = \frac{\hat{r}(d)v + C}{v + 1} \tag{3.1}$$

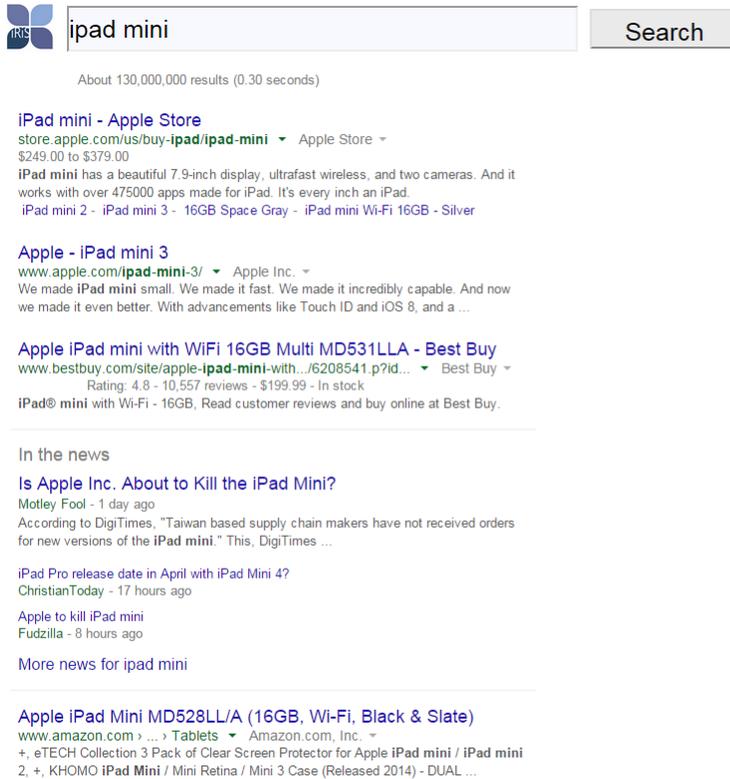


Figure 3: The mobile (right) and desktop (left) search interfaces for Jing.

3.3.5 User Study Procedure

User study I was conducted in the year of 2013, which began with an introduction to the study, and then followed by asking participants to fill out a background questionnaire about their demographic information (age, gender, major, general web search experience, and cross-device web search experiences, see Appendix B.1). Next, the participants were given a 5-minute tutorial and 15-minute training on using system Jing. In the training session, the participants worked on finding the news about *the causes and consequences of U.S. government shutdown in 2013*⁴. They were asked to explore this topic firstly on mobile devices and then continue their searches on desktop computers. This way helps users to get familiar with system interfaces and functionality on both devices. Considering that users might be unfamiliarity with different mobile operating systems, we provided both Android device (HTC Phone) and iOS device (iPod with iOS 7).

User study I employed a within-subject design [76], in which each participant worked on all six tasks, and each task was split into two sessions with eight minutes in the first session and seven minutes in the second. The extra one minute in the first session was designed for the participants to read task description. Therefore, the search time in both sessions was approximately equal. The orders of the task and device combinations were rotated based on the Latin Square. Each round of rotation consists of 12 participants. Figure 4 shows one of the 12 possible task and device combination orders. The task and device orders are rotated for other examples.

Similar to the illustration of Figure 4, each participant had three tasks started on mobile and the other three tasks started on desktop as the first session. In the second session, all six tasks were resumed on desktops and the task order was the same as in the first session to align the interruption time between two sessions. Therefore, there was an interruption of about 1 hour between the first and second session for each task. In total, three tasks were completed on mobile-to-desktop (M-D) and the other three tasks were completed on desktop-to-desktop (D-D) for each participant. In the example of Figure 4, the PD1, PE2, and NE1 were completed in M-D and PD2, PE2, and NE2 were completed in D-D.

⁴During the time I conducted the user study, U.S. government shutdown is the latest news. A detailed description of the training task can be found in Appendix B.2.

During the user study, participants were instructed to save the relevant web pages. After participants finished each search session (both the first and the second search session), they were asked to answer post-task questionnaires about their search experiences (see Appendix B.3.1 and Appendix B.3.2 for details). In addition to that, after the completion of the second search session, they were further requested to rate the relevance of each saved document using a five-point Likert scale with 1 denoting non-relevant and 5 representing highly relevant. This document relevance judgments will be applied to examine the utility of different cross-device search support algorithms.

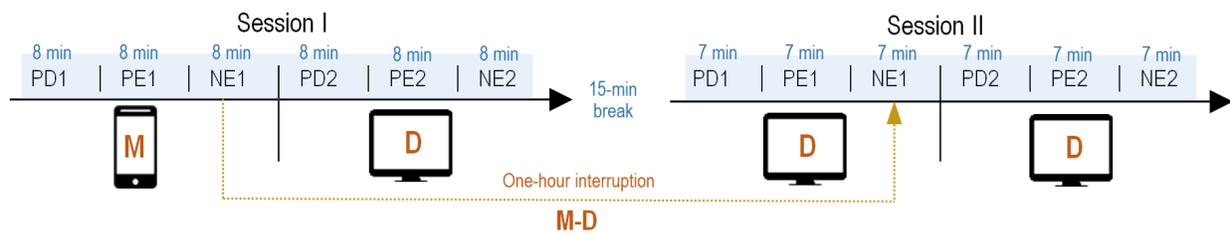


Figure 4: One example user experiment procedure for User Study I.

User Study I also included a post-task interview after each participant completed all of the six search tasks and before she left the experiment, in which I asked three open-ended questions: 1) what strategies did you adopt when searching in different sessions; 2) what strategies did you apply when searching in different devices; and 3) what were the main strategies for you to transfer information across different sessions. On average, each participant answered these questions around five minutes.

3.3.6 User Study Data

I recruited 24 participants from the University of Pittsburgh and Carnegie Mellon University by promoting my user study via flyers (15 females and 9 males; 16 undergraduates and 8 graduates). Their ages are between 18 and 30 with a median of 21 years old. All participants reported a 4-or-higher search experience level on a 7-point Likert scale (1 is the least skilled

and 7 is the most skilled). The participants also reported their experiences of cross-device web search in the background questionnaire. On average, 18% of their web searches conducted on mobile devices had follow-up desktop searches, which is similar to the results from the study on Microsoft Bings search logs.

In total, the recruited participants issued 1,419 queries (794 from the first session and 625 from the second), clicked 3,186 web pages (1,736 from the first session and 1,450 from the second) and saved 2,108 of them (1,191 from the first session and 917 from the second). Since my goal is to support user search in the second session, queries from the second search session are treated as the **test queries**.

3.3.7 Summary

To summarize, this section provides a complete description about the experiment design, search task design, research system, user study procedure, participant recruitment and the collected dataset for **User Study I**. The obtained data collection will be used for understanding (§4) and modeling (§5) users' cross-device and cross-session web search processes, and later, for examining the effectiveness of different cross-device search support algorithms through off-line simulations (§6.2).

3.4 SIMULATION EXPERIMENT DESIGN

Data collection from user study I is firstly applied to analyze cross-device and cross-session web search behaviors, and then used to validate our proposed search support algorithms through simulation experiments. Since the major goal of this dissertation is to assist user search in late stages of a cross-device search (i.e., the second session), the simulation experiments treat user behaviors in the second session as unknown and then attempt to forecast these behaviors. In my later analysis from Chapter 5.1, we discover two major types of search behaviors in the second session — the information re-finding behavior and the information exploration behavior, based on which I design two simulation experiments, as shown

in Figure 5, to support each of them.

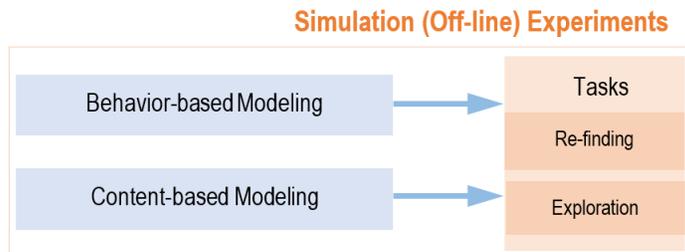


Figure 5: Tasks and search process models studied in the simulation experiment.

The first step to build an automatic search support algorithm is to *quantify* user behaviors and their intentions. This dissertation examines two quantitative approaches — behavior-based modeling and content-based modeling. A behavior-based model attempts to understand users’ search process based on recorded interactions. Instead of direct studying raw user actions, behavior models delve into the underlying intentions behind user behaviors. The discovered user intentions are further applied to personalize users’ search results. For example, if a user’s intention is to explore novel information, diversifying search results and demoting repeated information might be helpful. Otherwise, providing previously-accessed information might be more appropriate [68, 69]. A content-based model, on the other hand, aims to capture the underlying mechanism that drive the changes of search content (e.g., content of a query or a clicked webpage) across the whole search course. Follow the observations from several prior studies [38, 144, 147], where user knowledge acts as an important role in determining user behaviors, my content-based model focuses on modeling and applying user knowledge change during search. For example, if the model detects that a user has already known most of the knowledge regarding to the current task, it tends to return documents with much deeper or novel information.

To summarize, the major goal of the simulation experiments is to explore appropriate approaches (behavior-based and content-based) for assisting different cross-device search support tasks (here, I consider both information re-finding and information exploration support tasks). In addition, these experiments also help us fine-tune parameter settings for the

search support algorithms. The overall framework can be simplified as Figure 5.

3.5 ON-LINE EVALUATION EXPERIMENT DESIGN (USER STUDY II)

Off-line experiments are unable to simulate true user interactions in the system that provides the search support functions because users might only selectively adopt or even ignore these functions. To understand real-world user behaviors, I conduct a set of live experiments asking participants to search under a cross-device search support system. This on-line experiment is named to **User Study II**, which can be simply illustrated as Figure 6.

Following the main research goal, User Study II also attempts to support user search in the continued session. In addition, it further includes a baseline that does not provide any search support function, for the comparison purpose. The system with search support functions is named as Jing+ in order to distinguish the system Jing without any such support function. The following three sections provide more details about the design of search tasks, search conditions and user study procedure.

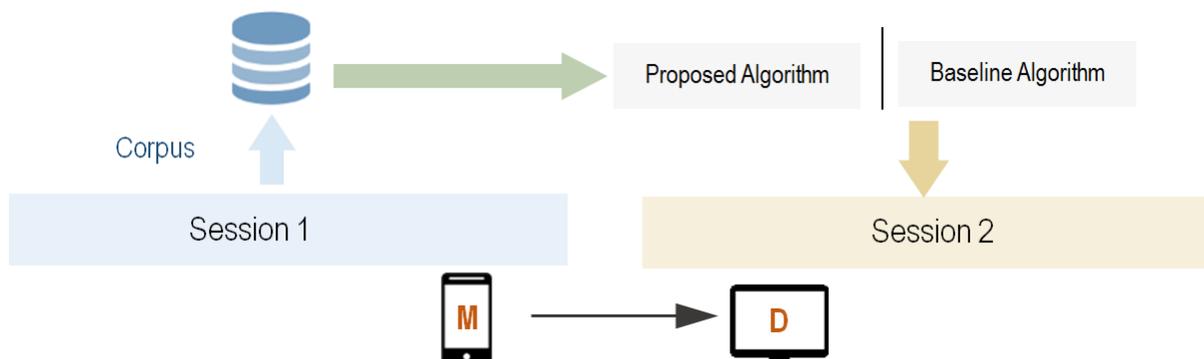


Figure 6: An illustration of the on-line evaluation experiment design.

3.5.1 Search Tasks & Conditions

User Study II reuses the web search tasks designed for User Study I. Instead of adopting all of the six search tasks, I only employ the people search and product search tasks (PE1, PE2, PD1 and PD2). The two news search tasks (NE1 and NE2) are not included for two reasons. First, news search in User Study I was conducted in the year of 2013. After four years (the time of User Study II), news search results change extensively. Thus, adopting the news search tasks makes it hard to reuse the previously-collected data in User Study I. Second, in addition to the experiment design of User Study I, User Study II further considers the control/treatment conditions, the pre-task, during-task and post-task questionnaires, all lead to a significantly high cognitive load for users. Therefore, excluding the two news search tasks can prevent users from being too fatigue during search.

In terms of the search condition, User Study II only focuses on the mobile-to-desktop (M-D) search (Section 1.4.1 describes the reason why I only choose M-D). On top of the M-D search, two types of cross-device web search systems including a baseline system (**B**) without any search support and an experimental system (**E**) providing multiple search support functions, are both examined for comparisons. A detailed description of the implementations of baseline and experimental systems are provided in Section 6.3.

In summary, a total of four cross-device web search tasks and two search support systems are taken into account in User Study II. The user study adopts a mixed within-subject and between-subject design. The order of four tasks and the order of experimental/baseline systems are rotated based on the Graeco-Latin Square design [76]. Figure 7 illustrates one round of the experiment orders consisting of eight different participants (P1 to P8). In total, we plan to recruit 24 participants in User Study II to obtain a full coverage of the task and system orders for three rounds. The Latin Square rotation ensures each task to occur equally in each order, which also helps eliminate the learning and fatigue effects.

3.5.2 User Study Procedure

3.5.2.1 Overall Procedure Same as User Study I, User Study II also began with the introduction to the study, then the filling of entry questionnaire about user demographic

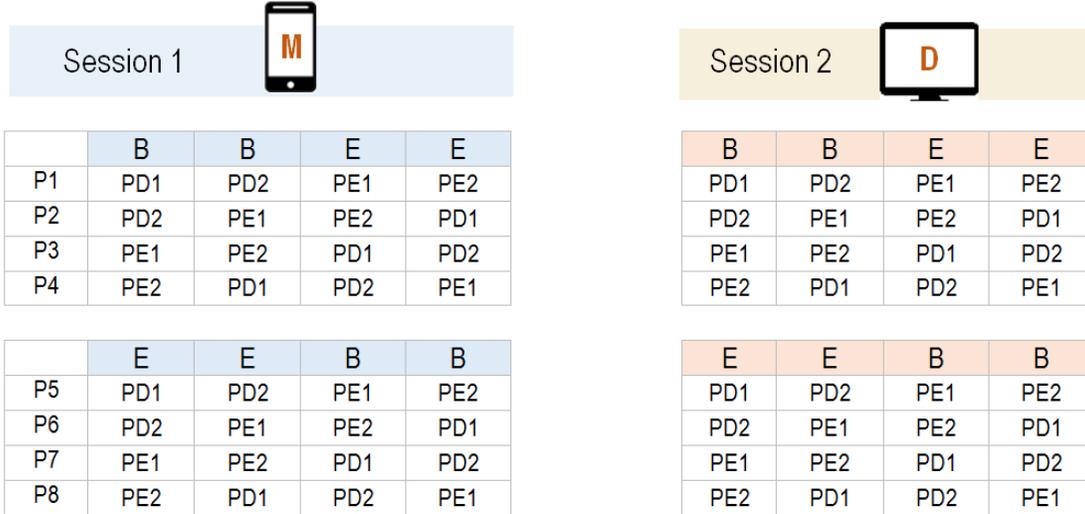


Figure 7: Task and condition orders of User Study II.

information and cross-device web search experiences (see Appendix B.1 for questionnaire). After that, each participant was given a 5-minute tutorial and 15-minute training on using the experimental system. The training task was the same as User Study I (see Appendix B.2 for task description), which asked the participant to explore the training task topic on a mobile device at first and then continue their searches on desktop to get familiar with system functionalities. After the training task, the participant was further instructed to search for four tasks with the goal of finding as many relevant web pages as possible in order to write a high-quality report (though the participant actually did not need to write a real report).

Time assignment for each search task in the first and second sessions are shown in Figure 8. Instead of allocating 7 minutes for each task in the second session, as what I did for User Study I, I assigned 10 minutes for each task in the continued session of User Study II. This is because User Study II included an in-situ questionnaire asking a set of questions regarding to the purposes of accessing a document each time before she is about to exit a web page (more details will be presented in the next paragraph). This adds a significant amount of

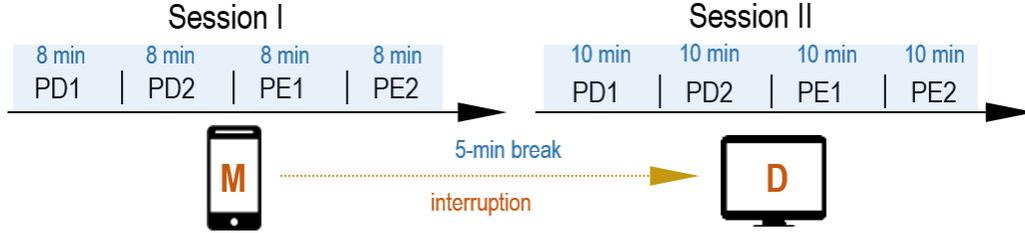


Figure 8: One example user experiment procedure for User Study II.

attention and effort during search in the second session. Therefore, three minutes were added to compensate the time for answering questions. Note that the task and device orders (in Figure 8) for other examples are rotated to leverage.

3.5.2.2 Pre-task and Post-task Questionnaire As mentioned above, each time before a participant starts a search task and immediately after she completes the task (for both the first session and the second session), she will be asked to answer a set of questions regarding her current familiarity with the given task topic (see Appendix B.4.1 and Appendix B.4.3 for details). The major goal of this questionnaire is to understand users’ pre-knowledge about the task topic before search, and the knowledge acquired during search. The questions are directly adopted from Liu et al. [90], which were purposely designed for measuring user knowledge in search. Participants’ responses to the questionnaire are used to study users’ knowledge gain in search.

3.5.2.3 In-task and Post-task Relevance Judgments In addition to surveying users’ knowledge states during search, participants are also asked to provide post-task relevance judgment for all clicked documents using a five-point Likert scale, with 1 indicating non-relevant and 5 denoting highly relevant. However, relevance judgment obtained in this way, often referred as to **user relevance**, might not reflect document utility at the page accessing time. A user may think one relevant document is not useful if she has already known most information from the document. This occurs frequently in interactive search systems where

users often encounter repeated documents under multiple similar queries [69, 75]. Researchers [104] named this type of document utility as **contextual relevance** in literature, and it can only be assessed at the time when a user reads the document. To collect the contextual relevance, I design an in-situ questionnaire asking participants to rate document usefulness. This is implemented through popping out a dialog window as shown in Figure 9 at the time when a participant intends to leave the current document. Only if the displayed questions are answered, the system will close the pop-up dialog and move forward.

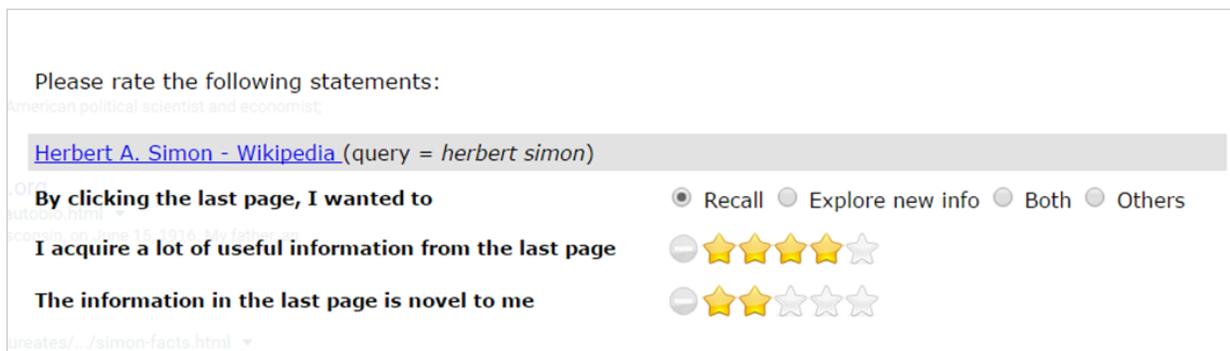


Figure 9: The contextual relevance judgment interface for User Study II.

Contextual relevance is a complex measurement related to multiple factors. A duplicate document can be contextually useful if the user’s purpose is to recall previously-accessed information. It can also be unhelpful if the user’s goal is to explore new documents. Similarly, a new document is useful if the user wants to explore novel information [69, 70, 137], whereas such document is less useful to rebuild the user’s prior knowledge state. As a result, when collecting the contextual relevance (with the question “*I acquire a lot of useful information from the last page*”), I further include a question regarding to the purpose of document access — *Recall* and *Exploration*. These two options are chosen based on their importance in the cross-device web search progress [55, 69, 70]. Furthermore, considering that novelty and diversity are important leveraging factors in complex search tasks [29], I also ask participants to evaluate the information **novelty** regarding to the clicked document.

Different from user relevance and contextual relevance, document relevance adopted in

User Study I is the aggregation of user relevance score from all participants (in §3.3.4). This is often referred as to topical relevance (see Jiang et al. [70]). In User Study II, all of these three types of document relevance, and the contextual novelty are employed and compared. To be specific, I will take into account the following four metrics.

- **Topical relevance** measures document relevance at the topical level, which is computed based on the aggregation of user relevance score with Formula 3.1.
- **User relevance** assesses the topical relevance of a document from one specific user’s point of view. It is obtained through the post-task judgment of document relevance.
- **Contextual relevance** examines document usefulness based on a user’s current search stage and search context. It is collected through in-task questionnaire in Figure 9. Contextual relevance is highly user-dependent and context-dependent.
- **Contextual novelty** focuses on measuring the amount of novel information a user can acquire from a document. Same as the contextual relevance, it is also obtained through the in-situ questionnaire in Figure 9.

It is worth noting that including the in-situ questionnaire also brings negative effect. Spending time answering the questionnaire breaks the continuity of a search process, which also affects users’ true search behaviors. However, according to my post-task interviews with participants, most of them did not feel that the pop-up dialog was distracting once they get used to it. In addition, this approach was adopted in another study when researching a similar problem [70]. Based on these evidences, I believe that my approach would reasonably reflect true user behaviors.

3.5.3 Evaluation of Search Performance

To understand how effective can users complete their search tasks over different conditions, I compared their search processes based on the following two dimensions — search performance and user perception. When evaluating the search performance, I focus on examining whether the proposed search support approaches can help participants quickly restore their task progresses, find more relevant and useful documents, collect more diverse documents and obtain deeper understanding of the search topics. Specifically, I adopt a list of search behavior

metrics such as the amount of queries or repeated queries, the complexity of search queries, the amount of document clicks and relevant document clicks, the average document relevance, the coverage (diversity) of the clicked documents, the average SERP ranking positions of the clicked documents (a higher ranking position means that the user explores deeper [158]), and so on. In addition to that, I also adopt several conventional evaluation metrics such as MAP and nDCG to evaluate result ranking or res-ranking for each individual search query. In terms of user perception, I combine both the quantitative analysis of users' post-task reports about search system efficiency and effectiveness, and the qualitative results from users' post-task interviews.

3.5.4 Summary

To examine the effectiveness of my proposed search support approaches with real-world search users, this section provides a detailed description about the user experiment design (Figure 6), search task design (Section 3.5.1), compared search condition design (Section 3.5.1), overall user study procedure (Section 3.5.2), and the expected search evaluation metrics (Section 3.5.3). More specifically, the experiment (i.e., User Study II) adopts a mixed within-subject and between-subject design (Figure 7), in which each participant needs to go through all of search tasks and two systems — one system is equipped with my proposed search support algorithm (Experimental), while the other one does not contain any search support (Baseline).

To better examine how real-world users interact with my proposed search support system, my user experiment further includes a set of pre-task, in-task and post-task questionnaires asking about users' knowledge change, user purposes of conducting certain search actions, user feelings on the search support functions and their search performances. In addition to the user-reported measurements from questionnaires, I further analyze user behaviors recorded in my search systems Jing (baseline system) and Jing+ (experimental system).

4.0 UNDERSTANDING CROSS-DEVICE WEB SEARCH

The major goal of this chapter is to answer RQ1. Each sub-question of RQ1 (from RQ 1.1 to RQ 1.3) is answered subsequently by one of the following sections. It is worth noting that I only employed the search behavior data from User Study I for data analysis.

4.1 CROSS-DEVICE SEARCH TOPICS, TRIGGERS AND INFORMATION TRANSFERRING

User responses to the Mechanical Turk survey answered my RQ 1.1. The survey was distributed in October 2013, and I received responses from 106 participants. Two of them were rejected because the corresponding respondents did not complete their answers. Among the rest participants, 40% of them are iPhone users, 54% are Android users and 6% use other types of phones. Besides, 14% of them reported experience on using iPad and 12% used other tablets. The average time to complete the survey was 8.86 minutes.

The main questions asked in the Mechanical Turk survey include the cross-device web search topics, the common triggers, and the ways of preserving and transferring information across different devices. In terms of the cross-device web search topics, Figure 10 provides the frequency distribution (in percentage) of topical categories. Product, News, Video and People are the top four cross-device information needs. This result was then applied for designing search tasks for both User Study I and User Study II. Among the four tasks, I chose three of them (Product, News and People) for user study but exclude the video search since it is difficult to record user behaviors for video websites.

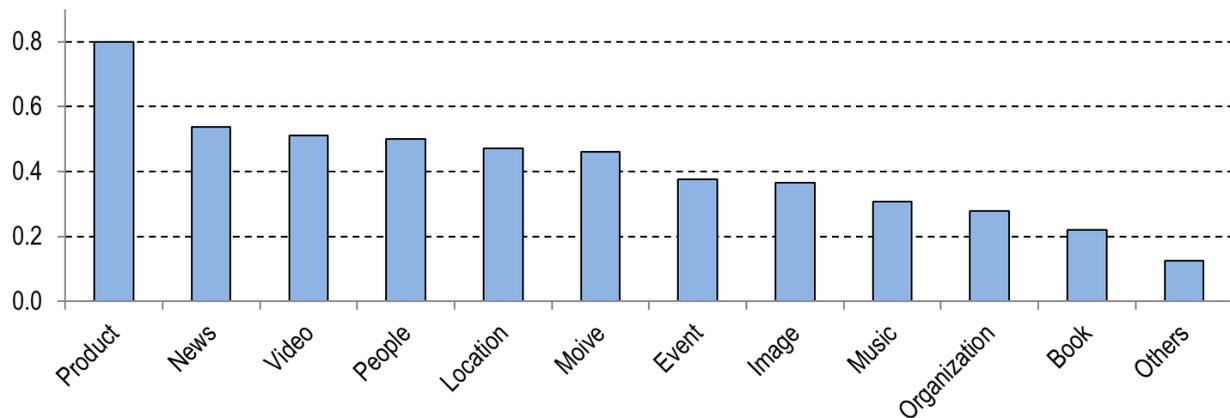


Figure 10: Cross-device search topic distribution based on Mechanic Turk responses.

Besides the above questions, I also asked about the methods participants adopted to preserve and transfer information across multiple devices so that their search tasks can be successfully continued. The survey results reveal that people preserved information through multiple ways, including sending emails (55.8%), bookmarking (52.9%), memorizing (52.9%), taking notes (18.3%), saving (17.3%), taking photos (16.3%), sharing on social media (7.69%) and others (7.69%). It is easy to see that a large proportion of respondents remain using simple preservation approaches such as emailing or even pure memorization. Some respondents even mentioned that they usually left the search result pages open on mobile devices so that these pages can be read later on. Several respondents mentioned that they sometimes employed the data synchronous functions from popular softwares such as using bookmarking function in Google Chrome, or using Evernote or Dropbox. However, they also indicated that these tools are not easy-to-use.

