EPHEMERAL RELEVANCE AND USER ACTIVITIES IN A SEARCH SESSION

by

Jiepu Jiang

B.S., Wuhan University, 2007

M.S., Wuhan University, 2009

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

2016

UNIVERSITY OF PITTSBURGH

SCHOOL OF INFORMATION SCIENCES

This dissertation was presented

by

Jiepu Jiang

It was defended on

May 11, 2016

and approved by

Daqing He, Ph.D., Associate Professor, Univeristy of Pittsburgh

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

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

Diane Kelly, Ph.D., Professor, University of North Carolina at Chapel Hill

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

EPHEMERAL RELEVANCE AND USER ACTIVITIES IN A SEARCH SESSION

Jiepu Jiang, PhD

University of Pittsburgh, 2016

We study relevance judgment and user activities in a search session. We focus on ephemeral relevance—a contextual measurement regarding the amount of useful information a searcher acquired from a clicked result at a particular time—and two primary types of search activities—query reformulation and click. The purpose of the study is both explanatory and practical. First, we examine the influence of different factors on ephemeral relevance and user activities in a search session. Second, we leverage short-term search history and implicit feedback in a session to predict ephemeral relevance and future search activities. The main findings include:

1. As a contextual usefulness measurement, ephemeral relevance differs from both topical relevance judgment and context-independent usefulness assessment. We show ephemeral relevance significantly relates to a wide range of factors, including topical relevance, novelty, understandability, reliability, effort spent, and search task. The difference between ephemeral relevance and context-independent usefulness assessment is linked to judgment criteria, novelty, effort spent, and changes in user's perceptions of a search result.
2. Ephemeral relevance can be predicted accurately using implicit feedback signals without any manual explicit judgments. We generalize existing implicit feedback methods from using information related to a single result to those based on user activities in a whole session, achieving a correlation as high as 0.5 between the predicted and real judgments.
3. We show choices of word changes in query reformulation and click decisions significantly relate to recent search history, such as the contents and effectiveness of previous search

queries, the contents of the results viewed and clicked in previous searches, etc.

4. Leveraging short-term search history in a session and other information, we can predict word changes in query reformulation and click decisions with different levels of accuracies.

These findings help disclose and explain the dynamics of relevance and user activities in a search session. The developed techniques provide effective support for developing interactive IR systems.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

This dissertation is intended to help researchers in the field of information retrieval (IR), especially those working on interactive IR, contextual IR, and personalized IR, understand the dynamics of relevance and user activities in a short-term search session (typically spanning several queries but no longer than 30 minutes). It examines the influence of different factors on relevance and user activities, develops techniques for predicting relevance and user activities in a search session, and offers suggestions to IR practice.

Much of the work has come from my Ph.D. study at the University of Pittsburgh from 2009 to 2013, although I was at the University of Massachusetts most of the time when I wrote the dissertation. I had a memorable experience at Pitt. I sincerely thank all the helps I received from the many wonderful mentors, friends, and colleagues during my Ph.D. study.

I deeply appreciate Daqing He, my advisor, for his endless help, support, guidance, encouragement, understanding, and patience, for which I am indebted forever. I missed him and our collaboration all the time after I left Pitt.

I also appreciate other dissertation committee members' help, related and unrelated to this dissertation. My initial exploration on user and social perspectives of IR started from a doctoral seminar taught by Peter Brusilovsky. I learned many knowledge of experiment design and user study from Diane Kelly's articles. Peter Brusilovsky, Diane Kelly, and Yu-ru Lin also provided many valuable suggestions in conceptualizing, formulating, and finalizing this dissertation.

I would also like to thank the much help I received from Ahmed Hassan Awadallah, Ryen White, Rosie Jones, Xiaolin Shi, Imed Zitouni, and many others during my internship at Microsoft Research. This wonderful experience had a great influence on my research and this dissertation. I also appreciate the many discussions with Ben Carterette, Grace Hui Yang,

Jiyun Luo, and others who actively involved in the TREC session tracks, where I started my research on topics related to search session.

I also appreciate my labmates and friends at Pitt, especially Shuguang Han, Zhen Yue, and Wei Jeng. Our discussions are always fruitful, and our collaborations are mostly productive.

My Ph.D. study and research also received much help from many others that I worked with, especially Cassidy Sugimoto, Sherry Koshman, and Jung Sun Oh. I would also like to thank Debbie Day for all her help in the program. I also appreciate ACM SIGIR for offering me generous travel grants a number of times, which enable my travels to SIGIR, CHIIR, and WSDM. I also thank James Allan and the University of Massachusetts for their consideration, which makes the completion of this dissertation possible.

Last but not the least, I thank my beloved wife Chaoqun Ni for everything she did. She helped me annotate data; she offered numerous suggestions on statistical analysis and writing; she made me delicious food and coffee; she kept our home cozy and clean . . . I also thank Aaron for coming to the family and his innocent smile, which comforted me all the time when I felt frustrated.