The answers to the questions about the cross-device web search triggers revealed that seeing something (70.2%), thinking about something (56.7%), talking about something (53.9%) are the top three categories, which are then followed by browsing something on the web (37.5%), receiving messages (9.62%) and others (3.85%). It is interesting to see that more than half respondents were triggered by thinking about something. This makes it difficult to predetermine users' information needs beforehand.

4.2 UNDERSTANDING SEARCH BEHAVIOR CHANGES

This section provides a comparison of user behaviors between M-D and D-D. Here, I use recorded search behaviors from User Study I. As for the statistical test, if the testing methods are not specifically mentioned, Generalized Linear Models (GLMs) are employed and the reported p-values are based on Wald Chi-Square test.

4.2.1 Behavioral Patterns in Cross-Device Web Search

An exploratory web search usually consists of both information exploration activities and sense-making activities [40]. Issuing queries, clicking on documents, and preserving documents are important information exploration actions [131, 155]. Sense-making actions are related to reviewing the current search progress and figuring out further search directions. Checking the workspace is a typical sense-making user behavior recorded in Jing. Information exploration and sense-making activities are both analyzed in this section.

Firstly, I compare user behaviors between different search sessions (the first session vs. the second session) and different search conditions (the cross-device search vs. the single-device cross-session search). The measures used for comparisons include number of queries (#query), average query length (#words in a query), number of visited web pages (#visited), number of saved web pages (#saved), average page dwell time (unit: second), and number of actions on workspace pages (clicking on a link in the workspace page or refreshing workspace to obtain the latest saved web pages). These measures, covering both information exploration and sense-making actions, are often employed for evaluating web search systems [130, 155].

Table 2 shows the comparison results. Numbers in bold denote the comparison between M-D and D-D, and numbers in italics represent the comparison between the two sessions (*: $p < 0.05$; **: $p < 0.01$). In the first session, participants issued more queries and visited and saved more web pages in D-D than in M-D, indicating that the search is relatively easier on desktop computers than on mobile devices. However, we did not find any significant difference on query length, despite that typing might be more difficult on mobile phones. This finding is consistent with the result from a previous study using Microsoft Bing’s search

logs [127]. Page dwell time in the first search session of M-D is significantly longer than that of D-D. This is mainly due to the difficulty of reading and navigating on the small screens of smartphones. Besides, in the first search session, participants conducted more sense-making behaviors on desktop computers than that on smartphones, which indicates that participants find relevant documents much easier on desktop computers, allowing them to spend extra time to conduct sense-making activities. However, the above-mentioned measurements seldom show significant differences in the second session, which suggests that user behaviors depend more on devices than session orders.

Table 2 also shows that the changes from the first session to the second session are quite different between M-D and D-D. For example, number of queries and number of visited and saved web pages remain at the same level in M-D, but the two measures drop significantly in D-D. This could be because the participants using desktop computers in the first session (D-D) have already explored enough documents in the result space, leaving fewer relevant documents for the second session. Also, sense-making activities increased significantly in the second session than the first session for both M-D and D-D, meaning that sense-making is more common in the middle or later search stages.

The above analysis shows that both M-D and D-D contain information exploration and sense-making behaviors. The device difference in the first session further leads to several important distinctions between M-D and D-D. Given the same amount of time, participants were able to explore information space more widely in the first session of D-D than that of M-D. However, the differences between M-D and D-D are less obvious in the second session if people use the same device (i.e., desktop computer). Besides, the insufficient exploration in the first session of M-D drove the participants to devote more effort in the second session, either to further explore the result space or conduct more sense-making actions. The complexity of user behaviors in the second session calls for a better support, particularly in the condition of M-D.

Table 2: Mean (S.D.) of several descriptive measures in cross-device search.

	M-D (1st)	D-D (1st)	M-D (2nd)	D-D (2nd)
#query	4.44(1.78)	6.22** (2.97)	4.51(2.75)	<i>4.22** (2.54)</i>
average query length	3.63(1.28)	3.65(1.21)	3.53(1.90)	3.54(1.50)
#visited	9.71(3.21)	16.0** (8.81)	11.0(8.09)	<i>10.6** (7.34)</i>
#saved	6.38(2.92)	9.78** (5.48)	6.53(5.98)	<i>6.10** (5.51)</i>
page dwell time	32.9(12.4)	22.8** (12.7)	<i>23.4** (21.4)</i>	23.9(18.2)
workspace	0.61(1.38)	1.10* (1.86)	<i>2.47** (3.04)</i>	<i>2.44** (3.31)</i>

4.2.2 Cross-Device Information Re-finding Behaviors

In addition to the above-mentioned behavioral patterns, I also discover that the information re-finding is another important cross-device behavior in the continued search session. The following section will provide a more detailed analysis of the re-finding behaviors. Since query and click are the two most important search activities in any web search system, I focus on analyzing how searchers reuse search queries and revisit documents in the continued (second) sessions. Formally, let \mathcal{S}_1 and \mathcal{S}_2 denote the two search sessions in a task, in which \mathcal{S}_1 involves m queries (q_1, q_2, \dots, q_m) and s clicks (c_1, c_2, \dots, c_s), and \mathcal{S}_2 contains $n - m$ queries ($q_{m+1}, q_{m+2}, \dots, q_n$) and $t - s$ clicks ($c_{s+1}, c_{s+2}, \dots, c_t$).

4.2.2.1 Usefulness of Information in \mathcal{S}_1 . I first look into how users rate the usefulness of information found in the first search session (\mathcal{S}_1). Each participant was asked to answer the question “*how do you rate the usefulness of the information from the first search session?*” after she completed the whole search task (i.e., at the time one finished the second session). Participants answered the question using a 5-point Likert scale, where 5 denotes the most useful and 1 the least. The average score (and standard deviation) of usefulness is **3.65 (1.13)** for M-D searches and **4.26 (0.80)** for D-D searches. In both types of searches, users

agreed that the information found in the first session are useful (with the mean score > 3).

However, participants rated information found in the first session of D-D searches to be significantly more useful than those from M-D searches ($p < 0.01$). One potential reason is that people can read and access more documents on the desktop computers than on mobile devices; thus, they may acquire more information from desktop. Besides, probably due to the presentation difference of the same information on mobile and desktop, users need to re-explore the same information on desktop after they have seen it on mobile. For example, I found that participants frequently re-read the desktop version of a Wikipedia page even though they have read the mobile version before. As a result, it is possible that people in the M-D searches will spend more efforts on task progress restoration in the continued sessions.

4.2.2.2 Cross-Session Information Re-finding In addition to a direct analysis of user report, I further analyze the cross-session information re-finding patterns, i.e., how people in \mathcal{S}_2 re-find the information from \mathcal{S}_1 , and how the re-finding patterns differ between M-D and D-D. When determining whether a query is reused (i.e., repeated), I normalized queries by converting all characters to lower case ones, trimming and discarding extra white spaces, removing stop words, and stemming words using the Porter Stemmer. Those pre-processing steps were also adopted in several prior studies [134]. Repeated clicks were determined through an exact match of a URL string, with anchor texts removed (i.e., the anchor text after a ‘#’ in a URL). Since many websites usually customize their display styles based on the accessed devices, when accessing the same content from different devices, users are often redirected to different display versions. For example, *en.m.wikipedia.org* denotes a mobile version of *en.wikipedia.org*. I manually crafted a list of matching rules to determine repeated clicks for this case.

To understand the proportion of repeated queries/clicks, I first compute the numbers of unique/repeated queries, query terms, clicks in both \mathcal{S}_1 and \mathcal{S}_2 . Then, the proportions of repeated queries, query terms, and clicks (i.e., those used in both the first and second sessions) over the first session are calculated. Table 3 reports the results. A number in bold

Table 3: Mean (S.D.) of several information re-finding measures in cross-device search.

		\mathcal{S}_1	\mathcal{S}_2	repeated	cross-session re-finding ratio (repeated/ \mathcal{S}_1)
#unique queries	D-D	6.69(3.11)	4.46(2.73)	0.53(0.84)	0.1107(0.1968)
	M-D	<i>5.01(1.81)</i>	4.61(2.78)	0.74(0.90)	<i>0.1825(0.2585)</i>
#unique query terms	D-D	8.43(3.62)	6.97(3.78)	3.52(2.24)	0.4298(0.2340)
	M-D	<i>6.79(2.62)</i>	7.26(3.82)	3.29(2.18)	0.4785(0.2812)
#unique clicks	D-D	13.46(6.68)	10.43(6.27)	2.01(3.14)	0.1479(0.1823)
	M-D	<i>8.54(3.22)</i>	11.07(6.84)	1.78(2.42)	<i>0.2132(0.2676)</i>

indicates a statistical significance (p-value < 0.05) between \mathcal{S}_1 and \mathcal{S}_2 , and a number in italics denotes a significance between M-D and D-D. Overall, I identify that around 15% of user queries (and clicks), and almost half of the query terms from \mathcal{S}_1 will be reused in \mathcal{S}_2 . This indicates that information reuse behaviors, particularly the reuse of query terms, is an indispensable part of a cross-session web search process.

I also observed that users in M-D reused a relatively higher percentage of queries, query terms, and clicked results from \mathcal{S}_1 . This is consistent with our expectation in the above subsection – people need to spend more efforts on re-exploring the same information that has been accessed in \mathcal{S}_1 for M-D. However, Table 3 only provides the aggregated statistics across all sessions. To further understand how the reuse behaviors change over time, I conducted a temporal analysis of cross-session information re-finding in the next section.

4.2.2.3 Temporal Analysis of Information Re-finding I study two research questions in this section. First, I want to explore how the cross-session information re-finding behaviors change over time in the continued session (i.e., \mathcal{S}_2). Second, since the re-finding behaviors can happen not only across different sessions but also within each session [29, 68], I further examine how the cross-session information re-finding pattern differs from the within-session information re-finding. To make a fair comparison, I only analyze the differences

between the last few search behaviors of \mathcal{S}_1 and the first few behaviors of \mathcal{S}_2 . Here, our goal is to understand how the re-finding behaviors are affected by a temporal interruption. Following the analysis provided in Table 3, I consider the reuse of search queries, query terms, and re-finding of clicks.

To understand the first question, I extract the first three queries from \mathcal{S}_2 . For each query q_i , I then compute whether it will repeat any query from \mathcal{S}_1 . After calculating the repeated percentages for all users and tasks, I average their values to obtain the repeated percentages for queries in each position (i.e., i). The results are plotted in Figure 11 (a). Note that the X-axis in Figure 11 denotes different positions in \mathcal{S}_2 . Last indicates the last one of \mathcal{S}_1 , 1st, 2nd and 3rd indicate their positions in \mathcal{S}_2 . ** (or *) indicates the significance at the p-value < 0.05 (or 0.10) level between the two neighbored data points. For example, the blue ** in (a) indicates a significance between Last (\mathcal{S}_1) and 1st (\mathcal{S}_2) of M-D.

Here, I only include the first three queries since half the \mathcal{S}_2 -sessions do not contain more than three queries. A similar procedure is applied to calculate repeated term percentages, except I compute and average the percentage of repeated terms for each query q_i (i.e., $\# \text{repeated terms in } q_i / \# \text{terms in } q_i$). The results are plotted in Figure 11 (b). In Figure 11 (c), I focus on the repeated clicks. Instead of only considering the first click c_{s+1} , I put the first three clicks together as the first group (i.e., c_{s+1} , c_{s+2} and c_{s+3}), then the next three clicks as the second group (i.e., c_{s+4} , c_{s+5} and c_{s+6}), and so on so forth. The repeated percentages are calculated over the first, second and third groups. In this way, the variance of repeated click percentage in each position is smoothed. All three plots suggest a unified information re-finding pattern - with time moving on, people are less likely to re-find the information from \mathcal{S}_1 , probably due to that they have already restored their task progresses after interruption and started to focusing on more advanced topics. This also reconfirms the finding from Tyler and Teevan [137], who discovered that cross-session re-finding is more likely to happen at the beginning of a search session than at the end.

To answer the second question, I extract the last query of \mathcal{S}_1 (i.e., q_m), and compute whether or not q_m repeats any previous query in \mathcal{S}_1 (Note that the repeated query percentage

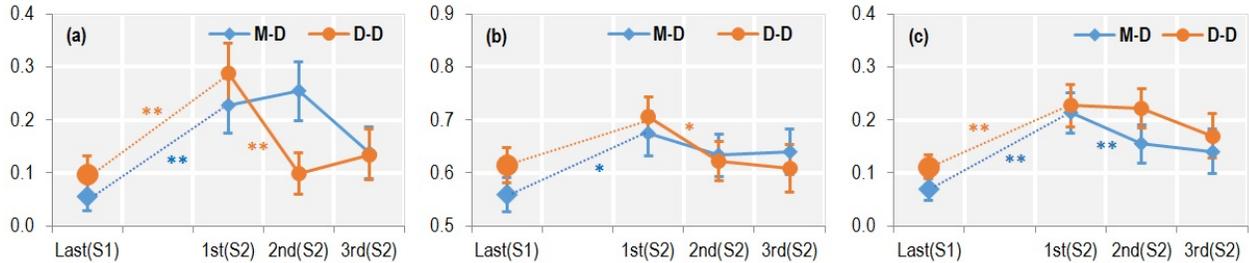


Figure 11: Information reuse percentages (S.E.) for queries/query terms and clicks.

for q_{m+1} in Figure 11 is computed over all queries in \mathcal{S}_1 instead of $\mathcal{S}_1 - q_m$. I also recalculate the percentages over $\mathcal{S}_1 - q_m$; however, there is not too much difference in terms of numbers). After averaging all users and tasks, I obtain the percentage of q_m being a repeated query. A similar procedure is then applied to compute repeated term and repeated click percentages for the last queries and the last three clicks in \mathcal{S}_1 . All these values are corresponding to the information re-finding behaviors within a single search session. The results are also plotted in Figure 11 for comparison, in which I find that with a long-time interruption (roughly one hour), people tend to conduct more repeated behaviors at the beginning of the continued session. One potential reason is that people forget due to interruption. Our result in Figure 11(a) indicates that around 30% of the first queries in \mathcal{S}_2 of D-D reused the queries from \mathcal{S}_1 . This is consistent with a study from Teevan et al. [134], who found that people failed to remember their original queries 30% of the time after a half-hour to one-hour interruption.

4.2.2.4 Summary. Results in this section clearly demonstrate the usefulness of information in \mathcal{S}_1 , and the necessity of transferring useful information across multiple sessions to support task resumption. Specifically, there are two interesting patterns regarding the information re-finding behaviors. First, I found that the information re-finding behavior is more likely to happen at the beginning of a continued session, mostly for resuming prior tasks. Second, information re-finding patterns for different search conditions (i.e., M-D and D-D) are different. In our case, although participants in M-D perceive less usefulness of the

information from \mathcal{S}_1 than that in D-D, they still re-find a significant amount of information from \mathcal{S}_1 , even more than that in D-D. This indicates that the device factor does significantly affect people’s information transition across sessions. Therefore, a proper cross-session search support should consider both conditions and I will provide more details in the section related to cross-device search support.

4.2.3 Understanding Search Strategy through Post-task Interviews

To understand what search strategies users adopted in different search sessions and search conditions, I summary the post-task interview for User Study I. The detailed design about the interview questions was provided in Section 3.3.5.

Participants’ responses about their adopted search strategies mostly lied on the following aspects. Many participants mentioned that they attempted to search for simple and easy-to-find information in the first session, and then explored or read deeper in the continued sessions. Others mentioned that they were trying to collect as many documents as possible in the first search session, and then read these documents in details at the continued session. These two strategies are consistent with our findings in Table 2, in which I found more sense-making behaviors in the second session while more information gathering activities in the first session. When answering the questions related to the search strategies they adopted in different devices, most participants complained that mobile phones are slower, the screen and keyboard are too small so that they tended to issue simpler queries on mobile while left the harder ones for desktop, particularly when the desktop computer was used for task continuation. For the third question, most of them answered that their main strategy is to memorize the the queries they searched and the webpages they clicked. And then reissue the same queries or re-visit the same webpages (i.e., the information re-finding) for task restoration. Several of them also pointed out the necessity of building alternative supporting tools such as providing a workspace for saving documents, designing a function for bookmarking or even photo-taking.

Overall, the post-task interviews revealed two important behavior patterns adopted in the continued search sessions — the information re-finding behavior for task restoration,

and the sense-making behavior for task advancement. In the future, I will explore these two search patterns in more details.

4.3 UNDERSTANDING SEARCH CONTENT CHANGE

The above search behavior analysis provides few understanding for the search content change. Previous studies [38, 90, 147] found that, with the search goes on, people tend to issue more difficult queries, adopt more complex search strategies and conduct more in-depth explorations. One likely reason is that human knowledge keeps growing during search, which further drives the adoption of more complex search content [38, 144, 158]. Eickhoff et al. [38] designed a set of metrics such as query complexity, content coverage and page dwell time to characterize human knowledge change patterns. My dissertation follows their methodology by considering the temporal changes of different metrics in a cross-device search process. Specifically, I evenly divide the whole search process (with two sessions) into six search stages - three in the first session and the other three in the second. Then, I calculate and plot the values of each metric at different stages. The metric change over different stages are then used to examine users' search content change.

4.3.1 Metrics Related to Search Content Change

A set of six metrics, through combining and tailoring the measurements used in both Eickhoff et al. [38] and Zhang et al. [158], are adopted in my study.

Query Complexity is an important indicator of users' domain knowledge, and it has been measured in several different ways [38, 144, 158]. Motivated by the finding that experts (with relatively higher domain knowledge) tend to use domain-specific query terms that are unknown to novices (with relatively lower knowledge), White et. al. [144] treat the percentage of technical query terms as query complexity. However, this method requires building a domain thesaurus based on a large data collection, which is difficult in our setting. Eickhoff et al. [38] took a different approach by measuring query complexity

through its readability score [31]. Specifically, they utilized a list of 30,000 English words compiled by Kuperman et al. [84]. Each word in the list is associated with an age at which native speakers typically learn the term. Therefore, the higher this score, the more complex a word is. Then, the query complexity is measured by the maximum age of acquisition across all query terms. In addition, because query length was found to be a strong predictor of domain knowledge [158], it can also be applied for measuring query complexity. To summarize, I choose both **readability level** and **query length** for measuring query complexity.

Click-through Complexity. Users examine relevant documents by clicking on them. Some relevant documents can be easily discovered and then clicked, whereas some others may require pre-acquiring more topical knowledge before users can realize their relevance. Therefore, the likelihood of discovery can be a good measure of users’ domain knowledge. In this dissertation, I borrow the idea of Inverse Document Frequency (IDF) for computing the *likelihood of discovery* for each click-through document. This approach is also applied in Shah et al. [123]. Since I am only interested in the relevant clicked documents, this measure only examines those documents with user relevance score higher than 3.0¹. Formally, for each document d_i that has a relevance score larger than 3.0, its complexity score $C(d_i)$ is calculated using Formula 4.1, where N_d is the total number of clicks on the relevant documents under the current task, and n_{d_i} indicates the number of participants who clicked d_i .

$$C(d_i) = \log \frac{N_d}{n_{d_i}} \quad (4.1)$$

Depth of Exploration. Zhang et al. [158] found that the average ranking position of the click-through documents in search engine result pages is an important indicator of users’ knowledge level. I hypothesize that people tend to conduct more in-depth explorations and seek more novel knowledge with their continuous searching of the same topic.

Website Domain Diversity. The above three metrics mainly measure users’ knowledge depth. This metric attempts to quantify users’ knowledge breadth [148]. Following several

¹Each participant is asked to rate the relevance of a saved document in a 5-point Likert scale, where a score of 3.0 means a neutral relevance.

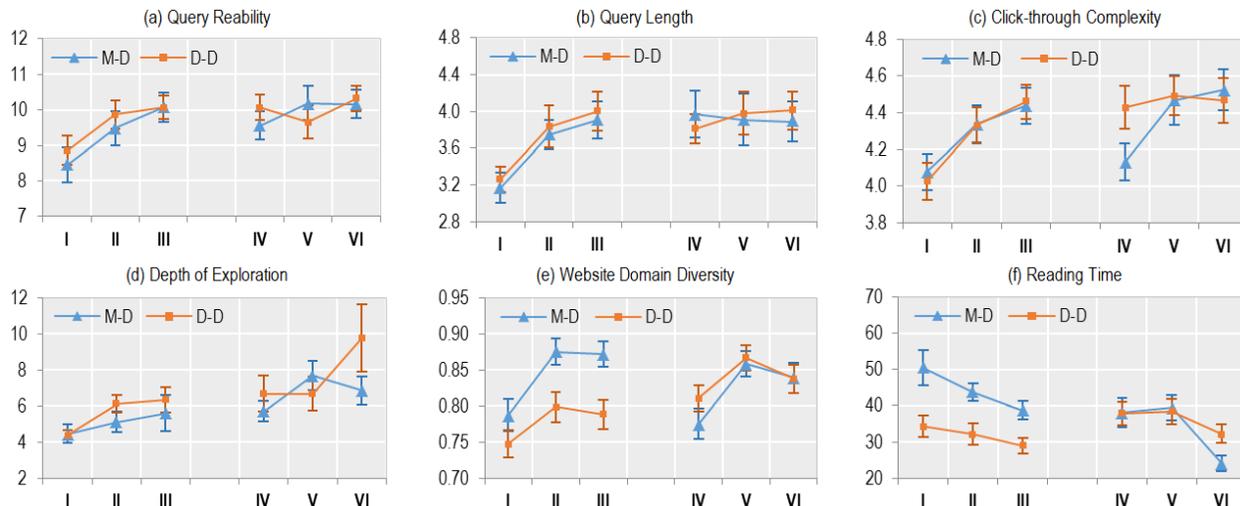


Figure 12: Temporal evolution of several content metrics in cross-session searches.

of previous studies [38, 144], I measure knowledge breadth through website domain diversity. It is calculated by the ratio between the number of unique website domains and the number of total domains appearing on search engine result pages.

Reading Time. The amount of time a user spent on reading click-through documents has been found to be an important signal of domain expertise. Both White et al. [144] and Kelly et al. [77] have discovered that users’ document reading time decreases with the increase of topical familiarity because they become more adept at reading domain-related content. Therefore, the average reading time for click-through documents is also employed as an indicator of users’ knowledge change.

4.3.2 Temporal Changes of Search Content Metrics

Based on the findings of previous studies [38, 144, 158], I expect that people will spend less time on reading clicked webpages and conduct more in-depth (i.e., higher query complexity, higher click-through complexity and higher SERP ranking position of click-through documents) and in-breadth explorations (i.e., higher website domain diversity) while acquiring more domain knowledge. Therefore, I compute the above six metrics and plot their values

over six evenly-divided search stages. The results are plotted in Figure 12, which plots the temporal development of (a) query readability, (b) query length, (c) click-through complexity, (d) depth of exploration, (e) website domain diversity and (f) reading time across six search stages in cross-session web searches. I, II, III come from the first search session, and IV, V, VI come from the second session.

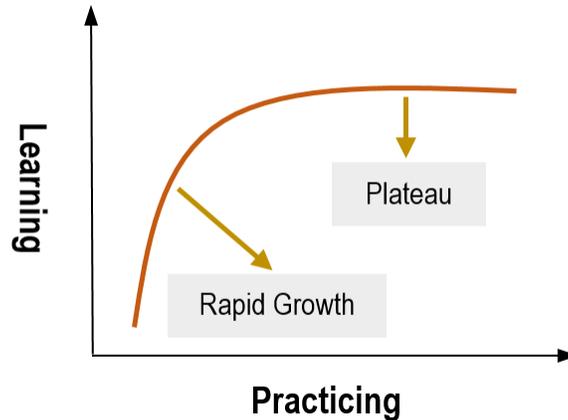


Figure 13: Knowledge growth pattern summarized from user study results.

4.3.2.1 Overall Temporal Patterns. As shown in Figure 12, there is an overall trend of value increase for all these measures (except an overall trend of decrease for the click-through reading time since it is expected to be negatively correlated with knowledge growth [77, 144]), either in each separate session or across two search sessions. This trend is consistent in both M-D and D-D. To understand the significance of metric change, I conduct the non-parametric Wilcoxon signed-rank tests for related samples since the data is not normally distributed. Statistical tests of two consecutive search stages show that $I < II$ (a, b, c, d and e on both M-D and D-D), $IV < V$ (e on D-D; c, d and e on M-D) and $VI < V$ (f on M-D), where $<$ stands for a significant difference being observed at the 0.05 level (i.e., $p\text{-value} < 0.05$). This suggests that domain knowledge gains (i.e., search content changes) dramatically at the beginning and gradually reaches a stable status, which is illustrated in Figure 13. Among existing studies of learning curve in education domain [64, 152], the learning curve with

logarithmic rise and a gradual saturation is one of the most common learn curve models. My findings in Figure 12 clearly align with the learning, where people learn quickly with few practices at the beginning and gradually arrive at a plateau as shown in Figure 13. It is worth noting that my finding on knowledge saturation is consistent with Liu et al. [90]’s finding of ceiling effect for domain knowledge growth in their lab study. Overall, the above findings suggest that knowledge learning is a legitimate approach for modeling human’s search process, and I would like discuss more modeling details in later sections (see §5.2).

4.3.2.2 Comparisons between M-D and D-D. Although Figure 12 shows an overall similar trend for cross-device (i.e., M-D) and same-device cross-session (i.e., D-D) web searches, I do find several significant differences between these two conditions. To compare M-D and D-D, I only focus on the first sessions, because the second sessions are all performed via desktop. Non-parametric Mann-Whitney tests for independent samples are adopted to examine statistical significance for two reasons: (1) our data is not normally distributed; and (2) the same user who searches on different devices also performs different tasks so that we cannot hold device difference as the only independent variable. Thus, we cannot use statistical tests for related samples.

I do not find any significance through statistical tests for query complexity measures (i.e., a, b) and click complexity measure (i.e., c) on different devices, which indicates that mobile devices do not limit users’ selections of queries and clicks. However, I do observe that people spend significantly more time reading ($M > D$ for f in all three stages) and exploring less in-depth on search engine result pages ($M < D$ for d in Stage II) on mobile devices, which is mainly due to relatively small screen and keyboard sizes. In addition, even with several limitations on mobile information access, people on mobile devices still encounter more diverse domains ($M > D$ for e in both Stage II and Stage III). Since website domain diversity is measured by the content diversity of search engine result pages, one potential explanation for such a phenomenon is that returned search results have been specifically tailored for displaying more diverse content on mobile searches. To test this assumption, I compute the average unique number of website domains on both mobile devices (9.33 ± 1.26) and desktops (9.35 ± 2.00) for the same queries. I do not find a significant difference, which

indicates that users tend to generate queries with more content diversity on mobile devices.

4.3.2.3 Forgetting Effects. Figure 12 shows significant drops of several knowledge learning metrics between Stage III and IV ($III > IV$ for c and e on M-D), which contradict the overall trend for knowledge growth. I hypothesize that this is due to forgetting: after a one-hour interruption, people may forget some information they had accessed. They need to recall information and reconstruct their memories in order to continue the previous search tasks. Such phenomena has been demonstrated in search behavior analysis (see Section 4.2.2). It is worth noting that result significance between Stage III and Stage IV is only discovered on M-D while all learning metrics stay at the same level on D-D. I suspect this is due to the significantly different information contexts (e.g., screen size, font size, webpage design) on different devices. Many prior studies have identified context-dependent forgetting effects in psychology [44, 126] and their influence on an individual’s memory is reduced. This explains the difference between M-D and D-D. However, there still needs a deep analysis for better understanding the assumption.

4.3.2.4 Direct Evidence from User Study II. The above results of search content change are analyzed on the data collection of User Study I, in which a set of implicit behavior metrics are adopted. However, user knowledge is not directly measured. To explicitly measure the change of human knowledge, a 7-scale Likert question² directly asking users’ current knowledge levels is designed for User Study II. The question was presented both before and after each search session; and thus was asked 2×2 times (which includes the pre-1st session knowledge, post-1st session knowledge, pre-2nd session knowledge and post-2nd session knowledge). Based on the above analysis, I expect an overall knowledge growth pattern as shown in Figure 13 and a forgetting effect.

The mean (and standard error) of user-reported knowledge levels are provided in the following Figure 14. Here, ** indicates a result significance between two adjacent search stages. I adopt the Wilcoxon Singed-rank test since the data is not normally distributed.

²which adopts from the questions designed by Liu et al. [90], whose study was dedicated to analyzing users’ knowledge change across different search sessions. The question is “How *familiar* are you with the topic of this task?”

The result confirms my previous expectations. First, user knowledge gains the most in the first session. On average, there is a 2.59 knowledge growth after the search of the first session, whereas such number is only 1.20 for the second search session. This implies that user knowledge tends to be more saturated in the second session even if users have access to desktop computers for searching. Second, after about half an hour break, user knowledge drops around 0.25, this is consistent with my previous analysis of the forgetting effect (see Figure 12 (c) and Figure 12 (e)).

Overall, both the direct and indirect measurements of user knowledge in the above sections provide a clear knowledge growth pattern as indicated in Figure 13. The interruption, combining with the device factor can further cause the forget and drop of user knowledge.

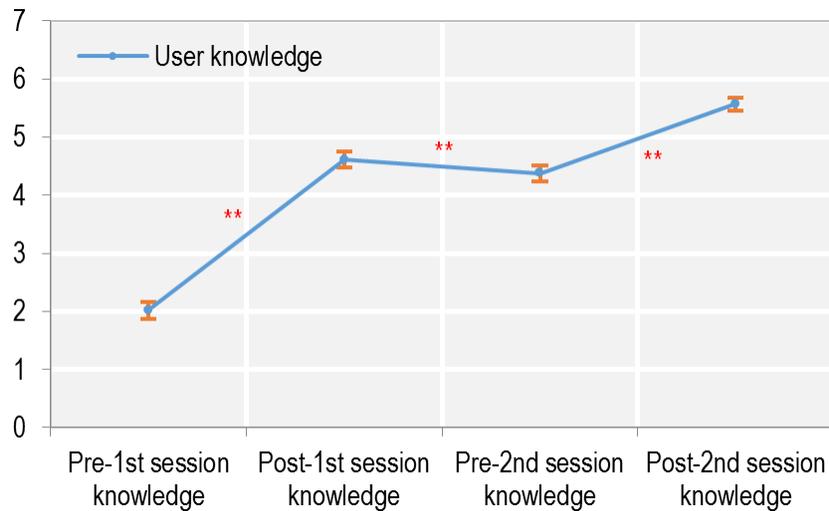


Figure 14: Mean (S.E.) of participant-reported knowledge in four search stages.

4.4 ANSWERS TO RQ1

Overall, this chapter answered the research question RQ1 regarding to the understanding of cross-device web search processes. Each of the above section attempted to answer one subquestion for RQ 1. Here, I list the answers to all of the three sub-questions.

- **Answers to RQ 1.1.** According to the responses of Mechanical Turkey survey, Product, News, Video and People are the top four categories for cross-device information needs. To transfer important search information across different devices, most of the respondents tend to send emails (55.8%), use bookmark functions (52.9%) or just memorize the information (52.9%). In terms of what triggered a cross-device web search, the top three categories are seeing something (70.2%), thinking about something (56.7%) and talking about something (53.9%). All of these results were applied to design my User Study I & II. The detailed description of the survey result can be found in Section 4.1.
- **Answers to RQ 1.2.** According to the search behavior analysis in Section 4.2, I found three important types of search behaviors during the whole search process — information exploration (search for relevant documents), sense-making (search for novel, diversity or reflect task progress) and information re-finding (seek for previously-access information). For both M-D and D-D, users in the first session tend to perform more information exploration behaviors, while they conduct more information re-finding and sense-making behaviors in the continued search session.

When comparing between M-D and D-D, I did find several differences. First, users in the first session of M-D issued fewer queries, clicked and saved fewer documents and stayed longer on each document, mainly due to the inconvenience of typing and reading on mobile devices. Second, although participants in M-D perceived less usefulness of the information from the first session, they still re-find a significant amount of information from the first session, even more than the participants on D-D.

- **Answers to RQ 1.3.** Based on the temporal analysis of a set of six implicit behavioral metrics, Section 4.3 discovered a consistent behavioral pattern where participants tended to seek for deeper and more diverse content with search goes on until reaching a plateau. This was hypothesized to be caused by participants' knowledge growth, and it was further validated through user-reported knowledge levels in User Study II.

When comparing M-D and D-D, I also found several differences. First, due to the relative difficulty of typing, reading and exploring, participants stayed longer in a page, explored deeper and issued more diverse queries. Second, the forgetting effect between two search sessions was more obvious on M-D which might be due to different information display styles across devices.

5.0 MODELING CROSS-DEVICE WEB SEARCH PROCESS

The above section has provided a clear understanding about users' search strategies and search content changes within a complete cross-device search process; however, such understanding was obtained through post-hoc data analysis, and mainly focused on qualitatively interpreting human search process, which might be difficult to be applied for the cross-device search support due to the lack of a quantitative approach that can automatically detect users' search strategies, model their search contexts and further be applied to personalize search results. The main focus of this chapter is to provide one such quantitative model.

Learning from the modeling rationale of previous studies [54, 95, 155, 158], I hypothesize that users' search activities (e.g., clicking a webpage or issuing certain content in queries) are driven by latent factors. For example, user-adopted search strategy might be a latent factor – the strategy of re-finding usually drives users to issue repeated queries or click the already-visited webpages, whereas the strategy of exploring new information leads to the opposite. The knowledge state can be another latent factor – saturated knowledge on one topic will push users to search for another topic while in the knowledge growing stage, users tend to stay in the same topic. Therefore, a hidden variable model, as illustrated in Figure 15, is needed for such modeling. The hidden factor may change/update with users continuing searching and exploring the result space. The goal of this chapter is to uncover the hidden factors based on the historical observed search behaviors (or search content).

In this chapter, I will emphasize on the modeling of search process for the continued sessions because of that 1) search behaviors in this session are more complicated, including both information re-finding and exploration activities (see the result analysis in Section 4.2.1); and 2) involving a continued session is the most unique characteristic of a cross-device web search, whereas other search formats seldom handle the similar problem. In

addition, when analyzing the modeling approaches for cross-device web searches, I will take into account both of the same-device cross-session web search and cross-device web search for comparison, particularly for modeling the search content since the behavior-based approach generate very similar results.

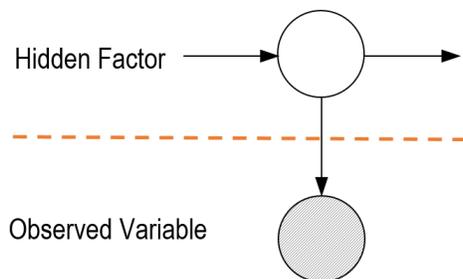


Figure 15: Hidden variable model for search behavior (content) modeling.

5.1 BEHAVIOR-BASED SEARCH PROCESS MODELING

This section focuses on modeling users’ search behaviors with the hidden variable model. The hidden factor can be thought as the search strategy and the observed variable indicates one type of search behaviors. This aligns with the quantitative modeling approach in several prior studies [54, 150, 155] on studying similar research topics, in which the Hidden Markov Model (HMM) was often adopted. The following sections will provide a more detailed description for this model.

5.1.1 Modeling Search Process with HMM

HMM is a commonly-adopted machine learning algorithm for modeling the sequential data patterns. It assumes a sequence of hidden states, each drives the adoption of an observed action. The hidden state sequence is further assumed to be a Markov chain, meaning that the next hidden state only relies on its nearest precedent state. An illustration of the HMM for this study is provided in Figure 16, in which users’ search actions (\mathcal{A}) such as query

and click are observed variables, and they are generated by hidden search states (\mathcal{H}) with different probabilities (emission probability). The hidden state sequence forms a Markov Chain, where each search strategy can transfer to or be transferred from another search strategy with different probabilities (transition probability).

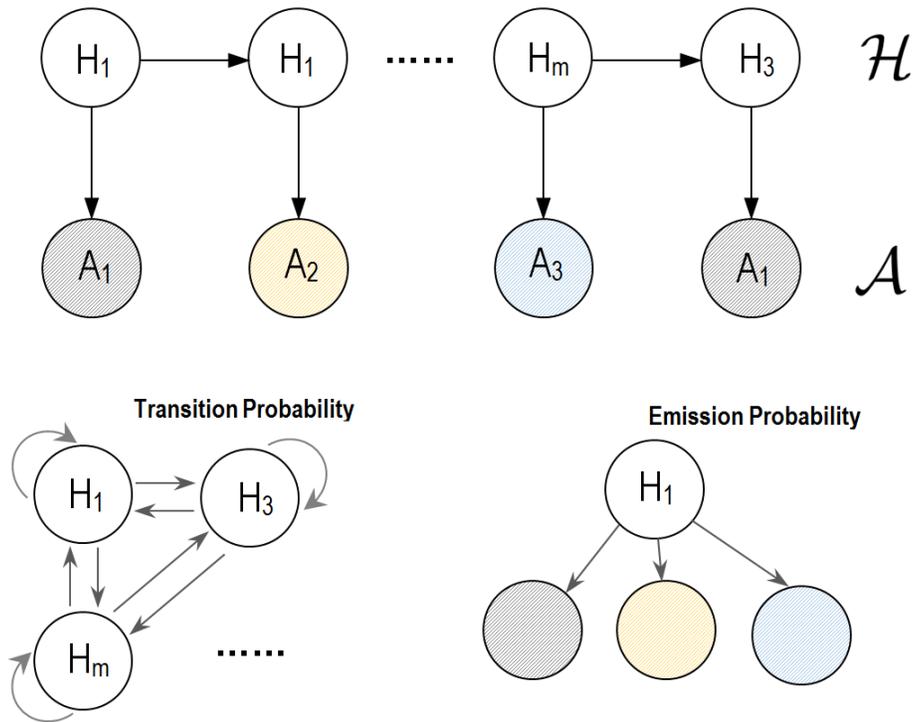


Figure 16: An illustration of HMM-based search process model.

Defining user actions. \mathcal{A} denotes a set of observed search behaviors. Han et al. [54] and Yue et al. [155] employed simple search behaviors such as query, click and save for the observed search behaviors. However, those behaviors are too coarse-grained and thus are unable to distinguish the same behaviors with different intents. For example, clicking a document with a long-time dwell might indicate that a relevant document is found, whereas a short-dwell click could mean that a document is of marginal relevance or irrelevance [53, 56]. Therefore, Luo et al. [95] and Guan et al. [46] divided a click behavior into two categories: SAT (SATisfactory) click (a click with more than 30-second dwell time) and non-SAT click

(a click with less than 30-second dwell time).

User interactions for cross-device web search are more complex. For example, two clicked documents with the same amount of dwell time might reflect different intentions – a short-dwell click on an already-visited document might be for recalling information, whereas a short-dwell click on a new document usually indicates that a document is non-relevant [56]. As a result, this dissertation considers a set of complicated user actions defined in Table 4.

Determining the number of hidden states. \mathcal{H} denotes the hidden states (i.e., search strategies) a user employs. In a fully-automated algorithm, the number of hidden states m is an important parameter to setup, which actually refers to the model selection problem. A complex model with a large number of hidden states can describe the data more accurately, but may have a higher risk of over-fitting. A simple model is less likely to over-fit, but it is usually unable to uncover the underlying data patterns. Information criterion such as the Akaike information criterion (AIC) or its variants [3] and Bayesian information criterion (BIC) [102] are often adopted to assist the selection of model parameters. This dissertation adopts the BIC because it takes into account the data sample size.

Specifically, suppose that the number of parameters in a HMM is \mathcal{P} , and the number of observed user behaviors is \mathcal{S} . BIC is defined as Formula 5.1. \mathcal{L} denotes the log-likelihood of the behavioral sequences, which can be computed through a learnt HMM. \mathcal{P} can be computed with $(m - 1) + (m - 1) \times (m - 1) + m \times (n - 1)$, in which n denotes the number of action types (in this dissertation, $n = 6$ as shown in Table 4). A larger log-likelihood (more fitting on the data) with fewer parameters (more generalizable) is preferred.

$$BIC = -2 \times \log \mathcal{L} + \mathcal{P} \times \log \mathcal{S} \quad (5.1)$$

5.1.2 Model Setup for HMM

The dataset for our HMM experiments includes both M-D and D-D searches. In this section, I mainly explains the HMM results for M-D. In total, there are 1,066 search actions collected from 24 participants \times 3 tasks on M-D = 72 search sessions. Figure 17 (b) pro-

Table 4: Search interactions adopted in this dissertation.

Abbr.	Interaction	Description
1. RQU	repeat query	a user issues a query that has been used in prior search sessions ¹
2. NQU	new query	a user issues a query that has not been used in prior search sessions
3. RCLD	repeat click, long dwell	a user clicks a document that has been accessed in the prior session, and stays more than 30 seconds
4. RCSD	repeat click, short dwell	a user clicks a document that has been accessed in the prior session, and stays less than 30 seconds
5. RCUN	repeat click, unknown dwell	a user clicks a document that has been accessed in the prior session, and stays with unknown time ²
6. NCLD	new click, long dwell	a user clicks a document that has not been accessed in the prior session, and stays more than 30 seconds
7. NCSD	new click, short dwell	a user clicks a document that has not been accessed in the prior session, and stays less than 30 seconds
8. NCUN	new click, unknown dwell	a user clicks a document that has not been accessed in the prior session, and stays with unknown time ²

Note¹: a query q_j is defined as a repeated query if it matches any query q_i from the prior session with more than 80% of similarity (I also tried other thresholds such as 75% and 85%, in both the HMM results looked quite similar). Here, the similarity is computed with the $\#$ overlapped query terms / $\max(\text{length of } q_i, \text{length of } q_j)$.

Note²: in this dissertation, the dwell time of a clicked document is computed based on the time interval between click and the next search action. If a click is the last logged search action in a session, the dwell time cannot be reasonably computed and thus is labeled as unknown. In addition, if a user stays more than 3 minutes on one webpage, it might indicate that she has left. I also label it as unknown.

vides the proportion of each behavior type averaged across all search sessions, from which I observed several interesting patterns. First, issuing new queries and clicking new web pages take relatively high proportions, indicating that participants were mainly exploring new information. Second, there is also a reasonable amount of information re-finding behaviors — around 21% of search actions are repeated behaviors (repeated clicks or queries), implying that participants may forget some information. The HMM requires setting the number of hidden states. Figure 17 (a) plots the BIC values under different settings of hidden states, in which I find that BIC reaches the optimal value when the number of hidden states is set to 4. As a result, this is the default setting in my following experiments. Note that such setting was also applied for the HMM model for D-D since the BIC evaluation also reached the best under 4.

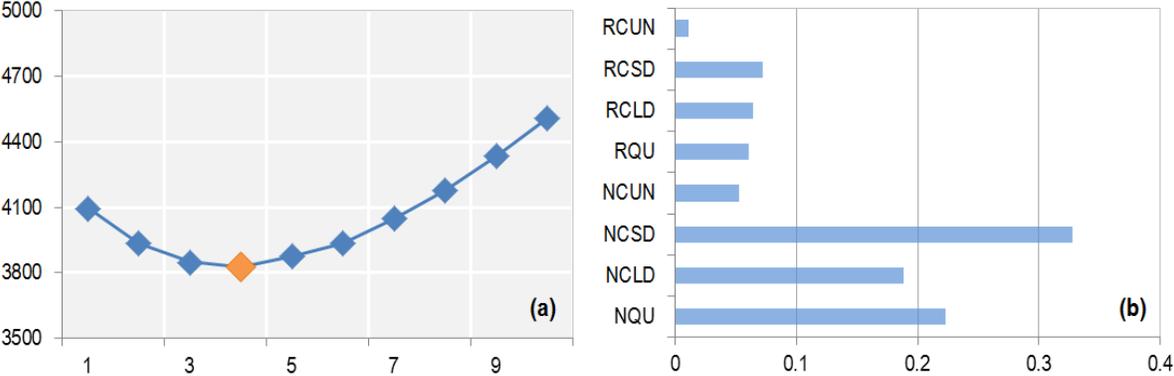


Figure 17: Experiment data (search behavior) and HMM parameter analysis.

5.1.3 Explaining HMM Outputs

An HMM consists of three important components — the prior probability (likelihood of a state being a starting state), the emission probability (probability of a hidden state to generate search actions), and the transition probability (probability of a hidden state transits to another hidden state). Those are all illustrated in the HMM results for the continued session of M-D search in Figure 18. The shaded circles with labels H1 to H4 denote the four hidden states, the double circle with label *start* indicate the start state and the numbers

from this state denote the prior probabilities, the probability table near each hidden state represents the emission probability and the arrows among hidden states reveal the state transition probability. Here, I only report the emission/transition probabilities that are larger than 5%. A detailed explanation of each search interaction can be found in Table 4. The label for each hidden state is manually crafted based on the qualitative understanding of emission and transition probabilities.

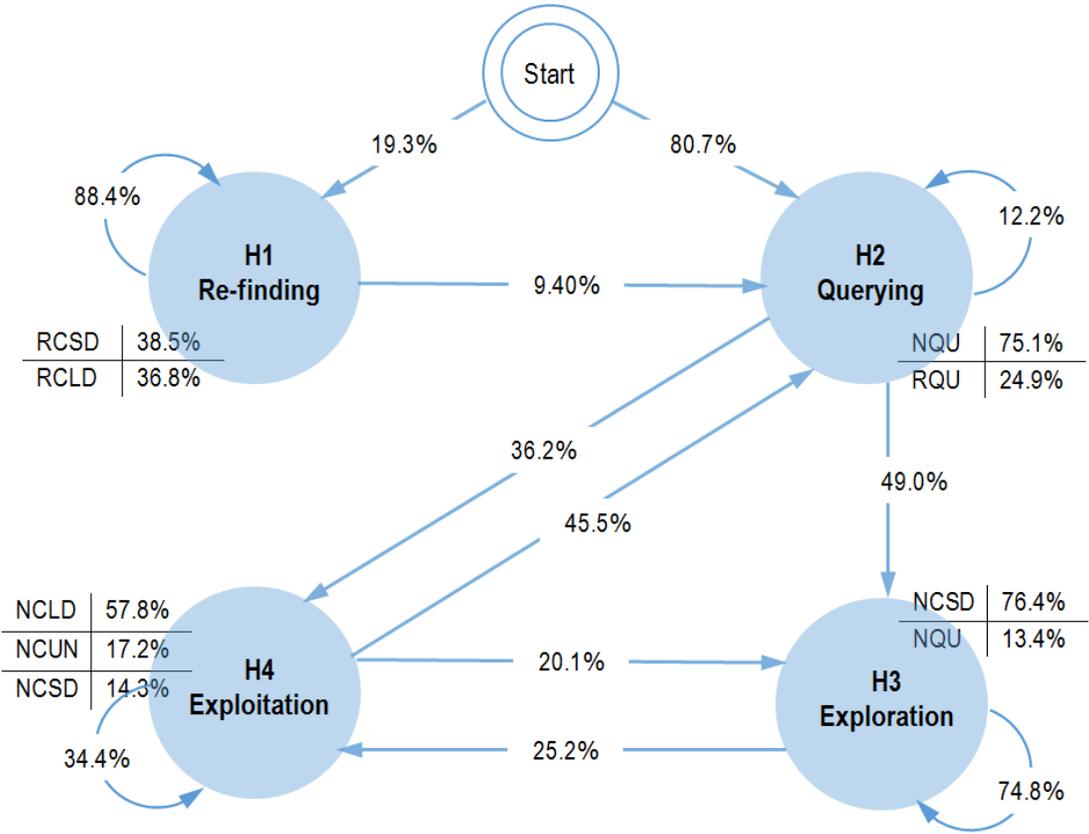


Figure 18: HMM results for the continued session of M-D.

Specifically, when adopting H1, a participant is likely to re-visit the previous webpages. Thus, I name it as *re-finding*. It is interesting to observe that participants tend to re-find information within a continuous time slot since the self-state transition probability is larger than cross-state transitions. H2 is another hidden state that mainly generates queries so that

I name it as *querying*. In state H3, a participant clicks new webpages but stays a relatively short period of time. I assume that it might be due to that the participant is exploring new information and encounter marginal relevant or non-relevant ones. In state H4, a participant clicks new webpages but stays quite a long time, which usually means that the participant finds a relevant document. As a result, I label H3 and H4 as *Exploration* and *Exploitation* respectively, to highlight their differences.

The HMM output for D-D looks alike to that of M-D, which is illustrated in Figure 19. Most of the hidden states, emission probabilities and transition probabilities are similar to M-D, except that there are fewer exploitation to querying but more exploitation to exploration on M-D, indicating a high cost of typing query on mobile devices.

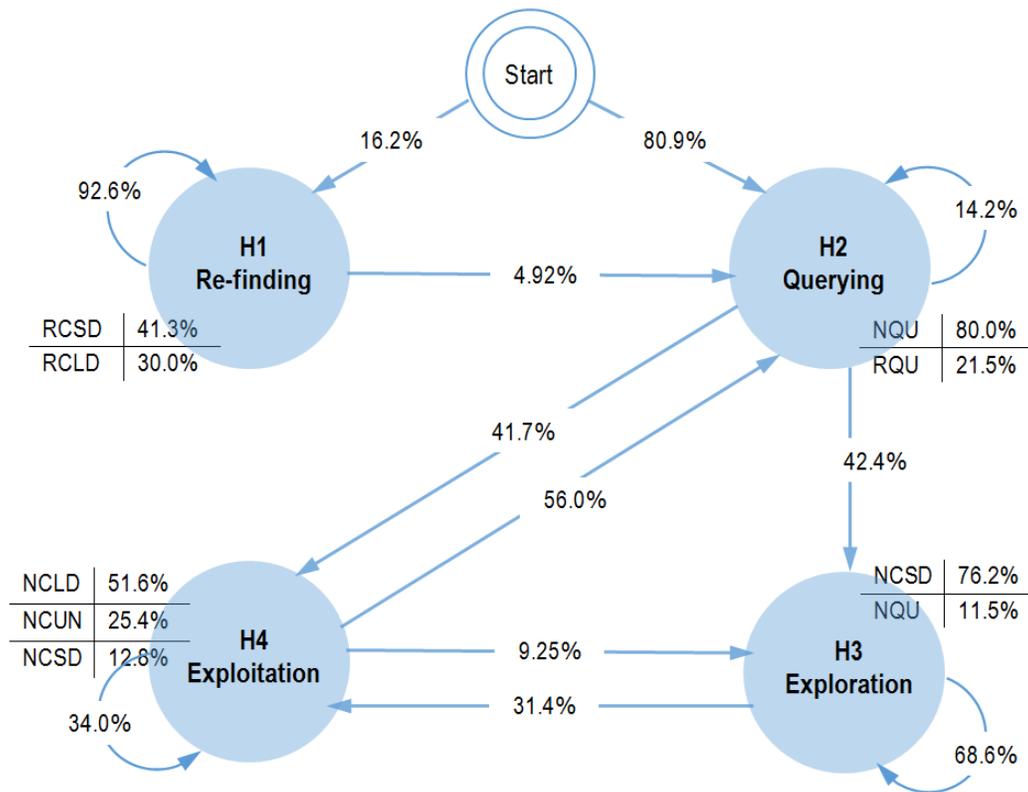


Figure 19: HMM results for the continued session of D-D.

5.1.4 Validating HMM Outputs

To understand the validity of the above HMM outputs, I further conduct several follow-up studies 1) to compare the consistence of HMM results with the findings reported from prior studies with similar topics [95] (a qualitative analysis); and 2) to examine whether such output can be applied for supporting cross-device search tasks (a quantitative analysis).

5.1.4.1 A Qualitative Analysis. When studying hidden behavioral patterns in a long search session, Luo et al. [95] defined four states with two dimensions (2-by-2 states): 1) the “relevant” dimension, describing whether a user finds a *relevant* or *non-relevant* document, and 2) the “exploration” dimension, denoting whether a user’s intention is to *explore* a new topic or *exploit* the current topical information. They found that *Exploration* is the dominating state for non-relevant documents and *Exploitation* is the dominating state for relevant documents. My HMM results (in Figure 18) are consistent with these findings, where H3 corresponds to the exploration state for short-dwell clicks (which usually indicates that a non-relevant or marginally relevant document is found), and H4 corresponds to the exploitation state for long-dwell clicks (which often implies the document is relevant).

In addition, Chapter 4 has identified that information re-finding for task restoration and sense-making for task advancement are the two most important search strategies participants adopted in the continued search sessions. Here, Figure 18 also discovered the corresponding hidden states — the re-finding state (H1) and the exploration & exploitation states (H3 tends to explore and select more diverse information, whereas H4 attempts to scrutinize certain new information in-depth). Meanwhile, I plot the temporal distribution of each search strategy over the whole search session in Figure 20 (X-axis represents the Nth minute for each session, and Y-axis denotes the percentage of such strategy in the whole search session). It clearly illustrates that re-finding and querying are more likely to occur at the beginning of continued sessions. Around 40% to 50% of H1 and H2 occurred within the first minute of the continued search session. This shares the same conclusion with my finding on information reuse in Section 4.2.2, and the conclusions in Tylor et al. [137] based on a large-scale analysis of Microsoft Bing’s search logs.

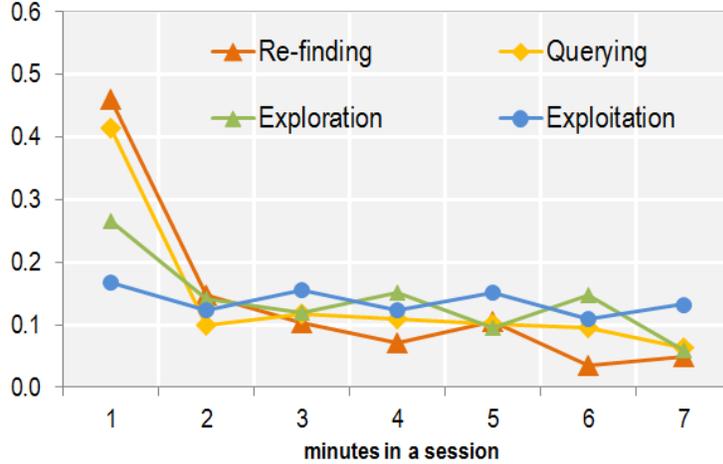


Figure 20: Temporal distribution of different search strategies in continued sessions.