1.0 INTRODUCTION

Many researchers have argued that there is not much room to improve *ad hoc search*, a classic but reduced form of information retrieval (IR) problem focusing on individual and one-off requests. For example, Armstrong, Moffat, Webber, and Zobel (2009) examined published results during 1998–2008 using TREC[1] collections but found no clear improvements over the best-reported results in past TREC evaluations. Trotman and Keeler (2011) reported that Okapi BM25 (S. E. Robertson, Walker, Jones, Hancock-Beaulieu, & Gatford, 1994), a popular ad hoc retrieval model, has already achieved a performance comparable to human ranking of search results.

One of the promising directions to further improve IR systems is to place search and user into contexts (Pitkow et al., 2002; Allan et al., 2003; Ingwersen & Järvelin, 2005; Bennett, Collins-Thompson, Kelly, White, & Zhang, 2015). We study interactive IR in the context of a search session, which usually involves multiple consecutive requests (queries) targeting the same problem. Compared with a single query (the focus of ad hoc search), a search session helps IR systems better understand and model users' information needs and interaction.

1.1 BACKGROUND AND PURPOSE

This study is related to the topic *contextual search* (Bennett et al., 2015) and *personalized search* (Pitkow et al., 2002) at large, where most related studies used some particular types of context information to improve search systems or examined the impact of some context factors on retrieval effectiveness, user activities, or search process.

[1] Text REtrieval Conference, the largest open information retrieval evaluation campaign at the time.

We study search problems in a *search session*, which is defined as a period that involves one or multiple consecutive rounds of searches by the same user targeting a consistent *task*. A task is an actual problem of the searcher that motivates the search process (such as "how to choose a dehumidifier"), and the search process may help the user address the task by retrieving useful information. § 2.1 reviews search session and task in detail.

A search session provides short-term and problem-related context information such as the user's recent search history and attributes of the search task. The context information studied here is similar to that in previous studies such as J. Liu et al. (2010), C. Liu, Belkin, and Cole (2012), D. Guan, Zhang, and Yang (2013), and Cole, Hendahewa, Belkin, and Shah (2014). The scope is different from studies that focused on other types of context information such as users' long-term interests (Gauch, Chaffee, & Pretschner, 2003; Teevan, Dumais, & Horvitz, 2005), locations (Bennett, Radlinski, White, & Yilmaz, 2011), domain expertise (White, Dumais, & Teevan, 2009), reading levels (Collins-Thompson, Bennett, White, de la Chica, & Sontag, 2011), social connections (Carmel et al., 2009), gender (Lorigo et al., 2006), and so on.

We study two core aspects of an IR system within the scope of a search session:

- **Search result relevance**. A primary goal of an IR system is to help searchers retrieve relevant results and acquire useful information. We examine the criteria and influencing factors of search result relevance in a search session and develops effective techniques for predicting search result relevance. More specifically, we examine *ephemeral relevance* (ERel), a contextual measurement regarding the amount of useful information a searcher acquired from a clicked result at a particular time of a session. We compare ERel with the state-of-the-art approaches for relevance judgments, which focus on topical relevance and do not involve search contexts into relevance judgments, such as the methods adopted by many TREC ad hoc search tasks (Harman, 1992a; Hawking, Voorhees, Craswell, & Bailey, 1999; Clarke, Craswell, & Soboroff, 2009).
- **User activities**. We study two specific types of user activities in a search session—*query reformulation*[2] and *click*[3]. They are the two primary types of activities that users rely on

[2] Query reformulation is the activity of formulating a search query where there was a previous one.
[3] Here a click specifically refers to a click on a search result displayed on a search result page (SERP).

to make progress and acquire useful information in a search session. It is also easy and practical for current search engines to record these two types of activities, which makes them feasible for a variety of purposes such as offering query suggestions (Baeza-Yates, Hurtado, & Mendoza, 2004; Bar-Yossef & Kraus, 2011; Cao et al., 2008; Mei, Zhou, & Church, 2008; Shokouhi & Radinsky, 2012) and inferring the quality of search results and the goodness of search services (Joachims, Granka, Pan, Hembrooke, & Gay, 2005; Chapelle & Zhang, 2009; Craswell, Zoeter, Taylor, & Ramsey, 2008; Guo, Liu, Kannan, et al., 2009; G. Dupret & Liao, 2010; Chao et al., 2015).

Our purpose is twofold, 1) to offer new understandings of search result relevance and the two types of user activities, especially their influencing factors and dynamics in a session, and 2) to develop effective approaches for predicting ERel judgments and the two types of user activities in a search session. This chapter elaborates the purpose of the study as four specific research questions.

1.2 PROBLEM STATEMENT

1.2.1 Understanding Ephemeral Relevance

Relevance is a key construct in information retrieval. Most current IR systems are designed and optimized to rank search results by relevance. Most offline methods for evaluating search systems are also based on relevance judgments. Previous studies had numerous discussions on the notion of relevance (Borlund, 2003; Mizzaro, 1997, 1998; Saracevic, 1996, 1975, 2007) and its measurement in IR (Xu & Chen, 2006; Xu & Wang, 2008; Y. Zhang, Zhang, Lease, & Gwizdka, 2014). Most studies acknowledged that the relevance of a search result depends on not only topic aboutness but also many other factors such as novelty, understandability, reliability, search task, search context, users' interaction with the search result, and so on.