5.1.4.2 A Quantitative Analysis. The above analysis demonstrates that HMM indeed uncovers important search patterns from observed user behaviors, and thus can be used for long-session search behavior modeling. However, there still lacks a clear message about whether and how HMM outputs can be used to support real-world cross-device search tasks, which is the focus of this section. Given the importance of information re-finding behaviors in cross-device search task resumption, this section attempts to apply HMM results for predicting information re-visitations. More specifically, I am trying to predict documents from the first session that are likely to be re-visited in the continued sessions. The following paragraphs describe the task setting for the M-D dataset at first, and then the same procedure is applied to the D-D dataset.

A Quantitative Analysis with HMM Results for M-D. The HMM model in Figure 18 cannot be directly applied for predicting information re-visitations because this model utilized user behaviors from continued sessions to infer the information reuse behaviors, which are unavailable if one wants to predict user behaviors in the continued session from the first session. Therefore, the information reuse behaviors such as RQU, RCLD, RCSD and RCUN (in Table 4) cannot be employed. To apply the HMM, I define a new categorize with four behavioral types for users’ search behaviors — query (QU), long-dwell click (LD), short-dwell click (SD) and click with unknown time (UN). Follow the above HMM process,

I obtain a new model (with 3 as the optimal number of hidden states) as shown in Figure 21. Again, labels for the three hidden states are manually crafted based on their emission and transition probabilities.

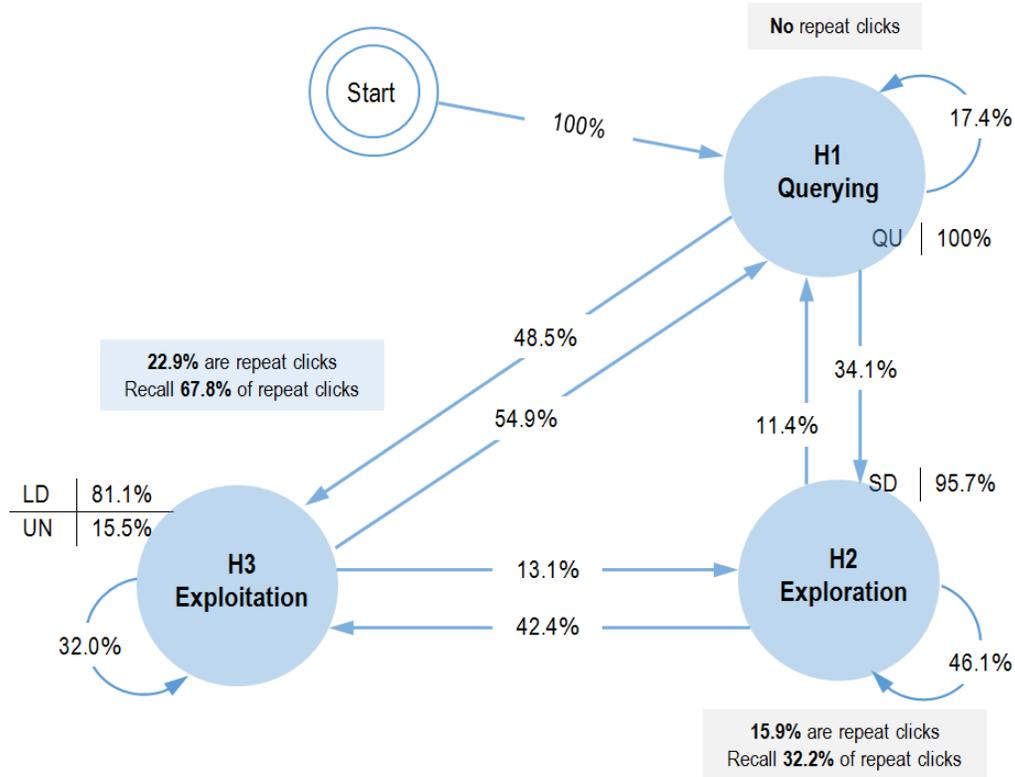


Figure 21: HMM results for the first session of M-D search.

Similar to the HMM model defined in Figure 18, the new model also detects querying (H1), exploration (H2) and exploitation (H3) as three important hidden states. Comparing to Figure 18, there are more querying to exploitation transitions (H1→H3), more exploration to exploitation (H2→H3) and less querying to exploration (H1→H2) transitions, indicating that participants in the second session tend to be more selective for relevant documents, probably because of that they have already explored several relevant documents in the first session, making the discovering of novel and relevant documents more difficult. Meanwhile, the new model does not involve an information re-finding state since users remain to have

fresh memories during a short-time period.

To understand the utility of this new HMM model, I conduct one further experiment to examine whether such model can help predetermine (before the second session starts) what search content will likely to be reused. Specifically, I assume that the hidden state is a good indicator for content reuse; therefore, my first experiment predicts web documents associated with each hidden state as the likely to-be-repeated resources. This experiment only focuses on predicting repeated documents because of its relatively higher percentage in terms of user behaviors (compared to query repeat). For repeated query prediction, I will provide more details in the following Chapter 6.

To understand the performance of this experiment, I compute two percentages – the proportion of predicted repeated documents (i.e., all documents under one hidden state) that are truly repeated by users in the continued session, and the proportion of true repeated documents that can be recalled from the predicted documents. These two measures are thus corresponding to *Precision* and *Recall*, respectively. An F1 measure is also applied to leverage both two metrics. To make a comprehensive analysis of the HMM-based approach, I include three baseline models. The *Random* approach randomly predicts a clicked document as a repeat document based on the average probability of repeat clicks over all clicks. The *Short-dwell (Long-dwell)* approach forecasts all short-dwell (long-dwell) clicks as repeat clicks. The results presented in Figure 21, which demonstrates that the HMM-based approach indeed brings positive results — comparing to other approaches, it improves precision, recall and F1. In the following section, I will include the hidden state as an additional feature to predict repeated documents.

A Quantitative Analysis with HMM Results for D-D. To further understand how such approach works for the D-D dataset. I repeat the above procedure and runs a HMM model solely based on the search behaviors of the first session on D-D. The new HMM model is illustrated by Figure 23. Since users performed their first sessions on desktop computers instead of mobile devices in this condition, there are several important differences between two models. For example, there are more querying to exploitation, fewer querying to

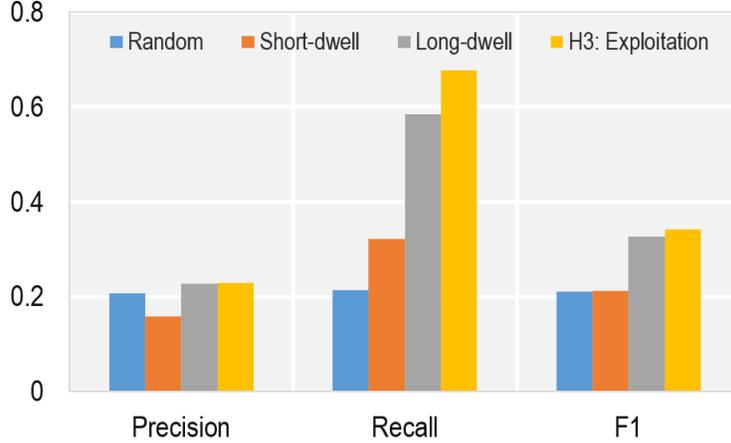


Figure 22: Performances of predicting repeated clicks in M-D for different approaches.

exploration and less exploration to exploitation, implying that users come out relatively high-quality queries, and tend to be more selective (keep exploring till find relevant documents) on desktop computers. To demonstrate the utility of applying HMM model for predicting content reuse information in desktop-to-desktop search, I repeat the same experiment as what I did for Figure 22. The new experimental results are provided in Figure 24. Again, the HMM output can help predict the reuse information with the highest recall and F1 and does not sacrifice precision.

5.1.4.3 Summary. The qualitative analysis confirms the utility of HMM. The hidden search states discovered by HMM not only align with prior studies [95, 137] but also can help accurately predict search content reuse. Both of them confirm the effectiveness of modeling cross-device search behaviors with HMM.

5.2 CONTENT-BASED SEARCH PROCESS MODELING

Several previous studies [26, 38] and my experimental results of Section 4.2 all identified a knowledge learning process during search. However, existing studies only focused on provid-

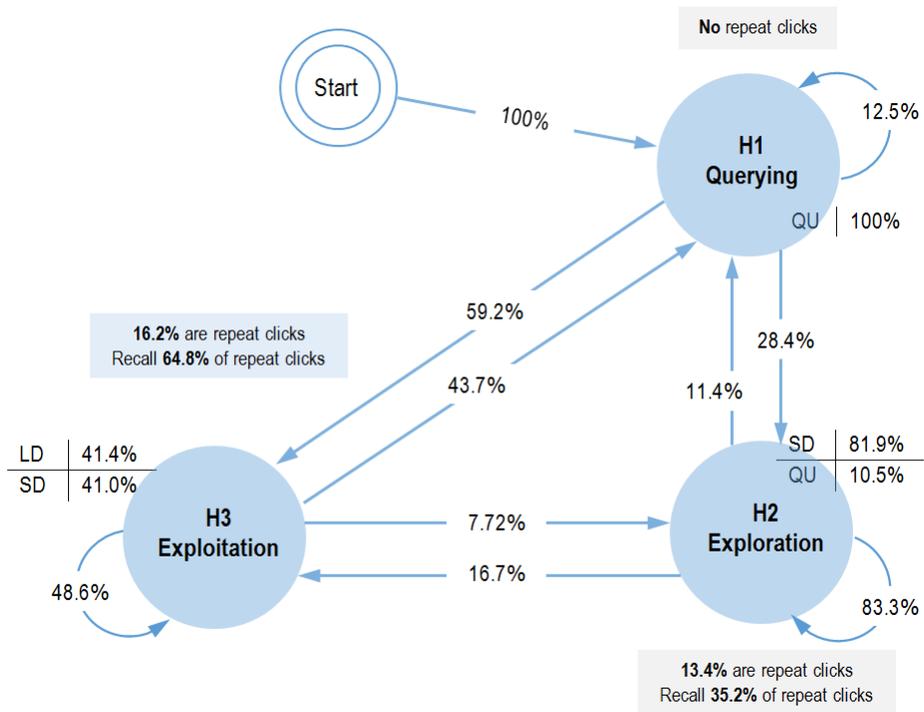


Figure 23: HMM results for the first session of D-D search.

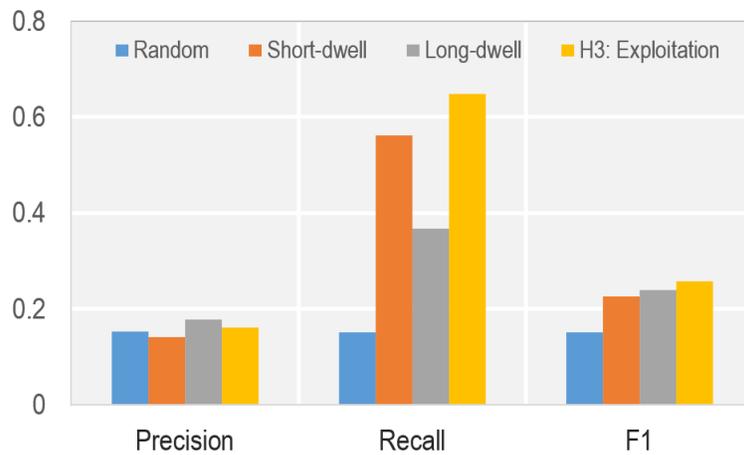


Figure 24: Performances of predicting repeated clicks in D-D for different approaches.

ing a qualitative analysis or a theoretical model to describe the learning process, whereas a quantitative model that captures the underlying knowledge learning mechanism is still missing. Therefore, based on the modeling rationale of existing studies [9, 10, 32], I propose a novel **K**nowledge-based **S**earch **P**rocess (KSP) model for this purpose.

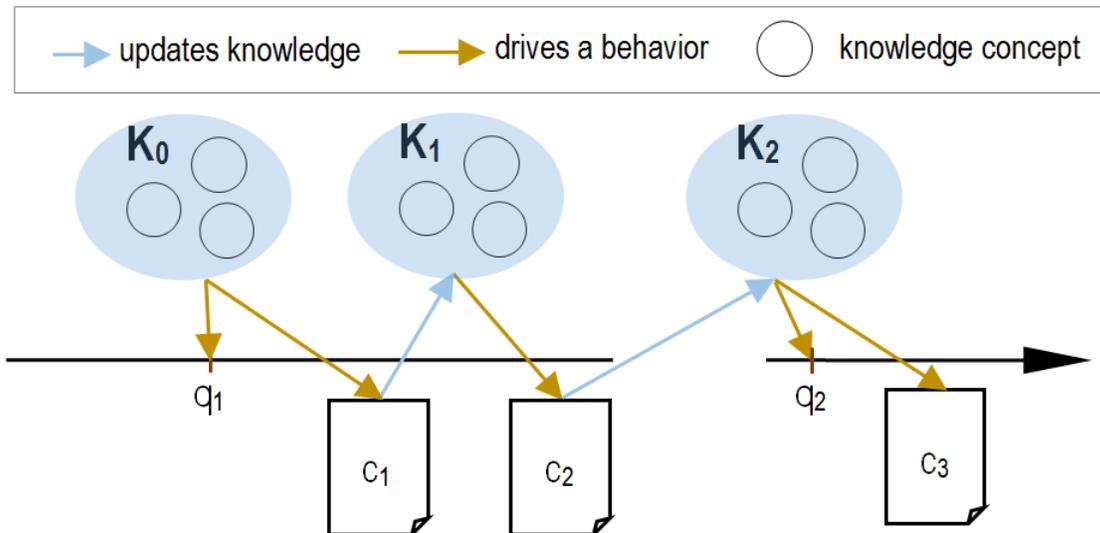


Figure 25: An illustration of the Knowledge-based Search Process (KSP) model.

5.2.1 Modeling Search Process with Knowledge Learning

Figure 25 illustrates the framework of this model, in which q_i refers to a query, c_i represents a clicked document and K_i denotes an overall user knowledge in the current stage. This model assumes that a user’s click and query behaviors are driven by her knowledge status. For example, the issue of q_1 and the click of c_1 are determined by the user’s initial knowledge K_0 , and the query q_2 and click c_3 are driven by the user’s learned knowledge status K_2 .

The KSP model is grounded on several well-known theories from both information science and education domains [9, 17, 32]. Following the ASK (*Anomalous State of Knowledge*) model proposed by Belkin [9], the KSP model views information retrieval as a process of learning new knowledge to resolve the mismatch between a user’s current knowledge and the

knowledge required by the search task. During search, a user can learn knowledge through reading clicked documents (this is consistent with Brookes’s statement [17] in his *fundamental equation* about information and knowledge). The updated knowledge status will then drive users to conduct follow-up search behaviors, including issuing appropriate search queries and evaluating documents for click. Quantifying such process requires implementing three important components — representing user knowledge, updating user knowledge structure and predicting next search behaviors based on knowledge status.

A common **representation of an individual’s knowledge**, as used in Koedinger et al. [80], is to divide knowledge into multiple small concepts and then analyze users’ understanding of each concept. This is a challenging topic from the first step of extracting knowledge concepts from a search task to the final step that aggregates users’ concept understanding to their knowledge. Prior studies have proposed several different approaches such as deriving knowledge units through concept maps [37] or expert-crafted subtask search topics [90]. However, they both require significant human effort and are thus lacking of scalability. The automatic approaches such as topic modeling, key-phrase extraction or matching with Wikipedia entries were also explored in the context of enhancing students’ knowledge learning process [92, 115, 141]; yet, the extraction performances were usually not good enough for use, and the effectiveness of these approaches were not extensively examined in information retrieval tasks. More importantly, the identified knowledge concepts are often not at the right level of granularity.

Therefore, instead of explicitly extracting knowledge concepts, this dissertation represents user knowledge with raw search behaviors — query and click. Specifically, assume that a user has a pre-knowledge K^i about a search task at the stage i . Then, she issues a query q . After spending T_{c_j} time on c_j , her new knowledge can be represented as $K^{i+1} = K^i + f(q, c_j, T_{c_j})$, in which f is a function that converts the triple (query content, click content, dwell time) into user knowledge. The function f will be introduced in the next paragraph. The knowledge growth process will iterate till the reach of last clicked document. One issue of this process is that it did not capture the knowledge growth pattern as shown in Figure 13. This is addressed in the next paragraph.

To model the **knowledge update process** and further compute a user’s understanding

(K_c^i) of content c after reading document c_i , I use the following computing steps learning from a well-studied Bayesian Knowledge Tracing (BKT) model [32] from the education domain. BKT is selected for two reasons: 1) BKT directly models knowledge states and their changes, which are consistent with my KSP model; and 2) as shown in [138], BKT can form a knowledge growth curve as shown in Figure 13, which is aligned with my findings in Figure 12. Specifically, BKT assumes that a user’s new knowledge relies on her prior knowledge (K_c^{i-1}), the knowledge about content c in the current reading document and the probability of converting the reading content into one’s own knowledge, $P(T)$. This process can be illustrated as Equation 5.2. Note that I assume a zero prior about user knowledge, i.e., $K_c^0 = 0$. I further assume that $P(T)$ is proportion with three factors: the similarity between content c and document c_i , the normalized dwell time and the personalized learning rate (i.e., the learning ability of an individual). In this dissertation, I use cosine similarity to compute $S(c, c_i)$, the normalization of dwell time is to make sure both $P(T)$ ¹ and K_c in the range of (0, 1), and the learning rate $P(T)$ is set to 0.6.

$$K_c^i = K_c^{i-1} + P(T) \cdot (1 - K_c^{i-1}), K_c^0 = 0 \quad (5.2)$$

$$P(T) = S(c, c_i) \cdot \frac{\log(T_{c_i} + 1)}{const} \cdot \eta \quad (5.3)$$

The most challenging part for KSP is to develop a principled mechanism that can **drive search behaviors based on users’ current knowledge**. Since the main goal of this dissertation is to support task continuation, here, I only focus on search behaviors in the continued sessions, including both information repetition behaviors and information exploration behaviors (see Figure 18). Thus, the KSP model makes two hypothesis regarding to each type of behaviors. The validity of these two assumptions will be further examined in several follow-up experiments.

- The **information repetition hypothesis** assumes that users tend to revisit the documents with their best knowledge understanding. This assumption is based on the following observations. In general, two types of documents would have the most user

¹The parameter *const* is set to 6.0 in Equation 5.3 to make sure the longest dwell time (7 minutes in the second session and 8 minutes in the first session) will also be rescaled into the range (0, 1).

knowledge. The first kind is the document about the common/core information of a task. Since every clicked document contains full (or partial) core task knowledge, users will have much deeper understanding about such document. Revisit such documents can facilitate users quickly grasp the main content of the task. Another kind could be the hub page [79] covering content for several subtopics, for which documents of related subtopics are aggregated when computing user knowledge. Revisit such document can provide a start point for future exploration. Based on this hypothesis, an effective algorithm should predict documents with the most user knowledge as to-be-revisited documents.

- The **information exploration hypothesis** assumes that users are more likely to adopt (click) a document in their knowledge growth stage, whereas they tend to skip document if related knowledge is saturated (see Figure 13). As a result, with users gaining more and more knowledge, they will tend to be more selective in clicking, and would explore SERPs in more depth. I do find, based on the results of Figure 12(d), that there is an increased ranking positions of clicked documents with time goes on. Meanwhile, I also expect a reversed clicking probability curve as shown in Figure 26. I will provide more details about the validity of this assumption in my later experiments.

5.2.2 Understanding and Validating the KSP Model

Knowledge update process in the KSP model, i.e., Formula 5.2, takes into account both document dwell time and its content similarity with historical clicks. Both are important to determine document relevance [53, 56, 124], and have been widely used in many information retrieval tasks. To start, I provide an initial analysis of the KSP model by correlating the computed user knowledge with document relevance. Here, I am only interested in the documents from the continued sessions; therefore, the correlation results might differ from prior research derived from single-session web search (e.g., the relations between document relevance and dwell time) [53, 56].

For the purpose of comparison, I consider two baselines: the correlation between document relevance and dwell time, and the correlation between document relevance and document similarity with historical clicks. Specifically, dwell time has been recognized as an

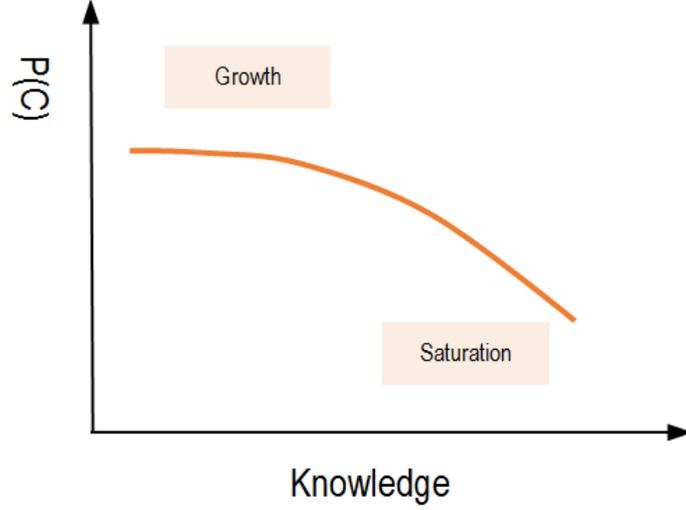


Figure 26: The expected probability of document click $P(C)$ over user knowledge.

important behavioral signal for measuring document relevance and user satisfaction [53, 56]. In my analysis of HMM (see Section 5.1.4), it was also an important predictor of repeat click. Besides, aggregated search history, particularly the click-through history, was commonly used for inferring users' search contexts [49, 55, 68, 124] and further improving search performances. To compute the search context for each individual, I adopted Shen et al.'s [124] approach to estimate an unigram click-through language model (θ_{H_c}) as shown in Formula 5.4. n is the total number of clicks in the first session, $c(w, C_i)$ denotes the count of word w in the clicked document c_i , and $|c_i|$ is the total word count of c_i . Later on, the estimated search context θ_{H_c} will be used for computing user's understanding of a new piece of content. I name these two baseline approaches as **dwell** and **context**, respectively.

$$P(w|H_c) = \frac{1}{n} \times \sum_{i=1}^n p(w|c_i) \quad (5.4)$$

$$P(w|c_i) = \frac{c(w|c_i)}{|c_i|} \quad (5.5)$$

The following subsections start with the understanding of the connections between the

computed user knowledge with document relevance. Then, I provide a detailed analysis and examination of the above-mentioned two hypotheses — the information repetition hypothesis and information exploration hypothesis.

5.2.2.1 Understanding User Knowledge in KSP. For each piece of content c in the continued session (e.g., the snippet of a document in an SERP), I can apply Formula 5.2 to compute a user’s understanding of c based on her historical clicks and page dwell time. To understand the correlation between document relevance and user knowledge in the continued session, I need to first obtain document relevance. Here, relevance is computed through aggregating all participants’ relevance judgments (see Formula 3.1 for computation details). Then, I calculate user knowledge of each clicked SERP snippet in the continued session². Meanwhile, I also compute the cosine similarity between content c and the user’s search context (θ_{H_c}) and assume larger similarity would indicate a higher document relevance. This refers to the baseline *context*. Another baseline *dwell time* is calculated with post-task analysis of the time spent on a webpage. The Spearman correlation results between user knowledge (or any baseline) and document relevance under each search condition (M-D or D-D) are provided in Table 5 in which ** indicates a significance at p-value < 0.01 level.

Different from prior studies [53, 56], Table 5 does not present a significantly positive relationship between dwell/context and relevance. This is mainly due to that prior studies only looked at the short-time single-session search, whereas my experiment emphasizes on exploring later search stages during a long search process. KSP-based user knowledge works effectively and highly correlates with the document relevance in both M-D and D-D, indicating the utility of modeling user knowledge with KSP. This is an interesting finding since it can potentially help predict document relevance even before a user click.

It is worth noting that the relevance here is measured by its overall topical relatedness (i.e., topical relevance) with the current task, but not consider user context. A user may not necessarily click a topically relevant document once she already knew it (or in the knowledge saturation stage as shown in Figure 13). I will leave such analysis for the next few subsections. This section identifies the connections between user knowledge and document relevance, the

²I ignore non-clicked search snippet since it has not dwell time.

next two subsections will focus on examining the information repetition hypothesis and information exploration hypothesis.

Table 5: Spearman correlation between topical relevance and several other measures.

	dwell	KSP knowledge	context
document relevance (M-D)	-0.016	0.125**	0.072
document relevance (D-D)	0.078	0.181**	0.053

5.2.2.2 Validating Information Repetition Hypothesis. The information repetition hypothesis states that users tend to re-visit the documents (from the first session) with relatively high user knowledge values, i.e., a user would have more knowledge understanding about the revisited documents than the non-revisited ones. Here, revisit or non-revisit behaviors are determined by users’ true repeated click behaviors in the continued sessions. To validate this hypothesis, I first compute user knowledge of each first-session click c_j before the second session starts since the future support would happen at this time. User knowledge of c_j , K_{c_j} , is computed using Formula 5.6, which considers all of the n clicks in the first session because each click may contain certain amount of knowledge related to c_j . The similarity between each clicked document c_j and user search context (θ_{H_c}) inferred from the first-session clicks will also be included as a baseline. Here, I assume users tend to revisit the documents with the highest similarities with their search contexts. Similarly, dwell time serves as another baseline.

$$K_{c_j} = \sum_{i=1}^n \left(K_{c_j}^{i-1} + P(T) \cdot (1 - K_{c_j}^{i-1}) \right), \text{ where, } K_{c_j}^0 = 0 \quad (5.6)$$

The comparisons of different approaches are provided in Table 6, which compared the average user knowledge (through the KSP model), dwell time and content similarity with search context for either repeated clicked documents or non-repeated clicked documents. In

Table 6, * (and **) denotes a statistically significant comparison between repeat and non-repeat clicks, at the p-value ≤ 0.05 (0.01) level. Results in this table reveal that the average user knowledge for repeat clicks is indeed significantly higher than the non-repeated clicks in both M-D and D-D, whereas baseline models did not exhibit any statistical significance. These results confirm my first hypothesis, and thus can be used to effectively predict the information re-finding behaviors. I will discuss more details in the next chapter.

Table 6: Mean (S.D.) of several measurements under different conditions.

Condition	Measurement	Repeat clicks	Non-repeat clicks
M-D	KSP knowledge	0.8862 (0.1395)**	0.8664 (0.0936)
	Dwell(unit: seconds)	36.350 (27.280)	37.154 (19.419)
	Context	0.6837 (0.1638)	0.6846 (0.1028)
D-D	KSP knowledge	0.9105 (0.0876)*	0.8848 (0.0627)
	Dwell(unit: seconds)	28.138 (23.249)	26.318 (14.100)
	Context	0.6330 (0.1499)	0.6639 (0.1029)

5.2.2.3 Validating Information Exploration Hypothesis. The information exploration hypothesis aims to uncover the underlying mechanism for user clicks. Under this hypothesis, there would be a high document click probability $P(C)$ at the knowledge growth stage, while a low probability at the knowledge saturation stage as illustrated in Figure 26. To validate this hypothesis, I plot users’ true click probabilities over different knowledge levels. Specifically, for each examined SERP snippet (here, I assume that a user will examine all of SERP results that are ranked above the position of the last click) in the continued session, I compute the corresponding user knowledge and categorize them into different knowledge levels. Since some snippets (each is associated with a document) are clicked while others are not, I further calculate the percentage of clicked snippets under each knowledge level ($\#$ snippets clicked under one knowledge level/ $\#$ total snippets user read under one knowl-

edge level). The true click probability distribution over different knowledge levels is provided in Figure 27, which shows a concave curve for both M-D and D-D.

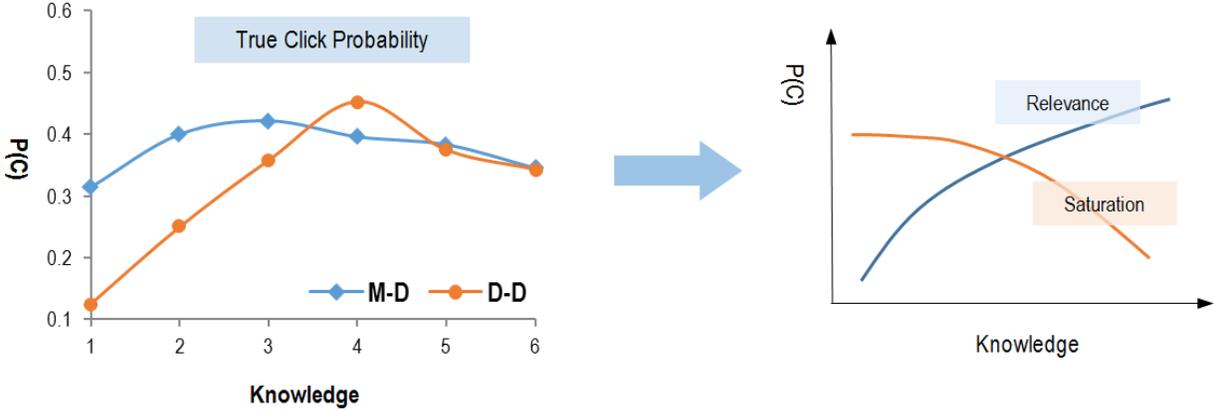


Figure 27: True click probability distribution (left), and its related factors (right).

However, this curve shape differs from my expectation in Figure 26, indicating that other factors also affect clicking. Prior studies [24, 28, 33] on click model identified two important factors that determine user clicks — ranking position and document relevance. In the above analysis, I only study user-examined SERP snippets (by assuming that users will read all SERP results before the last click), which reduced the ranking bias. Therefore, low document click probability at the low knowledge level might be related to document relevance. Indeed, my previous analysis identifies that user knowledge is a positive predictor of document relevance. The synergy effect between document relevance and knowledge saturation can be simply illustrated through Figure 27. In addition to that, I also identify an interesting pattern between different search condition M-D and D-D. Document click probability on M-D is larger than that on D-D, particularly at low knowledge levels. Since my analysis only targets on the continued session, it might be due to that mobile users on the first session do not explore enough documents. Therefore, they still need to read unknown (which have low knowledge levels) but relevant documents in the continued session.

Given the compound effect between document relevance and knowledge saturation, I assume that once controlling document relevance, knowledge saturation would be the major

effect. That is, given a relevant document, a user would click it if she has not obtained sufficient knowledge about it; otherwise, she may skip it. To understand the validity of this hypothesis, I conduct the following experiment. For each query q in the continued session, I compute user knowledge and relevance of each search result in the corresponding SERP. The user knowledge is computed in the similar manner as Formula 5.6 except including additional documents from the second session before issuing q . Document relevance is computed in the same way as Han et al. [51, 55]. After that, I calculate the average user knowledge for both relevant and non-relevant documents (relevance ≥ 3.0 is defined as relevant document; otherwise, it is a non-relevant document), and further differentiate them with click or not. Table 7 provides the comparisons between clicked and non-clicked documents with two approaches, and under two search conditions. Here, I did not include the dwell time as a baseline since non-clicked documents have no dwell time.

The results in table 7 confirm my assumption — for relevant documents, users tend to click those with lower user knowledge (likely to be in the knowledge growth stage), whereas it does not hold (even in the opposite) for non-relevant documents. Also, such effect is less obvious on M-D, which is not surprised since the click probability curve is flatter for M-D in Figure 27. In addition, this is only observed for the user knowledge metric, but no significant effects for search context-based metric. Again, it demonstrates the usefulness of KSP model. Note that ** (*) in Table 7 indicates a significant difference between clicked documents and non-clicked documents at p-value ≤ 0.05 (0.01) level. The statistical significances are examined through Wilcoxon signed-rank test since my data is not normally distributed.

5.2.3 Summary

This section proposed and described a Knowledge learning-based Search Process (KSP) model to capture users' search content change in the continued search sessions. KSP roots from several well-known theories from both the information seeking community and the education domain, which combines users' historical search content, click behaviors, as well as the knowledge learning patterns into one unified framework. KSP made two important assumptions regarding the two major search behaviors in the continued sessions — the

Table 7: Mean (S.D.) of several experimental metrics over different search conditions.

Relevance ≥ 3.0	Non-clicked documents	Clicked documents
Context (M-D)	0.5672(0.1344)	0.5528(0.1405)
KSP knowledge (M-D)	0.7622(0.2016)	0.7399(0.1948)
Context (D-D)	0.5911(0.1137)	0.5715(0.1348)
KSP knowledge (D-D)	0.8329(0.0.1390)	0.6330(0.7973)**
Context (both)	0.5791(0.1249)	0.5619(0.1379)
KSP knowledge (both)	0.7973(0.1767)	0.7679(0.1815)*
Relevance < 3.0	Non-clicked documents	Clicked documents
Context (M-D)	0.5181(0.1284)	0.5268(0.1406)
KSP knowledge (M-D)	0.6843(0.2333)	0.6794(0.2502)
Context (D-D)	0.5359(0.1297)	0.5693(0.1586)
KSP knowledge (D-D)	0.7373(0.1903)	0.7535(0.1729)*
Context (both)	0.5264(0.1291)	0.5471(0.1505)
KSP knowledge (both)	0.7089(0.2157)	0.7149(0.2193)*

information repetition behaviors and the information exploration behaviors. Through an extensive experimental result analysis, I find that the KSP model-based user knowledge is better aligned with document relevance, and the KSP model is demonstrated to be effective in predicting repeat clicks and differentiating non-clicked documents with clicked documents in SERPs. In the next chapter, I would like to explore the ways of applying such model for supporting cross-device search tasks.

5.3 ANSWER TO RQ2

This chapter answered the RQ2 by proposing two quantitative search process modeling approaches. Here, I list the answers to both sub-questions of RQ2.

- **Answer to RQ 2.1.** Section 5.1 proposed a two-layer behavior-based search process model, assuming that users’ observed search behaviors (the observed layer) are driven by their hidden search tactics (the hidden layer). Based on that, a standard HMM was applied for model construction. The answer to RQ 2.1 is positive since HMM outputs are aligned with several prior studies [95, 137] and participants’ self-reported search strategies (Section 4.2.3). In addition, HMM output was further proven to be effective when locating repeated clicks. However, there are no significant differences of HMM outputs between M-D and D-D.
- **Answer to RQ 2.2.** Section 5.2 proposed a hidden variable model to capture users’ search content change during the whole search process, in which the hidden factor refers to users’ knowledge states and the observations are user clicks. It is users’ current knowledge states that drive them to click or not. I named it as the Knowledge-learning based Search Process model (KSP). The answer to RQ 2.2 is also positive because the KSP model is effective in differentiating repeated clicks with non-repeated ones, and can also predict document clicks.

To compare M-D and D-D, I do observed several differences. Results from Table 5 and 7 both revealed that KSP model performed better on D-D. This might be due to several factors: 1) users in the first session of D-D explore more documents so that the model can be more robust; 2) the knowledge saturation assumption works better when users are exploring novel information instead of actively looking for relevant documents. Since users in the continue session of D-D have more sense-making activities, it might work better.

6.0 SUPPORTING CROSS-DEVICE WEB SEARCH

Based on the understanding of cross-device web search behaviors (Chapter 4) and the two proposed search process models (Chapter 5), this chapter moves forward to explore ways of supporting cross-device web searches using these outputs. Again, the cross-device search support mainly focuses on the continued sessions.

This chapter is organized in the following way. Section 6.1 starts with describing two important types of cross-device search behavior patterns. Based on that, I develop two corresponding search support functions (§6.1.1). Then, a prototype system is designed to integrate both functions (§6.1.2). Implementation details about the system are provided in §6.1.3 and §6.1.4. To understand the effectiveness of the developed system, §6.1.5 designs a set of off-line experiments (simulation study) and an on-line experiment (live user study). The detailed evaluation results are then provided in Section 6.2 (off-line evaluation) and Section 6.3 (on-line evaluation).

6.1 DESIGNING CROSS-DEVICE SEARCH SUPPORT

6.1.1 Designing Cross-Device Search Support Functions

Chapter 4 identified two important types of user behaviors in the continued sessions — the information re-finding behavior particularly in the beginning of a continued session, and information finding behavior (i.e., exploring novel information) in later stages of the continued sessions. The HMM-based search process model provides more evidences for these behavioral patterns since it automatically recognizes information re-finding (H1 in Figure

18) and information finding (exploration H3 and exploitation H4 in Figure 18) as important hidden states. Based on that, I design two cross-device search support functions, each attempts to facilitate one type of user behavior.

6.1.1.1 Function I: Supporting Information Re-finding. According to the analysis of search content change in Section 4.3, user knowledge drops (statistically significant) after a long-time interruption. As a result, the major goal of the information re-finding support is to help users quickly restore their previous knowledge so that they can resume their search tasks immediately. A straightforward support is to directly identify and display the most-likely reused search content to end users at the time when they plan to resume the task in continued sessions. Since most of the information re-finding behaviors occur at the beginning of a continued session, such support should be provided at this stage.

In addition, my analysis at Chapter 4 also identifies a certain amount of re-finding behaviors at later stages of the continued sessions, which also requires proper search support. Since users are often in active searching mode during the later stages, the re-finding support should minimize the interruption of users' normal search processes. These design rules are adopted when designing the information re-finding support system, which will be described in the next few sections.

6.1.1.2 Function II: Supporting Information Finding. Besides re-finding, most of user behaviors in the continued session are finding novel and relevant documents. Existing search engines seldom customize their results based on re-finding or finding so that searchers often encounter repeated content particularly for long-time search sessions [55, 69]. To understand how repeated information occurs in my user study I, I compute the percentage of information repeat for the top three SERP results in the continued sessions. Here, repeat means that a document either occurs in the top three SERP results of prior sessions, or was clicked in the previous session. Table 8 provides the percentages for three tasks under two search stages (1st-minute stage and rest-minute stage). I divide the whole search process into these two stages because information re-finding occurs the most in the first minute (see Figure 20).

Table 8: Repetition percentage for top three SERP results in different conditions.

Stage \ Tasks	NE	PD	PE
The 1st-minute stage	18.17%	32.99%	50.58%
The rest-minute stage	14.31%	29.41%	41.13%

According to Table 8, searchers do encounter a high proportion of repeated results on SERPs. Same as the 1st-minute stage, the rest-minute stage also has a high percentage of repeating behaviors (around 30%) even though users’ major intents, in this stage, are to explore novel documents. Moreover, since users in the continued sessions tend to explore more in-depth on SERPs (clicking documents on the bottom part of an SERP and issue more complex queries to find a different set of search results, as shown in Figure 12), it is obvious that the current topical relevance based result ranking has led to more efforts for searchers to filter out repeated documents. To support this type of search behaviors, I envision a search system that 1) can automatically detect users’ information finding state and then re-rank existing search engine results; and 2) can also provide useful information to indicate whether or not certain documents have been accessed before.

6.1.2 Building a Cross-Device Search Support System

This section designs a prototype system to integrate both functions as I mentioned above. The information re-finding support requires to provide information about users’ search histories at both the beginning and later stages of the continued sessions, and the information finding support calls for providing information to differentiate novel documents with repeated documents. Based on these two design principles, I build a prototype system as shown in Figure 28. The system aims to facilitate both information re-finding and information exploration behaviors in the continued search sessions. Besides the search box, a ranked click history ① is also offered in the index search page. Once a user issues a query, the search

results will be re-ranked as shown in ③ based on the system’s intelligent analysis of search intentions. The previously-clicked documents will also be displayed with a search result reminder, i.e., ②.



Figure 28: An illustration of the prototype system for cross-device search support.

Specifically, I design two modules to support information re-finding. First, instead of

solely displaying a search box in the index page, I also provide a ranked click history based on the likelihood of a document to be re-visited in the continued session (see the algorithm in Section 6.2). For each document, I list its title (with URL as a clickable link), accessed device and the most recent related search query. I expect that such information can help users recall their previous knowledge about the task topic. Second, during the process of exploring SERPs, I provide a result visitation reminder, as shown in Figure 29, to make users aware of their past interactions with a specific search result. Similar to the click history, here, I display its accessed time, accessed device and the associated query. If a user remains in the re-finding stage, such metadata information would help her recall previous knowledge.

The result visitation reminder is also helpful for information finding since it informs a user that the corresponding document was accessed before so that the user can allocate more efforts on exploring other documents. In my post-task interview for user study II, almost 80% of the participants stated that they tend to avoid clicking the document if there was a result reminder. In addition to the reminder, my system also provides the support for finding relevant but novel documents through re-ranking search engine results. An example of the re-ranked results can be illustrated in Figure 28. Since the user has already clicked the Wikipedia page (thus, it is not novel), my algorithm downgrades its importance and ranks it to position two in the SERP.

Providing the result visitation reminder is easy because it only requires a simple analysis of search histories. Re-ranking click history and SERPs in Figure 28 require developing intelligent result re-ranking algorithms. I will provide more details in the next two sections.



Figure 29: An illustration of search result reminder.

6.1.3 Re-ranking Click History to Support Re-finding

To provide a ranked click history in the index search page, I first extract users' click history from search logs, and then re-rank them based on their probabilities of re-visitation¹. Formally, I define the re-ranking task as following. Supposed that a user u is working on a cross-device search task. At time t , she needs to re-access certain clicked documents to recall her prior knowledge. Her search history, up to time t , is represented as \mathcal{S}_t , which consists of a list of m click triples. Each click triple $\mathcal{C}_i = (c_i, \Delta t_{c_i}, q_{c_i})$ includes a clicked document, the corresponding page dwell time and its associated query, respectively. Therefore, $\mathcal{S}_t = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_m\}$. The next step is to re-rank each \mathcal{C}_i based on its probability of re-finding.

I treat the re-ranking task as a learning-to-rank problem, and my goal is to learn a model to predict the historical documents that are likely to be revisited in the future, i.e., predicting the information re-finding behaviors. The re-ranking candidates are clicked documents before the current time t . The learning-to-rank algorithm is a supervised model; thus, requiring 1) ground-truth labels about re-visitation or non-revisitation; and 2) a set of ranking features. The ground-truth labels are obtained through users' true search behaviors in the future (i.e., the continued search session in my experiment setting). The ranking features are extracted from search behaviors in \mathcal{S}_t . In this dissertation, I consider a list of four features as described in below Table 9. My later experiments are based on either each of these features or their combinations. I will provide more details about each specific ranking model when used. For model selection, I choose LambdaMART [149] as the core model to build a pairwise learning-to-rank algorithm when ranking search history. Note that, here, I do not train a model per each individual or per each task, and then predict based on each model since the data for each individual/task is too sparse.

Table 9 lists four features — HMM and Knowledge are inferred from my proposed behavior-based (Section 5.1) and content-based (Section 5.2) search process models, and the rest are used as baselines. The HMM results are obtained from user behaviors of the first session. Here, I assume that if a clicked document c_i is in the Exploitation state, it will be

¹Here, I do not rank and display search queries because 1) query is usually too short for users to recall their search history; and 2) users might look for different information even under the same query. However, I do display associated search queries for each clicked document.

Table 9: Features to predict repeated click for any document c_i .

Features	Description	Feature type
Dwell	the time a user spends on c_i	Continuous
HMM	whether c_i is in a Exploitation state (i.e., H3 in Figure 21)	Nominal
Context	similarity between c_i and the user’s search context (§5.2.2)	Continuous
Knowledge	a user’s current knowledge about c_i	Continuous

more likely to be revisited in the future; thus, the HMM feature refers to a binary indicator about whether c_i is for Exploitation. The Exploitation state is chosen based on my previous analysis in Section 5.1.4.2. Knowledge feature is computed based on Formula 5.2. Dwell and context were introduced in Section 5.2 when examining the utility of the knowledge-based search process model. I acknowledge that the above features are far from enough to cover a complete set of behavioral signals for this ranking task. Here, the main purpose is to demonstrate the utility of the two proposed search process models; thus, employing a more comprehensive feature set is left for future.

6.1.4 Re-ranking Search Engine Result Pages

The other important function in Figure 28 is to re-rank search results in SERPs. Since users’ search states (search purposes) can either be looking for novel documents (if in the information finding state) or seeking for repeated documents (if in the re-finding state) in an SERP, the re-ranking algorithm should first detect users’ search states and then customize result ranking strategies for different states — it should prioritize repeated documents for re-finding but emphasize novel documents for finding. Therefore, the following subsections start with detecting search states and then describe the implementation of the re-ranking algorithm.

6.1.4.1 Detecting Search States. According to my analysis in Chapter 4, there are mainly two types of search states in the continued session of a cross-device web search process — finding and re-finding, which can be easily detected through a HMM. Previous chapters show that these two search states are only associated with clicked documents, whereas the SERP re-ranking requires to know the search states for queries. This is a challenging topic since querying only corresponds to the state query in HMM model outputs. Fortunately, after computed the percentages of finding/re-finding clicks under all search queries in the continued sessions, I find that 97.7% of queries are for finding and only 0.3% of queries are for re-finding. This is because of that 1) people might have already completed the information re-finding stage before searching any query; and 2) participants may not necessarily need to click any document for re-finding, reading search snippet could be sufficient.

Based on these facts, I assume that all search queries in the continued sessions are to find novel and relevant information. Therefore, the re-ranking algorithm needs to properly leverage both document novelty and document relevance. In the following subsection, I will discuss how such re-ranking model is implemented through the knowledge-based search process model. The HMM model is not applied here because there are available behavioral information for novel documents (since they are not clicked yet).

6.1.4.2 SERP Re-ranking based on the KSP Model. According to the analysis in Figure 12, participants explored SERPs much deeper in the continued sessions, indicating that users often need to spend more efforts, compared to the first search session, on locating useful documents. As a result, the main goal of the re-ranking algorithm is to minimize such effort by re-ranking useful documents to the top.

Formally, given a search query q and its default result ranking (from Google) $\mathcal{R} = (R_1, R_2, \dots, R_m, \dots, R_n)$. In most of the cases, users do not examine all of the n results and only click few of them [33]. I further assume a linear reading pattern for search engine results examination so that the user reads through (not necessarily clicks) all of the above results until reaching the m -th one (R_m). The linear reading assumption is a major component for many click modeling studies [28]. An SERP re-ranking algorithm attempts to re-arrange search results to minimize the examination effort and click effort.

The core component of implementing a re-ranking function is to compute the probability of clicking a search result $c_i = 1$ given one has examined the document $s_i = 1$, i.e., $p(c_i = 1|s_i = 1)$, and then readjust SERP results based on this probability. My previous analysis in Section 5.2.2 has plotted this probability over different levels of KSP-based user knowledge. The plot shows that this probability is a concave function of user knowledge (see Figure 27). Here, I summary it using Formula 6.1. The next step is to estimate $f(K_{s_i})$.

$$P(c_i = 1|s_i = 1) = f(K_{s_i}) \tag{6.1}$$

A principled way of estimating $f(K_{s_i})$ requires knowing its theoretical function format and then determine associated parameters from data. Purely based on the curve shape in Figure 27, it is hard to assess the true function format. As a result, I adopt a simple method that divides user knowledge levels into several small bins, and then estimate the click probability for each bin based on Maximum Likelihood Estimation (MLE). The bin size δ should not be too small or too large. If too small, there will be insufficient statistics for estimation; if too big, there would be no differences among different bins. For each bin b_k , the click probability is estimated by the number of clicked snippet over the number of examined snippets (both clicked and non-clicked) with user knowledge in this range, as shown in Formula 6.2. Here, the examination of a search snippet is determined by users' real click behaviors, for which I assume that a user will examine all of the search results above the rank of the lowest and clicked document (linear reading assumption). Note that these probabilities will be estimated from the training data. I will provide more details about training-testing split in the following sections.

$$P(c_i = 1|s_i = 1, K_{s_i} \in b_k) = \frac{||c_i|c_i = 1, K_{s_i} \in b_k||}{||s_i|s_i = 1, K_{s_i} \in b_k||} \tag{6.2}$$

The estimated probabilities will then be used to re-rank search results in the testing dataset. During re-ranking, I first compute user knowledge for each search result snippet and then look up its corresponding probability based on the knowledge level. Since two documents might have the same knowledge levels (a tie), I will pick the one ranked lower in an SERP once encountering a tie. In this way, it can better support user exploration of

novel documents².

In addition to the KSP-based approach, I also include the search context-based result re-ranking as a baseline. To make a fair comparison, the search context-based approach also adopted the same MLE estimation approach³.

6.1.4.3 Connection with Existing Studies. The main purpose of the proposed KSP-based SERP result re-ranking algorithm is to help users explore novel information in continued sessions. This corresponds to several important research topics in information retrieval community such as novelty/redundancy detection [4, 159] and search result diversification [22, 112]. Although sharing similar motivations, the proposed algorithm also significantly differs from prior studies in the following aspects. When determining novelty/diversity of a search result, previous studies (e.g., the Max Margin Relevance model in [22] and the redundancy detection approaches in [159]) only considered the content similarity between the search result and search history. The KSP-based approach, on the other hand, took into account both document content similarity and the document reading time. More importantly, this algorithm explicitly models users' knowledge growth processes based on a theoretical framework about human knowledge growth [32]. Search result ranking is then determined by users' current knowledge status. In the case of knowledge growth, documents similar to the current knowledge are preferred, whereas such documents will be demoted if user knowledge has already saturated. This perspective was never explored in existing studies.

6.1.5 Evaluating Cross-Device Search Support

The simplest way of examining the effectiveness of the above-mentioned cross-device search support functions is to run an **on-line** evaluation with two kinds of systems, one implements the support algorithm while the other does not. Then, the effectiveness of the proposed algorithm can be evaluated based on the comparison of search performance between two systems. The on-line experiment, on the other hand, usually has a high cost and is difficult

²I also experiment the other way by choosing top-ranked documents. My experiment results show that it is worse than selecting the bottom-ranked documents.

³A direct re-ranking of SERP results based on search context not through MLE will lead to much worse performance.

to experiment multiple system setups (e.g., different parameter settings in the proposed algorithm). Therefore, **off-line** evaluations are also adopted in this dissertation to choose the optimal search system setups in the first place.

According to Figure 28, there are two major search support functions in the proposed system interface; therefore, I develop two off-line evaluation experiments. The first experiment attempts to evaluate different settings for the information re-finding support and the second experiment aims to assess the information exploration support. To evaluate the performances for different search support algorithm settings, I need to collect the ground-truth data about users' true information finding and re-finding behaviors. These are obtained through partitioning my user study data temporally (data from User Study I) — suppose that the system can provide search support at time t , user behaviors after t will be treated as the ground-truth. Using this experiment setup, one can quickly test different parameter settings and pick up the optimal ones.

Based on the off-line experiments, I will select the best search support algorithm and implement it in my cross-device search system as shown Figure 28. To understand how real-world users perceive the provided search support mechanism, I further conduct an on-line user experiment (user study II) with certain tasks showing the search support functions (named as experimental system **E**) while other similar tasks masking the functions (named as baseline system **B**). User perceptions/performances under different conditions will be analyzed to understand system effectiveness. The detailed results about this user study will be provided in Section 6.3.

To summarize, I consider two types of evaluation approaches when assessing the effectiveness of cross-device search support — off-line evaluation and on-line evaluation. Both of them will be applied to assess the utility of the provided information re-finding/finding support functions. Moreover, since each supporting function can be implemented using the output either from the behavior-based search process model or the content-based search process model. Theoretically, there will a total $2 \times 2 \times 2 = 8$ search evaluation conditions. However, the HMM-based model cannot be applied to support information exploration as mentioned in Section 6.1.4.1. This will then result in total 6 evaluation conditions. This can be illustrated in Figure 30.

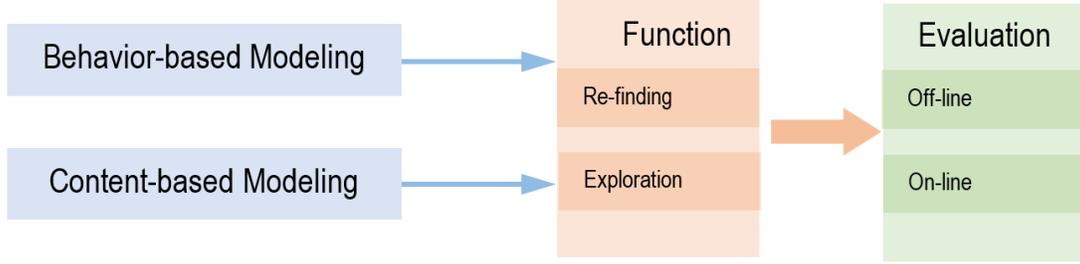


Figure 30: An illustration of the cross-device search support evaluation framework.

6.2 OFF-LINE EVALUATION FOR CROSS-DEVICE SEARCH SUPPORT

This section reports the off-line evaluation results for the two cross-device search support functions. The off-line evaluation utilizes the data from User Study I.

6.2.1 Off-line Evaluation for the Information Re-finding Support

To understand the effectiveness of different feature sets (Table 9) in this task, I experiment the following seven models, each with either one type of features or the combination of multiple features. *Dwell*, *HMM*, *Context* and *Knowledge* stand for each of the four features. *Knowledge + HMM* combines both knowledge and HMM features. *All* considers all of the four features. In addition, since not all continued search sessions include re-finding behaviors, the upper bound performance for re-finding prediction algorithms would not reach to 100%. Therefore, I add an oracle approach by assuming that I can always make a correct prediction if there is a repeated behavior.

6.2.1.1 Model Setup. The model *Dwell* utilizes the page dwell time as its feature, which can be simply obtained from the click tripe defined in Section 6.1.3. The model *context* employs the search context information computed through Formula 5.4. The HMM-based model uses the HMM outputs of the first-session search behaviors⁴. When computing user

⁴Users' search behavioral data comes from User Study I, and the number of optimal hidden states is set to 3 as it produces the best BIC value.

knowledge for the knowledge-based model, I set users’ initial knowledge about any content c as zero, i.e., $K_c^0 = 0$, the learning rate η as 0.6 and the normalized *const* as 6.0.

To compare the performances of the different approaches, I adopt a standard precision based measure — P@N. Here, N is set to three⁵. And the reported prediction performances in the following sections are averaged across a 5-fold cross-validation with 50 runs, with a random reassignment of five folds for each run.

6.2.1.2 Experimental Results. The comparison results among different feature sets are provided in Table 10. Numbers in bold indicates the best performance among all approaches. Note that there are three numbers for D-D are bolded because they have no significant differences. * indicates a statistical significance over the best baseline approach (i.e., *Dwell* for M-D and *Context* for D-D) at the p-value ≤ 0.05 level. The significance is tested with pairwised t-test since the P@3 performances are normally distributed.

Table 10 shows several interesting findings. The HMM based approach excels all other models on M-D, while the three knowledge-related approaches, including Knowledge, Knowledge + HMM and All, outperform all other models on D-D. However, they three did not show significant differences. This again demonstrates the usefulness of both HMM-based and Knowledge-based modeling of human search processes, though they might not work differently under different search conditions.

These findings are consistent with my previous analysis on the comparisons between M-D and D-D. HMM results for the first session of M-D (Figure 21) are better clustered than that of D-D (Figure 23) In M-D, H1 generates query, H2 is for short-dwell clicks and H3 for long-dwell clicks, whereas the hidden states for D-D mix up short-dwell clicks with queries in H2, long-dwell and short-dwell clicks in H3. This might account for the reason why HMM only works on M-D rather than D-D. In terms of the knowledge-based model, the model performance difference between D-D and M-D is also consistent with the findings from Figure 27 — there are larger click probability differences among different knowledge levels on D-D. Also, the computed user knowledge in Table 7 was found to have a larger correlation coefficient for D-D than that M-D.

⁵I also run evaluations over P@5 and NDCG@5, they provide similar conclusions.

Table 10: P@3 performances of different approaches under different conditions.

Models	M-D	↑ over Dwell	D-D	↑ over Context
Dwell	0.2156	-	0.1334	-16.8%*
Context	0.2139	-0.78%	0.1605	-
HMM	0.2569	+19.2%*	0.1552	-3.29%
Knowledge	0.2347	+8.89%*	0.2042	+27.2%*
Knowledge + HMM	0.2362	+9.58%*	0.2032	+26.7%*
All	0.2423	+12.4%*	0.2055	+28.0%*
Upper bound	0.4329	+101%*	0.4722	+194%*

6.2.2 Off-line Evaluation for the Information Finding Support

The second search support function attempts to assist users’ information exploration behaviors in continued sessions. As mentioned in Section 6.1.4.2, such support is achieved by re-ranking SERP results based on the estimated conditional probability of clicking $P(c_i = 1 | s_i = 1)$ in Formula 6.2. In this off-line experiment, I only re-rank user-examined search results to align with the conditional probability (i.e., probability of click given user examination)⁶. To understand the effectiveness of my re-ranking algorithm, I conducted the following experiment.

6.2.2.1 Experiment Setup

Sine the main goal of this search support algorithm is to re-rank SERP results so that the user can access useful documents with the minimum efforts; therefore, I evaluate the re-ranking performance using the **average ranking positions** (the smaller the better) of clicked documents in the re-ranked SERPs [38, 49, 158]. In total, I include four different approaches (three baselines) — original SERP ranking (low

⁶User examination is determined based on Section 6.1.4.2. Note that I will re-rank all SERP results in the on-line experiment since user examination behaviors are not available yet.

baseline), random SERP re-ranking (low baseline), search context based SERP re-ranking (high baseline) and KSP-based SERP re-ranking (proposed).

The original SERP ranking is a biased baseline since I am only re-ranking SERP results up to the last clicked document. Suppose that there is only one document clicked in a SERP (at position X), even a random re-ranking (which will rank the document to any position in the range [1, X]) will outperform this approach. Therefore, I include a random re-ranking approach as an additional baseline. Again, the search context is used as a high baseline. Both the search context-based re-ranking and the knowledge-based re-ranking use MLE to estimate click probability (as shown in Formula 6.2). Here, I divide the computed user knowledge (or search context) into six bins⁷.

Table 11: Average ranking positions of clicked documents under different approaches.

Models	M-D	D-D
Original ranking	4.6827(2.6649)	4.2905(2.5796)
Random re-ranking	3.4257(2.1937)	3.1321(2.1671)
Context-based re-ranking	2.8773(2.0782)	2.7620(2.0003)
KSP-based re-ranking	2.7768(1.9975)	<i>2.3894(1.6355)</i>

6.2.2.2 Experimental Results The comparisons among different re-ranking algorithms and different search conditions are provided in Table 11, in which the numbers in bold (italic) indicate a statistical significance over random (search context) baseline. The result significance is examined with Wilcoxon signed-rank test since the ranking positions are not normally distributed. Table 11 shows several interesting results. First, as expected, the random re-ranking of the examined SERP results outperform the original result ranking because of a biased selection of documents for re-ranking. Second, both the context-based and the KSP-based re-ranking algorithms work effectively well since they both reduced the

⁷A larger or smaller number of bins produce much worse results.

original rank positions by 1.5, whereas a random re-ranking only improve the original ranking position by 0.5. The results hold not only for M-D but also for D-D. Third, the KSP-based re-ranking model outperforms the context-based model on D-D but not on M-D though the average ranking positions for them have no (statistically) significant difference on M-D. This again confirms the usefulness of adopting knowledge-based modeling for human search processes. Furthermore, the differences regarding the KSP-based re-ranking between M-D and D-D are consistent with several my previous analysis on the knowledge-base models, which often show a relatively better performance on D-D than that on M-D (see Figure 27, Table 7 and Table 10).

6.3 ON-LINE EVALUATION FOR CROSS-DEVICE SEARCH SUPPORT

The above off-line experiments have already shown the usefulness of behavior-based and content-based search process models for cross-device search support. To understand whether and how these support algorithms would work for real-world users, I conduct an on-line evaluation experiment to compare user performances between the **E**xperimental system and a **B**aseline system. To be consistent with the experiment setting for user study I, and to primarily focus on the cross-device web search, the on-line experiment only takes into account the search condition of M-D.

6.3.1 Experimental System and Baseline System

The **E**xperimental System (**E**) provides two search support functions as shown in Figure 28 — the information re-finding support (click-history re-ranking and result visitation reminder) and the information finding support (SERP re-ranking). The click-through history is re-ranked based on the HMM feature since it generates the best performance in Table 10. The search results are re-ranked based on the KSP-based user knowledge in Formula 6.1.

The on-line evaluation experiment also includes a **B**aseline System (**B**) that does not provide any support, i.e., a system without the click-through through provided in the index

page, without the result visitation reminder and without the search result re-ranking. Here, I do not include the search context or page dwell as baselines for the following reasons.

First, to the best of my knowledge, a cross-device search support system with ranked click history, search history reminder and re-ranked search results has never been examined in previous studies. Therefore, the first-and-foremost step for this on-line evaluation experiment is to tell how user perceive and behave with and without these search supports. Including dwell or search context as baselines would be my future moves to study a more effective way of supporting cross-device web searches.

Second, based on the off-line experiments, I observe that the absolute search performance values, though having statistical significance, are not far away from each other. Such differences might not be able to reflect into real-world user behaviors and produce sufficient increase/decrease on human perceptions regarding to system effectiveness and efficiency.

Third, one potential way of accommodating the above two points is to include a baseline system (raw Google search), an experimental system with my proposed support algorithm (E1) and another experimental system with search context-based search support algorithm implementation (E2). However, including all three conditions will increase the complexity my experiment design and the task load for each user will increase significantly. These reasons motivate me to stick with my design in Figure 7.

6.3.2 Data Collection

6.3.2.1 Overall Description Follow the above experiment design, I recruited 24 participants (13 males and 11 females with an average of 21 year old) from both University of Pittsburgh and Carnegie Mellon University in March 2017. They are all college students with diverse background such as political science, neuroscience, information science, economics, physics, chemistry and so on. When rating their web search experience with a 7-point Likert scale, they all chose pretty high values and the average rating is 5.65. According to the entry questionnaire, these participants searched almost 45% of their queries on mobile phones and 28% of them have follow-up searches on desktop computers. This percentage is increased comparing to the number (18%) reported in User Study I.

In total, the recruited participants issued 987 queries (454 from the first session and 533 from the second), clicked 1,702 web pages (741 from the first session and 961 from the second session). A simple descriptive analysis of this new data collection is provided in Table 12, in which numbers in bold denote the comparisons between the first and second session (*: $p < 0.05$; **: $p < 0.01$). Wilcoxon signed-rank tests are adopted for the statistical significance tests since most of the data is not normally distributed. Most of the statistics are similar to what I obtained in User Study I (see corresponding results of User Study I in Table 2). Similarly, each participant issues more number of queries, clicks more number of documents and stay shorter in the continued search session since participants are staying on a desktop computer, which brings convenience for typing, reading and exploring. Consistent with my previous findings on Table 2, there are no significant differences regarding to query length between different sessions (devices).

Table 12: Mean (S.D.) of several metrics in the first/second sessions for user Study II.

	Mobile (1st session)	Desktop (2nd session)
#query	4.73 (2.47)	5.55 (3.24)*
average query length	3.83 (1.01)	3.91 (1.07)
#clicks	7.72(3.89)	10.01(3.08)**
average page dwell time	39.25 (22.63)	23.74 (20.05)**

6.3.2.2 Clicks for Recall and Exploration Following the on-line experiment design rationale as mentioned in Section 3.5, participants are asked to fill out an in-task questionnaire related to the purpose of click (recall, exploration or both), the utility and the novelty of the clicked document through a pop-up window. According to the results provided in Table 13, participants in the experimental system clicked significantly more documents from click history for recalling information, whereas there are no difference regarding to the number of clicks that are for exploration. This indicates that the participants indeed employ the

provided click history function during their search processes. Note that the numbers in bold denote the comparisons between the first and second session (*: $p < 0.05$; **: $p < 0.01$). Here, I adopt Generalized Estimation Equation (GEE) for system comparisons.

Table 13: Mean (and S.D.) of participants’ responses.

	Baseline	Experimental
clicks for recall (all)	0.92 (1.15)	2.12 (2.17)**
clicks for recall (from click history)	-	1.23 (1.88)
clicks for exploration	6.81 (3.75)	7.02 (3.18)
clicks for both	1.56 (2.01)	1.37 (1.76)

Table 14: Pearson correlations among four relevance and novelty measures.

	user relevance	topic relevance	context relevance	context novelty
user relevance	-	0.814	0.437	0.155
topic relevance	0.814	-	0.387	0.093
context relevance	0.437	0.387	-	0.383
context novelty	0.155	0.093	0.383	-

6.3.2.3 Participant-reported Document Utility Measures Besides collecting the purposes of document clicks, this experiment also collects four types of document utility measures in the continued search sessions (see Section 3.5.2.3 for more details), including user relevance, topical relevance, contextual relevance and contextual novelty. To understand the similarity and difference among these metrics, I compute their Pearson correlations and provide them on Table 14. Numbers in bold denote the corresponding correlation is statistically significant (two-tailed test) at p -value < 0.01 level.

According to the table, user relevance is highly correlated with topical relevance (aggregated from multiple users), meaning that users’ post-task relevance judgments focus on rating document utility at the topical level. More interestingly, I find that the contextual relevance does not fully align with post-task user relevance (0.437 correlation with user relevance and 0.387 correlation with topical relevance), indicating that participants do adjust their relevance criteria with the changing of search contexts. The contextual novelty is less correlated with document relevance (0.093 correlation with topical relevance and 0.155 correlation with user relevance) but still reflects the search context (0.383 correlation with contextual relevance). Overall, the differences among multiple document utility measures suggest the complexity of understanding a cross-device search process, and it is highly necessary to involve users and contexts when conducting search performance analysis.

To understand the relationship between document utility measures and document accessing purposes (recall, exploration or both), I compute the average document utility measures under each purpose group, and present their results on Table 15 (in which numbers in bold/italics indicates statistical significance comparing to the recall/exploration group at the p-value of 0.05 level. Statistical significances are examined through one-way ANOVA with post-hoc Bonferroni adjustments for multiple comparisons). As expected, when exploring new information, participants tend to relax their relevance criteria; thus, it shows that **both** \sim **recall** \succ **exploration** (with \sim indicating no significance while \succ denoting a statistical significance) for user relevance and topical relevance. Besides, the average novelty for *exploration* clicks is the highest among all three groups, and the *both* group have the second largest novelty. This also makes sense because both two groups contain the information exploration behaviors. The contextual relevance involves both relevance and novelty as the judging criteria. Therefore, the documents with ‘both’ purposes might be of high utility while either ‘recall’ or ‘exploration’ might have no differences. This can be illustrated as **both** \succ **recall** \sim **exploration**.

Table 15: Mean (S.D.) of several measures under different document access purposes.

	recall	exploration	both
user relevance	3.52 (1.39)	3.23 (1.40)	<i>3.84 (1.25)</i>
topic relevance	3.56 (0.75)	3.34 (0.73)	<i>3.66 (0.66)</i>
context novelty	2.15 (1.36)	3.48 (1.42)	2.90 (1.23)
context relevance	3.20 (1.44)	2.93 (1.40)	3.62 (1.14)

6.3.3 User Perception on the Baseline and Experimental Systems

The analysis of on-line experiments focus on the comparisons between the baseline system (named as **B**, no support) and the experimental system (named as **E**, with support). System performances are analyzed based on both user perceptions and user behaviors. User perceptions are collected through participants' responses in post-task questionnaire and user behaviors are extracted from recorded behavior logs. This section focuses on user perceptions and the next section analyzes system performance based on user behaviors.

Table 16: Mean (S.D.) of user perceptions (range: 1 - 7) on several measures.

Statement	Baseline	Experimental
Efficiency: I can quickly find the information I need	5.70 (1.12)	5.82 (1.01)
Effectiveness: The system provides me useful documents	5.91 (1.00)	6.01 (0.96)
Satisfaction: Overall, I am satisfy with my performance	5.58 (1.07)	5.63 (0.05)

I first provide an overall analysis about how participants perceive the usefulness of the experimental system. Each time after a participant completed a search task in the continued session, she is asked to answer a set of questions with responses from a 7-point Likert scale

(see Appendix B.4.2). Participants’ post-task responses on each question are summarized as the following Table 16. Although the experimental system receives relatively higher values on efficiency, effectiveness and satisfaction measures, no statistical significances are detected between the baseline and experimental systems. This might either because of the sample size is too small, or the system change is not substantial enough. Therefore, I provide more analysis of search behaviors in later sections.

Table 17: Mean (S.D.) of user perceptions on information re-finding support.

	Baseline	Experimental
Usefulness of the click history support ¹	-	5.12 (0.20)
Usefulness of the visitation reminder support ²	-	4.77 (0.25)
Usefulness of the Information from the first session	5.35 (0.19)	5.06 (0.18)
Tasks (in the experimental system)	PD	PE
Usefulness of the click history support ¹	<i>5.30 (1.45)*</i>	4.96 (1.19)
Usefulness of the visitation reminder support ²	4.30(1.78)	5.17 (1.40)

Note¹: ① in Figure 28. Note²: ② in Figure 28.

The experimental system provides three important functions for supporting information re-finding and information finding — click-through history, search result visitation reminder and SERP result re-ranking. Here, I will provide user perceptions on the provided click-through history and result visitation reminder since these two functions are directly visible to end users. In addition to ask user perceptions on these two function, to be consistent with User Study I, I also asked user feeling about the usefulness of information from the first search session. The results are provided in Table 17 (A number in italics indicates a significant result between the first and second questions using Generalized Estimating Equation (GEE)), which shows several important findings.

First, both click history and result visitation reminder are rated to be useful, with each

of them has an averaged score over 4.0, the neutral value in a 7-point scale. Second, although there is no overall difference between the usefulness of click history and the usefulness of result visitation reminder, there is a strong task effect. The click history is more useful for product search ($Z=-2.253$; $p=0.024$) while no difference for people search ($Z=-0.674$; $p=0.500$). One potential reason, based on the analysis of post-task interviews, is that participants often encounter hub pages (a webpage covering information of multiple topics) in product search (e.g., a webpage covering different types of tablets for PD1, or camera lens for PD2) and these pages are important sources for task resumption. Therefore, many clicks from the click history are looking for such webpages in PD tasks. Third, the usefulness of first-session information tends to be more important in the baseline system, and the trend stays the same across different tasks though no statistical differences are detected. This might be due to that the experimental system have already provided search support for task restoration so that participants do not need any further information from the first search session.

6.3.4 User Behaviors on Baseline and Experimental Systems

User perception offers a subjective evaluation of search performance from a user's perspective. To provide a relatively objective assessment, I further analyze user behaviors recorded in my system. Given the information re-finding support and the information finding support are the two most important components of the experimental system. This section provides a detailed analysis about how both search support functions affect user behaviors in continued sessions. Note that the following analysis adopts Generalized Estimating Equations (GEE) in SPSS to test for significant differences in the compared measures. The reported p-value is based on the Wald Chi-Square test, and we used the Bonferroni adjustment for multiple tests. The reason for using GEE is that the model does not require the normality of outcome variables and it also helps to explore the interaction between tasks and used systems.

6.3.4.1 Information Re-finding Behaviors I hypothesize that an effective information re-finding support should minimize user effort for task restoration in the continued sessions so that users can quickly resume their tasks and jump into the stage of information

finding. Under this assumption, I expect more re-finding (recall) clicks in the beginning stage for task restoration, fewer re-finding (recall) clicks in the middle or late stages and fewer re-finding queries in the experimental system. Here, a re-finding document refers to the document that has been clicked in the first session. A recall click is determined through participants' responses of the in-task questionnaire in Figure 9. A document with the selected option *recall* or *both* are both treated as recall clicks.

To validate the above assumptions, I compute the following measures in Table 18 and compare their differences between baseline and experimental systems. The beginning stage is defined as the first **three** minutes in the continued session⁸, and the rest **seven** minutes belong to the late stage. Experimental results in Table 18 (Statistical significances are examined with Generalized Estimation Equations (GEE) in SPSS. Numbers with †, * and ** indicate the p-value is at 0.1, 0.05 and 0.01 levels, respectively.) confirm my hypotheses — comparing to the baseline system, participants in the experiment system issued a smaller number and percentage of repeated queries (measures 1 and 2), they visited more documents for information re-finding in the beginning stage (measures 3 to 6), while fewer in the late stage (measures 7 to 10). These results indicate that the provided search supporting functions do work positively to push the information re-finding behaviors to the beginning of a search session so that people can focus on exploring novel information in later stages of a session.

The above conclusion can be simply illustrated with Figure 31, in which I plotted the temporal distributions of participants' repeated click percentages in (a), and participants' recall click percentages in (b). Again, the recall clicks are determined by participants' responses of in-task questionnaires on Figure 9 (X-axis represents time, and Y-axis denotes the percentage. Information about recall click is obtained through questionnaire in Figure 9). The two subplots in Figure 31 show a consistent message — participants in the experimental system conduct a large amount of information re-finding behaviors in the beginning stage and then these behaviors decrease rapidly for the rest of stages. This further indicates that the provided information re-finding support can push the task resumption behaviors to the beginning phase so that users can focus on exploring novel information in later stages.

⁸I also tried two minutes or four minutes and the results are very similar

Table 18: Mean (and S.D.) of information re-finding measures in the continued sessions.

	Baseline	Experimental
1: #repeated queries	0.96 (1.03)	0.46 (0.73)**
2: %repeated queries	0.21 (0.26)	0.13 (0.24)*
3: #repeated clicks (the beginning stage)	0.50 (1.67)	1.12 (1.76)*
4: %repeated clicks (the beginning stage)	0.18 (0.26)	0.29 (0.37) †
5: #recall ¹ clicks (the beginning stage)	0.91 (1.01)	1.88 (1.80)**
6: %recall clicks (in the beginning stage)	0.31 (0.36)	0.48 (0.36)**
7: #repeated clicks (the late stage)	0.79 (1.00)	0.42 (0.70)**
8: %repeated clicks (the late stage)	0.12 (0.16)	0.06 (0.10)*
9: #recall clicks (the late stage)	1.58 (1.72)	1.60 (1.54)
10: %recall clicks (the late stage)	0.26 (0.30)	0.26 (0.25)

Note¹: a document is a recall click once a participant clicks the document and selects either “recall” or “both” in Figure 9.

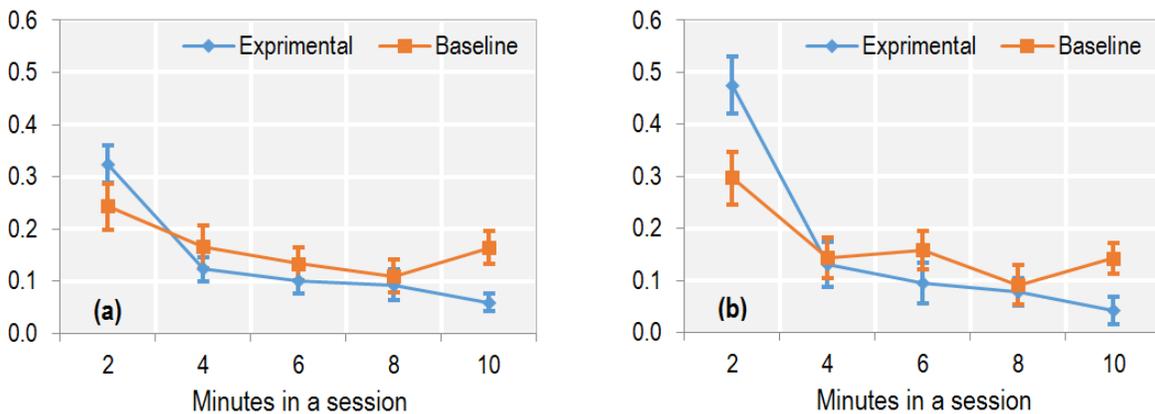


Figure 31: Temporal distribution of repeated/recall click percentage over time.

6.3.4.2 Information Finding Behaviors Besides the information re-finding support, the experimental system also provides two important functions to assist information finding — the search result visitation reminder that helps users easily skip their previously-clicked documents, and the re-ranked SERP results to push useful documents to top positions for quick access. With these support functions, I expect participants to spend less effort on searching and exploring novel documents in the experimental system comparing to the baseline system. To validate such hypothesis, I compute a set of nine metrics under four groups, which are described in details as following. Note that the following metrics only focus on the clicked documents with participant-rated purpose as “*exploration*” since the main goal of this section targets on the support for information finding.

- **Query-related metrics** measure participants’ search behaviors through user-issued queries. Since querying is the only way to pull results from the document collection, a system that produces similar search performance but requires more search queries might be less effective. Therefore, *#queries* is my first search metric. In addition, *query length* and *query readability* as mentioned in Section 4.3.1 are adopted to measure the complexity of a search query. I expect that complex queries can better support the exploration of a novel information.
- **Ranking-related metrics** attempt to assess the utility of SERP re-ranking algorithms by the ranking positions of the useful results. Here, the usefulness of a document can be simply measured by whether it is clicked or not. For each SERP, I consider two ranking-related metrics including the *average ranking positions* and the *nDCG* of the clicked documents.
- **Click-related metrics** measure the level of search exploration through clicked documents. Systems with better search support would result in clicking more documents (*#clicks*) and exploring more complex documents (*click-through complexity* as mentioned in Section 4.3.1).
- During the search course in continued sessions, each participant needs to report document relevance and novelty before she leaves the document. The **participant-reported metrics** aim to provide an overall assessment for document utility. Here, I consider both

document usefulness (*contextual relevance*) and document novelty (*contextual novelty*) measures.

Table 19 provides the assessment for the above-mentioned nine metrics under both baseline and experimental systems. The statistical significances in this table are tested by Generalized Estimating Equation (GEE) with Bonferroni adjustments for multiple tests. Numbers with †, * and ** indicate the p-value is at 0.1, 0.05 and 0.01 levels. For *query-related metrics*, participants in the experimental system issued fewer (metric 1) but longer (metric 2) queries. This demonstrates that the experimental system can effectively save search efforts by reducing repeated queries — after removing the repeated queries, there is no significant difference between the baseline and experimental systems. Meanwhile, the experimental system also helps participants come out more complex queries (metric 2, with a marginal significance).

Ranking-based metrics provide strong positive messages for the information finding support. Participants in the experimental system do not need to reach as deep as the baseline system (metric 4), to seek for useful documents (metric 5). At the same time, such search support function does not sacrifice user exploration of relevant (metric 8) and novel documents (metric 9) from users' perspective, i.e., *participant-reported metrics*.

Regarding the *click-based metrics* (metric 6 and metric 7), though the absolute values for the experimental system tend to be larger, I do not find any statistical difference. One of the reasons might be that participants in the experimental system usually spend longer time on task restoration (according to Table 13 and Table 18, there are more repeated and recall clicks in the experimental system) so that the amount of time spent on exploring new information would be less. Even in this case, the experimental system does not show a significant drop on #clicks for exploration, which is also a positive signal to demonstrate the usefulness of the experimental system.

6.3.4.3 Summary and Discussion The comparison of user behaviors (both information finding and re-finding behaviors) between the experimental system and the baseline system have shown the effectiveness of my search support function — it helps push the re-finding behaviors to the beginning of the continued session so that users can focus on

Table 19: Mean (S.D.) of information exploration metrics for system effectiveness.

	Baseline	Experimental
1: #queries	6.08 (3.30)	5.06 (2.95)**
#non-repeated queries	5.12 (3.34)	4.60 (2.99)
2: average query length	3.78 (0.95)	4.02 (1.12)†
3: query readability	9.53 (2.88)	9.47 (3.66)
4: average ranking position (of clicked documents)	4.55 (2.27)	3.89 (1.72)*
5: nDCG (of clicked documents)	0.50 (0.19)	0.57 (0.17)**
6: #clicks	6.81 (3.75)	7.02 (3.18)
7: click-through complexity	4.91 (1.22)	5.02 (0.88)
8: contextual relevance	2.95 (0.93)	2.89 (0.88)
9: contextual novelty	3.35 (1.15)	3.33 (1.05)

exploring novel information in the rest of time, and it further facilitates the issue of more complex queries and reduce users’ effort on exploring deeper information. However, it is also to my surprise that the overall contextual novelty is not improved since the initial goal for SERP re-ranking is to assist the exploring of novel information. I believe that there are two major reasons for this. First, my algorithm only re-ranks the ten search results within one SERP and does not include other documents, which might have very limited power to bring totally novel information. Second, search tasks considered in my user study are all exploratory and users in the first search session (spent 8 minutes) might have already read documents that covered multiple sub-topics such as the Wikipedia page. Therefore, it becomes much harder for users to find novel information in the continued session. I would like to explore further about this topic in the future.

6.4 ANSWERS TO RQ 3

This chapter developed two important search support functions in the continued sessions, one for supporting information re-finding and the other for assisting information finding. Both of the behavior-based and content-based search process models (discussed in Chapter 5) were employed and demonstrated to be effective for supporting cross-device web search. Overall, this chapter answered research question RQ3. Here, I list the answers to the three sub-questions.

- **Answer to RQ 3.1.** Based on the off-line experiments (Table 10), the behavior-based search process model was proven to be useful in predicting repeated clicks, and it is particularly useful in the case of cross-device search (compared to the same-device cross-session search). However, this model cannot be used for SERP re-ranking due to the lacking of behavioral information for non-clicked documents.
- **Answer to RQ 3.2.** Based on the off-line experiments (Table 10 and Table 11), the content-based search process model was proven to be useful in both predicting repeated clicks and assisting SERP results re-ranking. Content-based search process model worked particularly well on D-D (compared to M-D), which is consistent with my previous findings on Section 5.2, and was further explained in Section 5.3.
- **Answers to RQ 3.3.** An on-line evaluation experiment (with recruited participants) was designed and carried out to assess the effectiveness of cross-device search support functions. The experimental search system with properly-designed search support functions were demonstrated to be effective in the following two aspects. First, the information re-finding support facilitates users' information re-access behaviors so that they do not need extra efforts on issuing repeated queries and revisiting prior documents. This also makes users to be more focused on information finding in later search stages and explore in more depth. Second, the information finding support through SERP re-ranking also reduces user efforts when seeking for deeper information.

7.0 DISCUSSION, CONCLUSION AND FUTURE WORK

The above three chapters (Chapter 4, Chapter 5 and Chapter 6) have clearly answered each of the three research questions raised in the Introduction (Chapter 1). In this chapter, I will summarize all of the experimental results and findings discovered in the above sections, and further discuss the implications of those results. Specifically, I first provide the conclusions and contributions of this dissertation through answering each of the three research questions. Then, Section 7.2 presents a comprehensive discussion about the results. After that, I talk about the limitations of this dissertation in Section 7.3. Finally, Section 7.4 envisions potential future research topics.

7.1 CONCLUSION & CONTRIBUTIONS

7.1.1 Conclusions

Observing the increasing popularity of cross-device web search but a lack of sufficient understanding on this complex search condition, my whole dissertation targets to provide a comprehensive understanding of this new search process. To obtain a complete knowledge about user behaviors, I further include a baseline condition that goes across multiple sessions but using the same device.

Based on the datasets collected from two user studies with 48 participants, I discovered several important behavioral patterns for cross-device web searches: 1) comparing to the initial search session, users tend to conduct more sense-making behaviors in the continued session; 2) information re-finding and information finding are the two most important be-

haviors in the continued search sessions; 3) information re-finding behaviors are likely to happen in the beginning of a continued search session; and 4) user knowledge grows during the whole search course. Particularly, user knowledge grows rapidly in the beginning of a continued session and then gets saturated gradually. All of these conclusions hold for both cross-session same-device search and cross-device search. However, I find that search device does affect user behaviors in a considerable way. Due to the device difference, user behaviors in the initial sessions look very different, which mainly due to the convenience of device usage and the difference of information displays. This further affects user behaviors in continued sessions — users with the initialized session on mobile device often need to spend more efforts on the continued devices because of insufficient exploration (or mismatched returned search results) in the first session.

Known the importance of information re-finding and information finding behaviors in continued sessions, I further explored the automatic ways of supporting those behaviors. To quantitatively capture users' behavior patterns, I adopted a hidden variable model in which I assumed that user behaviors are driven by hidden factors; and developed two kinds of search process models. One model treated search tactic as the hidden factor, while the other one viewed user knowledge as the hidden factor. Both of them were demonstrated to be valid for modeling cross-device web search processes.

Finally, these two search process models were applied to support cross-device web search. With extensive evaluation results from both off-line (through data simulation) and on-line experiments (with real-world users, see §6.3), I discovered that both search process models can be properly utilized for supporting cross-device web searches.

7.1.2 Contributions

In general, I believe that there are three important contributions of my study.

First, to the best of my knowledge, this is the first comprehensive study about the understanding of cross-device search processes for exploratory tasks. My dissertation not only examined and compared how people search within and across multiple devices, but also studied automated ways of modeling and supporting important behavioral patterns in a

cross-device web search. It identified two critical behavioral patterns in the continued session of a cross-device web search — information re-finding and information finding. Since each of them has a different (or even contradictory) search purpose, a search system should properly differentiate these two behaviors when providing support. More importantly, my dissertation also proposed quantitative search process models to capture these behavioral patterns and further applied them in real-world search support scenarios. All of these findings and models could benefit future researchers who are interested in studying similar topics, or industry practitioners who planned to design and develop cross-device search systems.

Second, my dissertation devised two quantitative search process models, both target to bridge the hidden factor with observed search behaviors. In the behavior-based search process model, a hidden variable can be thought as the search tactic; while it is treated as a user’s knowledge state in content-based search process model. Both two assumptions were demonstrated to be reasonable through my experimental analysis. These models might be interesting for researchers who are studying information seeking process (ISP) and interactive information retrieval. The search tactic and user knowledge have been long studied in information retrieval community as main drivers for user behaviors [10, 39, 82, 99]. However, they were mainly studied as theoretical frameworks in prior studies, mainly due to the difficulty of qualifying search tactic and user knowledge. The quantitative model proposed in my dissertation might open a new direction for existing studies.

Third, the proposed search process models might also beneficial other fields. Behavior-based search process model attempted to model sequential user behaviors, which can be extended to other domains when analyzing similar sequential behavioral data. Indeed, recent studies from psychology [125] and human-robot interaction [15, 27] have already applied the HMM for discovering important human interaction patterns. Similarly, the knowledge-based search process model is to model human reading and knowledge learning, which shares similar characteristics as many other tasks such as textbook-based knowledge learning [63]. I expect my model to be useful in that task.

7.2 DISCUSSION OF RESULTS

7.2.1 Re-finding and Finding Behaviors in Cross-device Web Search

Consistent with prior studies [133, 134, 137], I also discovered a plenty of *information re-finding* behaviors in cross-device (and cross-session) web searches. Table 3 (data from User Study I) showed that around 15% of search queries, 20% of document clicks and more than 45% of the issued query terms in the continued sessions are repeated from previous search sessions. Table 18 (data from User Study II) reported similar behavioral patterns for repeated clicks and repeated queries — 20% of search queries are repeated from the first session, 15% (or 30%) of document clicks are repeated clicks (or user-reported re-finding clicks).

Beside re-finding, Chapter 4 and Chapter 5 also discovered that the exploration of relevant documents (i.e., the *information finding* behavior) take a large proportion of user behaviors, which is even more than re-finding in the continued session. Note that the information finding behavior does not exactly map with the exploration state of HMM outputs in Figure 18. Here, finding can be viewed as the counterpart of re-finding. Therefore, it covers the hidden states of exploration, exploitation and querying (if the purpose of a query is not to find repeated documents). Overall, the results from this dissertation clearly demonstrate the importance of both finding and re-finding behaviors in cross-device web searches, which are also my major focuses when developing the search support functions.

Experimental results from the HMM outputs (Figure 18, data from User Study I) and SERP re-ranking evaluation results (Table 19, data from User Study II) both show that information re-finding and information finding have clear differences. In information finding, users tend to seek for documents with more novelty and willing to sacrifice relevance. However, for information re-finding, users are more likely to examine relevant documents other than novel documents.

The difference between finding and re-finding calls for the development of a search support algorithm that can automatically identify user purposes and customize search results based on them. The HMM-based approach in Section 5.1 cannot be directly employed since it relies on the post-task behavior analysis. Fortunately, in this dissertation, I identify that

timing is a strong behavioral that can differentiate re-finding from exploration. User study results from Figure 11 (User Study I) and Figure 31 (User Study II) both showed that the majority of information re-finding behaviors occur in the beginning of the continued sessions. Providing necessary support for re-finding, such as displaying search history, in the beginning stage of the second session would probably satisfy most of user needs. This is actually the design rationale for my cross-device search support system, and the experimental results show that such support is sufficient to generate better search performance.

However, I do think that the investigation on automatic detection of search intentions remain to be an important topic. Particularly when users conduct the within-session information re-finding behaviors [137], which was not effectively supported in my system.

7.2.2 Knowledge Growth during Searching

When modeling users' search content change during search, I assumed that user knowledge plays the key role, and thus proposed a knowledge-based search process (KSP) model. Results from both Figure 12 (implicit measure of knowledge) and Figure 14 (explicit measure of knowledge) confirmed my hypothesis, and further discovered a unified pattern for knowledge growth as shown in Figure 32: user knowledge grows rapidly at the beginning and gradually gets saturated; and due to interruption (between two search sessions/devices), it also causes user knowledge drop. The knowledge growth pattern looks like to be the combination of the learning curve [152] and forgetting curve [93].

Quantifying the knowledge growth pattern is a challenging topic since it requires a proper representation of user knowledge and the simulation of human learning process. A principled approach, as illustrated in Corbett and Anderson [32], is to represent user knowledge as multiple knowledge units and accumulate user understanding of each unit across time. The KSP model simulates this principle except representing user knowledge as knowledge unit, which is due to the following reasons. First, different from the factual-based search topics, exploratory search tasks are often open-ended and are hardly to construct a fixed representation of knowledge units for all users. Second, the KSP model is built to provide

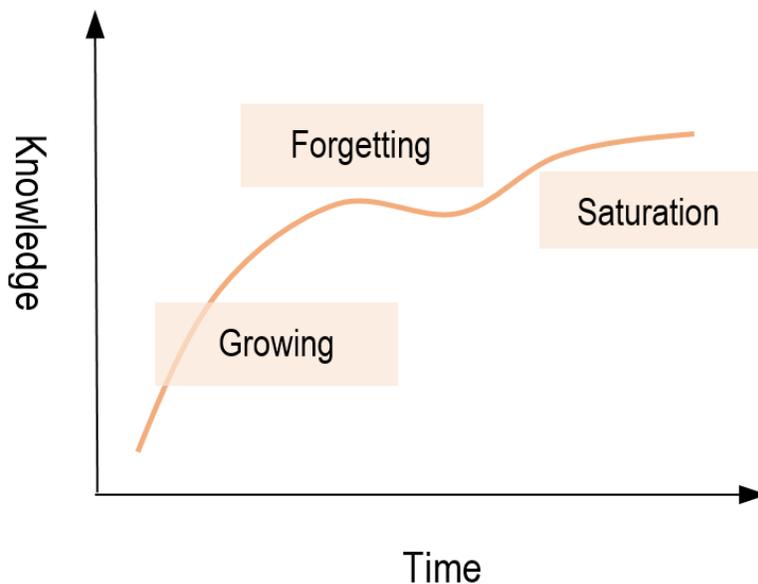


Figure 32: Knowledge growth and saturation pattern identified in our user studies.

two important search support functions for information re-finding and finding, both targeted to predict future user interactions at the document level instead of knowledge unit level. Therefore, directly modeling the document-level understanding at the very beginning would be sufficient. Note that the KSP model only simulates the learning curve whereas there is no further consideration for modeling forgetting. This will leave of the future study.

Some of the existing studies [38, 90] did find a similar knowledge growth pattern as my study; however, they did not further consider applying such findings for real-world search support. In this dissertation, I explored ways of predicting future user clicks with obtained user knowledge. According to the results in Section 5.2.2, user knowledge is a compound factor when predicting user clicks. On the one hand, a larger value of user knowledge may indicate that the user has already known most of them; thus, there is no need to click the corresponding content (i.e., knowledge saturation). On the other hand, a larger value of user knowledge also represents content relevance; thus, comparing to non-relevant content, a relevant content will get a higher probability of clicking. Therefore, I further leverage this compound factor in Section 6.1.4 when re-ranking SERP results.

7.2.3 Modeling Search Process with Hidden Variables

Existing information retrieval research [100, 145] still lacks a sufficient understanding of human search process for long-time complex search tasks, for which users' relevance judgment criteria and their search needs change dynamically, whereas its causes remain unknown. Recent studies have found that they are closely related to multiple factors such as user knowledge [10, 38, 158], adopted search tactics [8, 54, 155], search skills and experience [60], and so on. As an initial step to study the quantifying of complex search processes, this dissertation explored ways of modeling the above factors with hidden variables and assumed that user behaviors are all driven by the hidden variables.

Particularly, this dissertation took into account two types of hidden factors — knowledge state and search tactics, and therefore developed two hidden factor models — the HMM-based search process model (Section 5.1) and the knowledge-based search process model (Section 5.2). The HMM-based approach focuses on uncovering hidden search tactics from system recorded user behaviors and the knowledge-based approach attempts to trace and predict user behaviors based on their knowledge changes. Through an extensive analysis with two search support tasks and several on-line and off-line evaluation experiments, this dissertation demonstrated the validity and effectiveness of both search process models.

There are several advantages of modeling search process with the hidden variable model. First, different from the prior studies that only considered simple behavior counts or behavior couples [25, 150, 158], the hidden factor model integrated observed user behaviors with hidden factors into one unified framework, and further assumed a generative process from hidden factor to observed behaviors. This mechanism makes the proposed models more interpretable. Second, the hidden variable model is a generic framework, which can be easily extended or integrated with other hidden factors. For example, my dissertation did not consider the integration of both search behavior and search content in one unified framework, but I do think it is a future direction to explore. Under the hidden variable model framework, it might be easy to merge two hidden factors with one unified factor. Third, hidden variable model is an active area in machine learning which provided us a list of examples on modeling, inference and applying such model.

7.2.4 Cross-Device Search VS. Same-Device Cross-Session Search

When studying user behaviors for cross-device web search, a same-device cross-session search condition was included in User Study I for comparisons. Through a comprehensive experimental analysis from all of the above three chapters (Chapter 4, Chapter 5 and Chapter 6), I found that although many user behavior patterns look quite similar, there are also several significant differences between these two search conditions, which can be summarized as the following three aspects.

First, comparing to D-D, users on M-D issued fewer queries, clicked and saved fewer documents but stayed longer time in each clicked document in the first session (see Table 2). This is mainly due to the inconvenience of typing and reading on mobile devices. In addition, because of their insufficient exploration on the first session, users on M-D need to conduct mixed behaviors including both information exploration and sense-making in the continued sessions. However, users on D-D can primarily focus on the sense-making activities since most of the relevant documents might have already examined in the prior sessions.

Second, in terms of the information re-finding behavior in the continued sessions, users on M-D conducted more such behaviors (issuing duplicate queries or revisit past clicks) than that on D-D. One potential reason, as I inferred in both Section 4.2.2.1 and Section 4.3.2.3, is that the information display styles between mobile and desktop are different. For example, the same Wikipedia entry is presented differently depending on the type of accessing device. Users might need to re-read or re-confirm the same content once using another device.

Third, when modeling users' search processes with knowledge-based approach (i.e., KSP model), I discovered a consistent difference between M-D and D-D — results from both Section 5.2 and Section 6.2 illustrated that the knowledge-based modeling on D-D consistently outperforms M-D. This, again, might be due to 1) users in the first session of D-D can explore more thoroughly so that the knowledge modeling might be more robust; and 2) the knowledge saturation assumption might work better for the cases when users are conducting sense-making activities instead of still actively looking for relevant documents. In the future, I would like to conduct a further analysis about the reasons.

Those differences reflect that search device plays an important role when determining user

behaviors. As a result, it might be hard to directly reuse the experiment findings from one search condition to the other. However, as one can see from Chapter 5 (modeling), the overall procedure to apply the two proposed search process models (HMM-based behavior model and KSP-based behavior model) remains the same, although the outputs and performances are significantly different. The generalizability of search process models is one of the main contributions I mentioned in Section 7.1. The two search conditions do show very similar behavioral patterns. For example, I discovered that the major behavioral patterns such as information re-finding and information finding in the continued sessions, and the search support mechanism for both re-finding and finding are the same. In the future, I would like to explore more on the connections between the two search conditions.

7.3 LIMITATION

Despite the above contributions and insights, I do acknowledge that my dissertation study also has several limitations, which are provided with details in the following subsections.

7.3.1 Representativeness of User Study

As many user study based research, the representativeness of users, tasks, conditions (e.g., M-D and D-D) are always the major limitation and are hard to be fixed. In this dissertation, I believe that the following factors can affect the result representativeness.

First, the conclusions and findings of my dissertation were purely based on the data collections obtained from lab-controlled user studies, which are usually limited in scale and hard to be replicated. Second, the search tasks adopted in my user studies (both User Study I & II) were predefined and assigned to the participants, instead of being initialized by participants. Although my tasks were designed based on Mechanical Turkey survey responses, and intended to avoid task effect through designing diverse search topics, they might still be unrepresentative for the participants I recruited. This may affect the ways of how the participants perform their searches. Third, when studying the cross-device and

cross-session web searches, I only took into account two types of search conditions — Mobile-to-Desktop (M-D) and Desktop-to-Desktop (D-D). I do believe that users would perform differently on Desktop-to-Mobile (D-M) and Mobile-to-Mobile (M-M), and I certainly know that my results cannot simply be generalized to these two new search conditions. I would like to explore more search conditions in the future. Fourth, User Study II included an in-task questionnaire asking about participants' to answer in the middle of their searches. This design will surely affect real-world user behaviors. Finally, the recruited participants are all college students, which are apparently not representative for other groups of searchers. The influence of the above factors are unknown and hard to measure; thus, I suggest future researchers to be more careful when drawing conclusions from this work.

7.3.2 Over-simplified Model Assumptions

When modeling users' search processes, I made several strong assumptions particularly for the KSP-based search process model. To be specific, there are three important assumptions/simplifications I made for the KSP model — assuming a zero initial user knowledge, setting a constant learning rate for all individuals and representing user knowledge with Bag-Of-Words (BOW) instead of semantic knowledge units. Those assumptions do make the computation of user knowledge much more effective but are obviously inaccurate, which might even lead to unreliable results.

According to Figure 14, even for the complex and exploratory search tasks I specifically design for this study, the initial user knowledge is not zero. In real-world scenarios, users might deal with more common and much simpler search tasks; thus, their initial knowledge should be even higher. In terms of learning rate, I did observe that different participants behave differently during search. Some of them learn fast and come out lots of effective queries, whereas some others search and read slowly. Apparently, setting a fixed learning rate cannot be applied in this case. As for the BOW-based knowledge representation, although I have provided several good reasons in Section 5.2.1 to justify my approach, there is still a mismatch between words and knowledge units. One important problem for BOW-based representation is that it does not capture the underlying semantics behind words if they are

not in the same form. In this future, I would like to explore more advanced approaches to improve those simplifications.

7.3.3 Increased User Load with Complex Interface

Comparing to the organic search system, my search support system (see Figure 28) provided a much more complex interface. Although there are three major modules in the interface, only the result visitation reminder and re-ranked click-through history directly changed the interface displays. Based on the observation of user behaviors, the new interface does benefit user search but also brings some confusions particularly for the re-ranked click-through history. By summarizing user feedbacks for the new interface, I find that there are two major problems that might potentially increase users' cognitive load. First, the re-ranking criteria is unclear to users. Several participants asked about how those documents were ranked, and their connections with visitation recency. Second, several participants also complained that they could not access the re-ranked history if not returning to the index page, and they suggest providing such functionality in SERPs so that they can leverage the current search context with past search histories. In summary, complex interface also increases users' task load. Even including a certain amount of training time in my dissertation, users still have an increased task load; therefore, I expect more cognitive load from users who are not familiar with the interface. In the future, I would like to explore simpler interface to leverage task load and interface complexity.

7.4 FUTURE WORK

In the future, I would like to explore the following three research directions.

7.4.1 Generalizing to Real-world Search Tasks

Although the conclusions and findings in this dissertation were mainly based on the data collected from lab-controlled user studies, I do believe that several of the discovered behavioral

patterns (e.g., information re-finding and information finding) can be easily generalized to real-world search tasks. Once obtaining large-scale cross-device search logs, I would like to replicate several of my experiments designed in this dissertation. Despite some of prior studies [134, 137, 142] have examined the behavioral patterns for information re-finding, they did not specifically focused on cross-session and even cross-device search scenarios, and they also did not compare the differences among different search scenarios. More importantly, their analysis and their experiments did not attempt to model the whole cross-session and cross-device search processes, and did not further consider applying their findings for supporting complex and exploratory web search tasks.

Besides, although the above-mentioned search process models were only analyzed for cross-device and cross-session web search processes. I expect that both models could be easily applicable for other search conditions, such as single-session long-time search, multiple-session web search (with more than two search sessions). I would also like to explore the generalizability of the proposed search model to other search conditions in the future.

7.4.2 Improving Search Process Models

My dissertation designed and developed two quantitative search process models, one focusing on bridging hidden search tactics and observed user behaviors, and the other emphasizing on quantifying users' hidden knowledge learning process through simulating the human learning process. Although they were both demonstrated to be useful in modeling and supporting cross-device and cross-session search processes, there are still much room to improve.

As mentioned in Section 7.3.2, the KSP model is built on top of several simplified assumptions such as zero initial knowledge, constant learning rate and Bag-of-Word (BOW) representation of user knowledge. To fill the KSP model with accurate initial knowledge, one potential approach is to infer users' initial knowledge from their search and interaction histories if such data is available; otherwise, in a user study, one might need to collect such information using pre-task questionnaire as I did for my user study II. Deriving learning rate for each individual is a challenging topic due to data insufficiency for one single user. One possible remedy is to cluster users into different cohorts and then assume that all users

within the same cohort share the same learning rate. In this way, all behaviors within the same cohort can be used to infer the learning rate. In terms of knowledge representation. I would like to extend the current Bag-Of-Words mechanism to semantic knowledge units, for which I plan to experiment several advanced approaches such as Latent Dirichlet Allocation (LDA) and word embeddings.

In addition, the HMM-based search process model and the KSP-based search process model were applied and analyzed independently in my dissertation. Although the two models try to capture different aspects of the same search process, it might be still reasonable to integrate them together within one framework since user behaviors (such as query or click) always couple with certain content (query content or click content). The simplest approach is to treat the two model outputs as separate features so that future machine learning models can directly include them in their feature set. A more strict way is to put both models in one unified umbrella such as the hidden variable model framework, in which one can define two hidden variable — one variable to represent search tactic and the other to denote user knowledge, they can be coupled together to determine both search behaviors and search content with different weights [107].

APPENDIX A

MECHANICAL TURK QUESTIONNAIRE

We are conducting a research study on cross-device information search. Particularly, we are interested in situations where you use your mobile device to perform a search first and then continue the search later on a computer. An example scenario can be: you see on a bus an advertisement about spending a holiday at a Greek island, so you perform a search using your mobile device (smartphone or tablet) first to look for the island and find it interesting. Later when you have an access to a computer, you continue your search on the island and on the planning of a trip to the island.

By mobile devices, we refer to a device that: 1) you would carry with you for most of the time; 2) it does not have a physical keyboard or a fully functional keyboard for you to type; and 3) the screen is usually smaller than 10 inches. Examples include your smartphones or tablets. By computers, we mean a desktop or laptop computer that has a real keyboard and a screen larger than 11 inches.

1. Have you ever searched for the same information need using a mobile device first and then transfer to use a computer to continue the search?

Yes

No

2. What mobile devices did you use in cross-device search? (select all that apply)

Iphone

Android phone

Other phone, please specify _____

Ipad (iOS-based tablet)

Android-based tablet

Other tablet, please specify _____

3. What computer did you use in cross-device search? (select all that apply)

Desktop

Laptop

Netbook

Other, please specify _____

4. How many searches did you perform using your mobile device in the past week?

None

1-5 times

6-10 times

11-20 times

20-50 times

More than 50 times

5. How many searches did you perform using your computer in the past week?

None

1-5 times

6-10 times

11-20 times

20-50 times

More than 50 times

6. What triggered you to perform a search using mobile device in the first place? (select all that apply)

I saw something and I need to know more about it

I talked with someone and I need to know more about something mentioned in the conversation

I thought about something and I need to know more about it

I need to do something and I need to know more about how to do it

I received a call/message/email and I need to know more about something in the call/message/email

I was browsing the Web and came to something that I wanted to know more about it

Other, please describe

7. Have you ever searched for one or more of the following topics using your mobile device first and then continue your search using your computer? (select all that apply)

Products/items for purchase

People and organizations

Images

Videos

News

Books

Movies

Music

Locations

Events

Other, please describe

8. How did you search for the information using your mobile device? (select all that apply)

Used a search engine

Visited a specific Website (other than the search engine page)

Used an app (other than a search app)

9. How did you preserve the information you found on your mobile device so that you can continue the search later using a computer? (please select all that apply)

Bookmark the information found

Save the information found

Memorize the information found

Take a photo of the information found

Email myself /other of the information found

Share the information to a social media (facebook, twitter, linkedIn and etc)

Other, please describe

10. How did you search for the information using your computer? (select all that apply)

Used a search engine

Visited a specific Website (other than the search engine page)

11. Can you describe at least one situation that you searched on your mobile device first and then performed a follow-up search using your computer?

What topic did you search for?

What triggered the information need?

What did you search using your mobile device?

Did you use a search engine? What was it?

What are the queries you used for search if you can recall or what are the sub-topics you searched for if you cannot recall the specific query?

Did you visit any Website? What were those Websites if you can recall?

How did you preserve the information you found on mobile device?

What did you search using your computer?

Did you use a search engine? What was it?

What are the queries you used for search if you can recall or what are the sub-topics

you searched for if you cannot recall the specific query?

Did you visit any Website? What were those Websites if you can recall?

12. How many searches can you recall using your smartphone? _____. Among these mobile searches you can recall, how many did you perform a follow-up search using your computer? _____

APPENDIX B

QUESTIONNAIRES FOR USER STUDIES

B.1 ENTRY QUESTIONNAIRE

Participant ID: _____

1. Your age: _____

2. Gender: _____

3. Your program of study: _____ Undergraduate _____ Graduate _____ Other: _____

4. What is your major course of study or profession? _____

5. How would you describe your search experience?

(Very Inexperienced) 1 2 3 4 5 6 7 (Very Experienced)

6. How many searches did you perform using your mobile device in the past week?

None

1-5 times

6-10 times

11-20 times

20-50 times

More than 50 times

7. How many searches did you perform using your computer in the past week?

None

1-5 times

6-10 times

11-20 times

20-50 times

More than 50 times

8. Among all the searches you can recall, what percentage were performed on mobile?

9. Among all the mobile searches you can recall, what percentage did you perform a follow-up search using your computer? _____

B.2 TRAINING TASK

Training task: You and your friend were talking about the recent government shutdown. You are interested in finding more information about it. What were the causes? What were the consequences? How it was ended?

B.3 POST-TASK QUESTIONNAIRES FOR USER STUDY I

B.3.1 Post-task Questionnaire after the 1st Session

1. How fast can you find the information you need?
(Extremely slow) 1 2 3 4 5 (Extremely fast)
2. How easy was it to find the information you need?
(Extremely easy) 1 2 3 4 5 (Extremely difficult)
3. How satisfied are you with the information you found?
(Extremely dissatisfied) 1 2 3 4 5 (Extremely satisfied)

B.3.2 Post-task Questionnaire after the 2nd Session

1. How fast can you find the information you need?
(Extremely not useful) 1 2 3 4 5 (Extremely useful)
2. How fast can you find the information you need?
(Extremely slow) 1 2 3 4 5 (Extremely fast)
3. How easy was it to find the information you need?
(Extremely easy) 1 2 3 4 5 (Extremely difficult)
4. How satisfied are you with the information you found?
(Extremely dissatisfied) 1 2 3 4 5 (Extremely satisfied)

B.4 PRE-TASK AND POST-TASK QUESTIONNAIRES FOR USER STUDY II

B.4.1 Pre-task Questionnaire before the 1st Session

How familiar are you with the topic of this task?

(Not At All) 1 2 3 4 5 6 7 (Extremely familiar)

B.4.2 Post-task Questionnaire after the 1st Session

At this point, how familiar are you with the topic of this task?

(Not At All) 1 2 3 4 5 6 7 (Extremely familiar)

I can quickly find the information I need.

(Extremely disagree) 1 2 3 4 5 6 7 (Extremely agree)

The system can provide me with useful documents.

(Extremely disagree) 1 2 3 4 5 6 7 (Extremely agree)

Overall, I am satisfied with my performance.

(Extremely disagree) 1 2 3 4 5 6 7 (Extremely agree)

B.4.3 Pre-task Questionnaire before the 2nd Session

At this point, how familiar are you with the topic of this task?

(Not At All) 1 2 3 4 5 6 7 (Extremely familiar)

B.4.4 Post-task Questionnaire after the 2nd Session

At this point, how familiar are you with the topic of this task?

(Not At All) 1 2 3 4 5 6 7 (Extremely familiar)

How useful is the first session of search on this topic?

(Extremely useful) 1 2 3 4 5 6 7 (Extremely not useful)

I can quickly find the information I need.

(Extremely disagree) 1 2 3 4 5 6 7 (Extremely agree)

The system can provide me with useful documents.

(Extremely disagree) 1 2 3 4 5 6 7 (Extremely agree)

Overall, I am satisfied with my performance.

(Extremely disagree) 1 2 3 4 5 6 7 (Extremely agree)

How do you rate the usefulness of click history (explain to the user)¹?

(Extremely not useful) 1 2 3 4 5 6 7 (Extremely useful)

How do you rate the usefulness of click history (explain to the user)²?

(Extremely disagree) 1 2 3 4 5 6 7 (Extremely agree)

¹This refers to the click history in the index search page of the experimental system. Therefore, this question is only asked in the experimental system

²The same as above.

BIBLIOGRAPHY

- [1] Eytan Adar, Jaime Teevan, and Susan T Dumais. Large scale analysis of web re-visitation patterns. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 1197–1206. ACM, 2008.
- [2] Eugene Agichtein, Ryen W White, Susan T Dumais, and Paul N Bennet. Search, interrupted: understanding and predicting search task continuation. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 315–324. ACM, 2012.
- [3] Hirotugu Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.
- [4] James Allan, Courtney Wade, and Alvaro Bolivar. Retrieval and novelty detection at the sentence level. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 314–321. ACM, 2003.
- [5] Bryce Allen. Topic knowledge and online catalog search formulation. *The Library Quarterly*, pages 188–213, 1991.
- [6] Erik M Altmann and J Gregory Trafton. Task interruption: Resumption lag and the role of cues. Technical report, DTIC Document, 2004.
- [7] Anne Aula, Natalie Jhaveri, and Mika Käksi. Information search and re-access strategies of experienced web users. In *Proceedings of the 14th international conference on World Wide Web*, pages 583–592. ACM, 2005.
- [8] Marcia J Bates. Information search tactics. *Journal of the American Society for information Science*, 30(4):205–214, 1979.
- [9] Nicholas J Belkin. Anomalous states of knowledge as a basis for information-retrieval. *Canadian Journal of Information Science-Revue Canadienne Des Sciences De L Information*, 5(MAY):133–143, 1980.
- [10] Nicholas J Belkin, Robert N Oddy, and Helen M Brooks. Ask for information retrieval: Part i. background and theory. *Journal of documentation*, 38(2):61–71, 1982.

- [11] Paul N Bennett, Ryen W White, Wei Chu, Susan T Dumais, Peter Bailey, Fedor Borisjuk, and Xiaoyuan Cui. Modeling the impact of short-and long-term behavior on search personalization. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 185–194. ACM, 2012.
- [12] Krishna Bharat. Searchpad: Explicit capture of search context to support web search. *Computer Networks*, 33(1):493–501, 2000.
- [13] Ralf Biedert, Andreas Dengel, Georg Buscher, and Arman Vartan. Reading and estimating gaze on smart phones. In *Proceedings of the symposium on eye tracking research and applications*, pages 385–388. ACM, 2012.
- [14] Pia Borlund. Experimental components for the evaluation of interactive information retrieval systems. *Journal of documentation*, 56(1):71–90, 2000.
- [15] Yves Boussemart and Mary L Cummings. Predictive models of human supervisory control behavioral patterns using hidden semi-markov models. *Engineering Applications of Artificial Intelligence*, 24(7):1252–1262, 2011.
- [16] Marc Bron, Jasmijn Van Gorp, Frank Nack, Lotte Belice Baltussen, and Maarten de Rijke. Aggregated search interface preferences in multi-session search tasks. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 123–132. ACM, 2013.
- [17] Bertram C Brookes. The foundations of information science part i. philosophical aspects. *Journal of information science*, 2(3-4):125–133, 1980.
- [18] Georg Buscher, Andreas Dengel, and Ludger van Elst. Query expansion using gaze-based feedback on the subdocument level. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 387–394. ACM, 2008.
- [19] Georg Buscher, Ludger van Elst, and Andreas Dengel. Segment-level display time as implicit feedback: a comparison to eye tracking. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 67–74. ACM, 2009.
- [20] Huanhuan Cao, Daxin Jiang, Jian Pei, Enhong Chen, and Hang Li. Towards context-aware search by learning a very large variable length hidden markov model from search logs. In *Proceedings of the 18th international conference on World wide web*, pages 191–200. ACM, 2009.
- [21] Robert G Capra III and Manuel A Pérez-Quiñones. Using web search engines to find and refind information. *Computer*, 38(10):36–42, 2005.
- [22] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual*

- international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336. ACM, 1998.
- [23] Tsung-Hsiang Chang and Yang Li. Deep shot: a framework for migrating tasks across devices using mobile phone cameras. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 2163–2172. ACM, 2011.
- [24] Olivier Chapelle and Ya Zhang. A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th international conference on World wide web*, pages 1–10. ACM, 2009.
- [25] Hui-Min Chen and Michael D Cooper. Stochastic modeling of usage patterns in a web-based information system. *Journal of the American Society for Information Science and Technology*, 53(7):536–548, 2002.
- [26] Yu Chi, Shuguang Han, Daqing He, and Rui Meng. Exploring knowledge learning in collaborative information seeking process.
- [27] Shih-Yi Chien, Michael Lewis, Siddharth Mehrotra, Shuguang Han, Nathan Brooks, Huadong Wang, and Katia Sycara. Task switching for supervisory control of multi-robot teams. *IEEE Transactions on Human-Machine Systems*, 2016.
- [28] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. Click models for web search. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 7(3):1–115, 2015.
- [29] Charles LA Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666. ACM, 2008.
- [30] Cyril Cleverdon. The cranfield tests on index language devices. In *Aslib proceedings*, volume 19, pages 173–194. MCB UP Ltd, 1967.
- [31] Kevyn Collins-Thompson and Jamie Callan. Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13):1448–1462, 2005.
- [32] Albert T Corbett and John R Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1994.
- [33] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 87–94. ACM, 2008.

- [34] Laura Dabbish and Robert E Kraut. Controlling interruptions: awareness displays and social motivation for coordination. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 182–191. ACM, 2004.
- [35] Peng Dai, Jeffrey M Rzeszutarski, Praveen Paritosh, and Ed H Chi. And now for something completely different: Improving crowdsourcing workflows with micro-diversions. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 628–638. ACM, 2015.
- [36] David Dearman and Jeffery S Pierce. It’s on my other computer!: computing with multiple devices. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*, pages 767–776. ACM, 2008.
- [37] Yuka Egusa, Hitomi Saito, Masao Takaku, Hitoshi Terai, Makiko Miwa, and Noriko Kando. Using a concept map to evaluate exploratory search. In *Proceedings of the third symposium on Information interaction in context*, pages 175–184, New York, 2010. ACM.
- [38] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. Lessons from the journey: a query log analysis of within-session learning. In *WSDM*, pages 223–232, 2014.
- [39] David Ellis. Modeling the information-seeking patterns of academic researchers: A grounded theory approach. *The Library Quarterly*, pages 469–486, 1993.
- [40] Brynn M Evans and Ed H Chi. Towards a model of understanding social search. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 485–494. ACM, 2008.
- [41] Facebook. Finding simplicity in a multi-device world. <https://www.facebook.com/business/news/Finding-simplicity-in-a-multi-device-world>, 2014. Accessed: 2016-11-14.
- [42] Henry Feild and James Allan. Task-aware query recommendation. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 83–92. ACM, 2013.
- [43] Norbert Fuhr. A probability ranking principle for interactive information retrieval. *Information Retrieval*, 11(3):251–265, 2008.
- [44] Duncan R Godden and Alan D Baddeley. Context-dependent memory in two natural environments: On land and underwater. *British Journal of psychology*, 66(3):325–331, 1975.
- [45] Saul Greenberg and Andy Cockburn. Getting back to back: Alternate behaviors for a web browser’s back button. In *Proceedings of the 5th Annual Human Factors and the Web Conference*. NIST, 1999.

- [46] Dongyi Guan, Sicong Zhang, and Hui Yang. Utilizing query change for session search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 453–462. ACM, 2013.
- [47] Qi Guo and Eugene Agichtein. Ready to buy or just browsing?: detecting web searcher goals from interaction data. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 130–137. ACM, 2010.
- [48] Qi Guo, Haojian Jin, Dmitry Lagun, Shuai Yuan, and Eugene Agichtein. Mining touch interaction data on mobile devices to predict web search result relevance. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 153–162. ACM, 2013.
- [49] Shuguang Han, Daqing He, Jiepu Jiang, and Zhen Yue. Supporting exploratory people search: a study of factor transparency and user control. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 449–458. ACM, 2013.
- [50] Shuguang Han, Daqing He, Zhen Yue, and Peter Brusilovsky. Supporting cross-device web search with social navigation-based mobile touch interactions. In *User Modeling, Adaptation and Personalization - 23rd International Conference, UMAP 2015, Dublin, Ireland, June 29 - July 3, 2015. Proceedings*, pages 143–155, 2015.
- [51] Shuguang Han, Daqing He, Zhen Yue, and Jiepu Jiang. Contextual support for collaborative information retrieval. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval, CHIIR '16*, pages 33–42, New York, NY, USA, 2016. ACM.
- [52] Shuguang Han, I-Han Hsiao, and Denis Parra. A study of mobile information exploration with multi-touch interactions. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 269–276. Springer, 2014.
- [53] Shuguang Han, Xing Yi, Zhen Yue, Zhigeng Geng, and Alyssa Glass. Framing mobile information needs: An investigation of hierarchical query sequence structure. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 2131–2136. ACM, 2016.
- [54] Shuguang Han, Zhen Yue, and Daqing He. Automatic detection of search tactic in individual information seeking: A hidden markov model approach. *arXiv preprint arXiv:1304.1924*, 2013.
- [55] Shuguang Han, Zhen Yue, and Daqing He. Understanding and supporting cross-device web search for exploratory tasks with mobile touch interactions. *ACM Transactions on Information Systems (TOIS)*, 33(4):16, 2015.

- [56] Ahmed Hassan, Xiaolin Shi, Nick Craswell, and Bill Ramsey. Beyond clicks: Query reformulation as a predictor of search satisfaction. In *CIKM*, pages 2019–2028, 2013.
- [57] Jiyin He, Pernilla Qvarfordt, Martin Halvey, and Gene Golovchinsky. Beyond actions: Exploring the discovery of tactics from user logs. *Information Processing & Management*, 52(6):1200–1226, 2016.
- [58] Marti Hearst. *Search user interfaces*. Cambridge University Press, 2009.
- [59] Sandra G Hirsh. Children’s relevance criteria and information seeking on electronic resources. *Journal of the Association for Information Science and Technology*, 50(14):1265, 1999.
- [60] Christoph Hölscher and Gerhard Strube. Web search behavior of internet experts and newbies. *Computer networks*, 33(1):337–346, 2000.
- [61] Jeff Huang, Ryen W White, Georg Buscher, and Kuansan Wang. Improving searcher models using mouse cursor activity. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 195–204. ACM, 2012.
- [62] Jeff Huang, Ryen W White, and Susan Dumais. No clicks, no problem: using cursor movements to understand and improve search. In *CHI*, pages 1225–1234, 2011.
- [63] Yun Huang, Michael Yudelson, Shuguang Han, Daqing He, and Peter Brusilovsky. A framework for dynamic knowledge modeling in textbook-based learning. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pages 141–150. ACM, 2016.
- [64] Clark Leonard Hull. *Principles of behavior: an introduction to behavior theory*. 1943.
- [65] Shamsi T Iqbal and Eric Horvitz. Disruption and recovery of computing tasks: field study, analysis, and directions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 677–686. ACM, 2007.
- [66] Kalervo Järvelin, Susan L Price, Lois ML Delcambre, and Marianne Lykke Nielsen. Discounted cumulated gain based evaluation of multiple-query ir sessions. In *Advances in Information Retrieval*, pages 4–15. Springer, 2008.
- [67] Natalie Jhaveri and Kari-Jouko Räihä. The advantages of a cross-session web workspace. In *CHI’05 extended abstracts on human factors in computing systems*, pages 1949–1952. ACM, 2005.
- [68] Jiepu Jiang, Shuguang Han, Jia Wu, and Daqing He. Pitt at trec 2011 session track. In *TREC*, 2011.
- [69] Jiepu Jiang, Daqing He, and Shuguang Han. On duplicate results in a search session. Technical report, DTIC Document, 2012.

- [70] Jiepu Jiang, Daqing He, Diane Kelly, and James Allan. Understanding ephemeral state of relevance. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pages 137–146. ACM, 2017.
- [71] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161. Acm, 2005.
- [72] Rosie Jones and Kristina Lisa Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 699–708, 2008.
- [73] Maryam Kamvar and Shumeet Baluja. Deciphering trends in mobile search. *Computer*, (8):58–62, 2007.
- [74] Shaun K Kane, Amy K Karlson, Brian R Meyers, Paul Johns, Andy Jacobs, and Greg Smith. Exploring cross-device web use on pcs and mobile devices. In *Human-Computer Interaction-INTERACT 2009*, pages 722–735. Springer, 2009.
- [75] Evangelos Kanoulas, Ben Carterette, Paul D Clough, and Mark Sanderson. Evaluating multi-query sessions. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1053–1062. ACM, 2011.
- [76] Diane Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1-2):1–224, 2009.
- [77] Diane Kelly and Colleen Cool. The effects of topic familiarity on information search behavior. In *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 74–75, 2002.
- [78] Paul A Kirschner and Aryn C Karpinski. Facebook® and academic performance. *Computers in human behavior*, 26(6):1237–1245, 2010.
- [79] Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.
- [80] Kenneth R Koedinger, Albert T Corbett, and Charles Perfetti. The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science*, 36(5):757–798, 2012.
- [81] Alexander Kotov, Paul N Bennett, Ryen W White, Susan T Dumais, and Jaime Teevan. Modeling and analysis of cross-session search tasks. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 5–14. ACM, 2011.
- [82] Carol C Kuhlthau. Inside the search process: Information seeking from the user’s perspective. *Journal of the American society for information science*, 42(5):361, 1991.

- [83] Carol Collier Kuhlthau. Accommodating the user’s information search process: challenges for information retrieval system designers. *Bulletin of the American Society for Information Science and Technology*, 25(3):12–16, 1999.
- [84] Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 44(4):978–990, 2012.
- [85] Dmitry Lagun, Chih-Hung Hsieh, Dale Webster, and Vidhya Navalpakkam. Towards better measurement of attention and satisfaction in mobile search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 113–122. ACM, 2014.
- [86] Liangda Li, Hongbo Deng, Yunlong He, Anlei Dong, Yi Chang, and Hongyuan Zha. Behavior driven topic transition for search task identification. In *Proceedings of the 25th International Conference on World Wide Web*, pages 555–565. International World Wide Web Conferences Steering Committee, 2016.
- [87] Zhen Liao, Yang Song, Li-wei He, and Yalou Huang. Evaluating the effectiveness of search task trails. In *Proceedings of the 21st international conference on World Wide Web*, pages 489–498. ACM, 2012.
- [88] Jie Liu, Wenchang Xu, and Yuanchun Shi. Autoweb: automatic classification of mobile web pages for revisitation. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*, pages 191–200. ACM, 2012.
- [89] Jingjing Liu and Nicholas J Belkin. Personalizing information retrieval for multi-session tasks: The roles of task stage and task type. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 26–33. ACM, 2010.
- [90] Jingjing Liu, Nicholas J Belkin, Xiangmin Zhang, and Xiaojun Yuan. Examining users knowledge change in the task completion process. *Information Processing & Management*, 49(5):1058–1074, 2013.
- [91] Xiaoyong Liu and W Bruce Croft. Passage retrieval based on language models. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 375–382. ACM, 2002.
- [92] Xiaozhong Liu, Zhuoren Jiang, and Liangcai Gao. Scientific information understanding via open educational resources (oer). In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 645–654. ACM, 2015.
- [93] Geoffrey R Loftus. Evaluating forgetting curves. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11(2):397, 1985.

- [94] Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri, and Gabriele Tolomei. Identifying task-based sessions in search engine query logs. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 277–286. ACM, 2011.
- [95] Jiyun Luo, Sicong Zhang, and Hui Yang. Win-win search: Dual-agent stochastic game in session search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 587–596. ACM, 2014.
- [96] Yuanhua Lv and ChengXiang Zhai. Positional language models for information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 299–306. ACM, 2009.
- [97] Bonnie Ma Kay and Carolyn Watters. Exploring multi-session web tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1187–1196. ACM, 2008.
- [98] Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.
- [99] Gary Marchionini. *Information seeking in electronic environments*. Number 9. Cambridge university press, 1997.
- [100] Gary Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006.
- [101] Gloria Mark, Victor M Gonzalez, and Justin Harris. No task left behind?: examining the nature of fragmented work. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 321–330. ACM, 2005.
- [102] Allan DR McQuarrie and Chih-Ling Tsai. *Regression and time series model selection*. World Scientific, 1998.
- [103] Kevin Meehan, Tom Lunney, Kevin Curran, and Aiden McCaughey. Context-aware intelligent recommendation system for tourism. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on*, pages 328–331. IEEE, 2013.
- [104] Stefano Mizzaro. Relevance: The whole history. *JASIS*, 48(9):810–832, 1997.
- [105] George D Montañez, Ryen W White, and Xiao Huang. Cross-device search. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1669–1678. ACM, 2014.
- [106] Dan Morris, Meredith Ringel Morris, and Gina Venolia. Searchbar: a search-centric web history for task resumption and information re-finding. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1207–1216. ACM, 2008.

- [107] Ara V Nefian, Luhong Liang, Xiaobo Pi, Liu Xiaoxiang, Crusoe Mao, and Kevin Murphy. A coupled hmm for audio-visual speech recognition. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 2, pages II-2013. IEEE, 2002.
- [108] Hartmut Obendorf, Harald Weinreich, Eelco Herder, and Matthias Mayer. Web page revisitation revisited: implications of a long-term click-stream study of browser usage. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 597–606. ACM, 2007.
- [109] Antti Oulasvirta and Lauri Sumari. Mobile kits and laptop trays: managing multiple devices in mobile information work. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1127–1136. ACM, 2007.
- [110] Zachary A Pardos and Neil T Heffernan. Modeling individualization in a bayesian networks implementation of knowledge tracing. In *User Modeling, Adaptation, and Personalization*, pages 255–266. 2010.
- [111] James Pitkow, Hinrich Schütze, Todd Cass, Rob Cooley, Don Turnbull, Andy Edmonds, Eytan Adar, and Thomas Breuel. Personalized search. *Commun. ACM*, 45(9):50–55, September 2002.
- [112] Filip Radlinski and Susan Dumais. Improving personalized web search using result diversification. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 691–692. ACM, 2006.
- [113] Karthik Raman, Paul N Bennett, and Kevyn Collins-Thompson. Toward whole-session relevance: Exploring intrinsic diversity in web search. In *SIGIR*, pages 463–472, 2013.
- [114] Soo Young Rieh, Jacek Gwizdka, Luanne Freund, and Kevyn Collins-Thompson. Searching as learning: Novel measures for information interaction research. *Proceedings of the American Society for Information Science and Technology*, 51(1):1–4, 2014.
- [115] Meng Rui, Han Shuguang, Huang Yun, He Daqing, and Brusilovsky Peter. Knowledge-based content linking for online textbooks. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 18–25. The Institute of Electrical and Electronics Engineers, 2016.
- [116] Jeffrey M Rzeszutarski, Ed Chi, Praveen Paritosh, and Peng Dai. Inserting micro-breaks into crowdsourcing workflows. In *First AAAI Conference on Human Computation and Crowdsourcing*, 2013.
- [117] Sargol Sadeghi, Roi Blanco, Peter Mika, Mark Sanderson, Falk Scholer, and David Vallet. Predicting re-finding activity and difficulty. In *European Conference on Information Retrieval*, pages 715–727. Springer, 2015.

- [118] Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *Readings in information retrieval*, 24(5):355–363, 1997.
- [119] Dario D Salvucci. On reconstruction of task context after interruption. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 89–92. ACM, 2010.
- [120] Dario D Salvucci, Niels A Taatgen, and Jelmer P Borst. Toward a unified theory of the multitasking continuum: From concurrent performance to task switching, interruption, and resumption. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1819–1828. ACM, 2009.
- [121] Abigail J Sellen, Rachel Murphy, and Kate L Shaw. How knowledge workers use the web. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 227–234. ACM, 2002.
- [122] Chirag Shah and Roberto González-Ibáñez. Exploring information seeking processes in collaborative search tasks. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–7, 2010.
- [123] Chirag Shah and Roberto González-Ibáñez. Evaluating the synergic effect of collaboration in information seeking. In *SIGIR*, pages 913–922, 2011.
- [124] Xuehua Shen, Bin Tan, and ChengXiang Zhai. Context-sensitive information retrieval using implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 43–50. ACM, 2005.
- [125] Jaana Simola, Jarkko Salojärvi, and Ilpo Kojo. Using hidden markov model to uncover processing states from eye movements in information search tasks. *Cognitive Systems Research*, 9(4):237–251, 2008.
- [126] Steven M Smith and Edward Vela. Environmental context-dependent memory: A review and meta-analysis. *Psychonomic bulletin & review*, 8(2):203–220, 2001.
- [127] Yang Song, Hao Ma, Hongning Wang, and Kuansan Wang. Exploring and exploiting user search behavior on mobile and tablet devices to improve search relevance. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1201–1212. International World Wide Web Conferences Steering Committee, 2013.
- [128] David Sontag, Kevyn Collins-Thompson, Paul N Bennett, Ryen W White, Susan Dumais, and Bodo Billerbeck. Probabilistic models for personalizing web search. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 433–442. ACM, 2012.
- [129] Amanda Spink. Multiple search sessions model of end-user behavior: An exploratory study. *Journal of the American Society for Information Science*, 47(8):603–609, 1996.

- [130] Amanda Spink. A user-centered approach to evaluating human interaction with web search engines: an exploratory study. *Information processing & management*, 38(3):401–426, 2002.
- [131] Amanda Spink, Minsoo Park, Bernard J Jansen, and Jan Pedersen. Multitasking during web search sessions. *Information Processing & Management*, 42(1):264–275, 2006.
- [132] Bin Tan, Xuehua Shen, and ChengXiang Zhai. Mining long-term search history to improve search accuracy. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 718–723. ACM, 2006.
- [133] Jaime Teevan. The re: search engine: simultaneous support for finding and re-finding. In *Proceedings of the 20th annual ACM symposium on User interface software and technology*, pages 23–32. ACM, 2007.
- [134] Jaime Teevan, Eytan Adar, Rosie Jones, and Michael AS Potts. Information re-retrieval: repeat queries in yahoo’s logs. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 151–158. ACM, 2007.
- [135] Jaime Teevan, Amy Karlson, Shahriyar Amini, AJ Brush, and John Krumm. Understanding the importance of location, time, and people in mobile local search behavior. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, pages 77–80. ACM, 2011.
- [136] Jaime Teevan, Daniel J Liebling, and Walter S Lasecki. Selfsourcing personal tasks. In *CHI’14 Extended Abstracts on Human Factors in Computing Systems*, pages 2527–2532. ACM, 2014.
- [137] Sarah K Tyler and Jaime Teevan. Large scale query log analysis of re-finding. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 191–200. ACM, 2010.
- [138] Brett van De Sande. Properties of the bayesian knowledge tracing model. *JEDM- Journal of Educational Data Mining*, 5(2):1–10, 2013.
- [139] Hongning Wang, Yang Song, Ming-Wei Chang, Xiaodong He, Ryen W White, and Wei Chu. Learning to extract cross-session search tasks. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1353–1364. International World Wide Web Conferences Steering Committee, 2013.
- [140] Kuansan Wang, Nikolas Gloy, and Xiaolong Li. Inferring search behaviors using partially observable markov (pom) model. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 211–220. ACM, 2010.

- [141] Shuting Wang, Alexander Ororbia, Zhaohui Wu, Kyle Williams, Chen Liang, Bart Pursel, and C Lee Giles. Using prerequisites to extract concept maps from textbooks. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 317–326. ACM, 2016.
- [142] Yu Wang, Xiao Huang, and Ryen W White. Characterizing and supporting cross-device search tasks. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 707–716. ACM, 2013.
- [143] Ryen W White, Paul N Bennett, and Susan T Dumais. Predicting short-term interests using activity-based search context. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1009–1018. ACM, 2010.
- [144] Ryen W White, Susan T Dumais, and Jaime Teevan. Characterizing the influence of domain expertise on web search behavior. In *WSDM*, pages 132–141, 2009.
- [145] Ryen W White, Bill Kules, Steven M Drucker, et al. Supporting exploratory search, introduction, special issue, communications of the acm. *Communications of the ACM*, 49(4):36–39, 2006.
- [146] Ryen W White and Resa A Roth. Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–98, 2009.
- [147] Barbara M Wildemuth. The effects of domain knowledge on search tactic formulation. *Journal of the american society for information science and technology*, 55(3):246–258, 2004.
- [148] Mathew J Wilson and Max L Wilson. A comparison of techniques for measuring sense-making and learning within participant-generated summaries. *Journal of the American Society for Information Science and Technology*, 64(2):291–306, 2013.
- [149] Qiang Wu, Chris JC Burges, Krysta M Svore, and Jianfeng Gao. Ranking, boosting, and model adaptation. Technical report, Microsoft Research, 2008.
- [150] Iris Xie and Soohyung Joo. Transitions in search tactics during the web-based search process. *Journal of the american society for information science and technology*, 61(11):2188–2205, 2010.
- [151] Yunjie Xu and Hainan Yin. Novelty and topicality in interactive information retrieval. *Journal of the American Society for Information Science and Technology*, 59(2):201–215, 2008.
- [152] Louis E Yelle. The learning curve: Historical review and comprehensive survey. *Decision Sciences*, 10(2):302–328, 1979.

- [153] Jeonghee Yi, Farzin Maghoul, and Jan Pedersen. Deciphering mobile search patterns: a study of yahoo! mobile search queries. In *Proceedings of the 17th international conference on World Wide Web*, pages 257–266. ACM, 2008.
- [154] Michael V Yudelson, Kenneth R Koedinger, and Geoffrey J Gordon. Individualized bayesian knowledge tracing models. In *Artificial Intelligence in Education*, pages 171–180, 2013.
- [155] Zhen Yue, Shuguang Han, and Daqing He. Modeling search processes using hidden states in collaborative exploratory web search. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 820–830. ACM, 2014.
- [156] Zhen Yue, Shuguang Han, Jiepu Jiang, and Daqing He. Search tactics as means of examining search processes in collaborative exploratory web search. In *Proceedings of the 5th Ph. D. workshop on Information and knowledge*, pages 59–66. ACM, 2012.
- [157] Chengxiang Zhai and John Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 403–410. ACM, 2001.
- [158] Xiangmin Zhang, Jingjing Liu, Michael Cole, and Nicholas Belkin. Predicting users’ domain knowledge in information retrieval using multiple regression analysis of search behaviors. *Journal of the Association for Information Science and Technology*, 66(5):980–1000, 2015.
- [159] Yi Zhang, Jamie Callan, and Thomas Minka. Novelty and redundancy detection in adaptive filtering. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 81–88. ACM, 2002.