However, in practice, relevance is usually assessed by external annotators without search contexts using criteria that focus on topicality. A typical example is the methods used in various TREC ad hoc search tasks in the past two decades (Harman, 1992a; Hawking et al., 1999; Clarke et al., 2009). For example, the TREC 2014 web track (Collins-Thompson, Macdonald, Bennett, Diaz, & Voorhees, 2014) asked external assessors to judge search result relevance using the following criteria, where the numerical numbers refer to priority scores of each level in test collection-based IR evaluation:

- **Nav** (4): the correct home page for a navigational query (the purpose of the query is to locate a home page).
- **Key** (3): a key page or site that is dedicated to the topic and authoritative.
- **HRel** (2): the web page provides substantial information on the topic.
- **Rel** (1): the page provides some information on the topic, which may be minimal.
- **Non** (0): the page does not provide useful information on the topic, and the page is not a spam or junk web page.
- **Junk** (-2): a spam or junk web page.

The past decade witnessed a flourish of contextual search techniques. However, many of them were evaluated in a context-independent manner, based on static and topical relevance judgments similar to the one presented above. For example, the TREC session track (Carterette, Clough, Hall, Kanoulas, & Sanderson, 2016) aimed to improve the effectiveness of search systems based on a session's search history. They adopted the same procedure and criteria of relevance judgments used by the TREC ad hoc search tracks. Such a topical and context-independent approach for relevance judgments has three key limitations:

- First, this approach fails to address the dynamic nature of search result relevance in a session. Many previous studies suggested that the relevance of a search result changes during a search process. For example, Taylor, Cool, Belkin, and Amadio (2007) collected users' self-reported relevance judgments in long search processes. They found that users reported criteria such as "Interest", "Ability to understand", and "Teaches" more frequently at early stages of a search process than in later stages. They also reported that other criteria such as "Specificity", "Recency", and "Depth/Scope" are less frequently at early stages of a session.
- Second, topical relevance is not necessarily pertinent to the actual usefulness of a result regarding the helpfulness for addressing an actual task. For example, for a query "*dehumidifier*", an article discussing how dehumidifier works may not be helpful to a user

who just wants to buy a cheap dehumidifier. Note that the goal of a search task may also evolve during a search session (e.g., the user may first explore how dehumidifier works and then look for an appropriate one to purchase). Thus this problem may also intertwine with the first limitation.

- Third, the relevance judgments collected without a context do not take into account the actual process of acquiring information from the result. Relevance is not only a characteristic or property of a result, but also a measurement of the effectiveness of communication (Saracevic, 2007)—in the case of IR and a search result, relevance measures the effectiveness of the communication between a searcher and a result—the process that the searcher acquires useful information from the result.

To address these issues, we measure search result relevance in a contextual manner and focus on usefulness as the primary criterion. We define *ephemeral relevance* (ERel) as the amount of useful information a user acquired from a clicked result at a specific time of a search session. We collected ERel through a laboratory user study (User Study 1). In User Study 1, the experimental search system elicited users to provide ERel judgments after they visited a search result and switched back to the system. This was the time point that the user had just acquired relevant information (if any) from the visited result. The user study also collected searchers' relevance judgments after a search session for comparison. § 3.4 introduces the design of User Study 1 in detail.

We use data collected from User Study 1 to answer the following research question:

RQ1 *– On what criteria do users assess ephemeral relevance (ERel)? How does ephemeral relevance judgment differ from topical relevance judgment and context-independent usefulness judgment?*

The purpose of RQ1 is to offer a better understanding of ERel and its differences with topical relevance judgments and context-independent usefulness judgments. RQ1 also helps to understand the dynamics of relevance in a search session.

1.2.2 Predicting Ephemeral Relevance

RQ1 aims to enhance current understandings of search result relevance in a search session. RQ2 further studies methods for predicting ERel judgments, which can be applied to search result ranking and search engine evaluation.

RQ2 – *What information other than users' explicit judgment tells the ERel of a search result? How well do these information predict ERel judgments?*

RQ2 is motivated by the cost and selection bias of collecting ERel judgments. Despite a few theoretical advantages, ERel needs to be assessed by real users in a genuine search context, which requires a more complex setting than collecting TREC-style relevance judgments. In addition, to maintain a genuine search context for ERel judgments, we can only rely on the searchers themselves to click on a result and harvest its ERel judgment. It is difficult to manipulate the experiment setting to collect ERel judgments for an arbitrary set of documents.

Enlightened by the success of implicit feedback techniques in web search (Kelly & Teevan, 2003; Kelly & Belkin, 2004; Joachims et al., 2005; Agichtein, Brill, & Dumais, 2006; Agichtein, Brill, Dumais, & Ragno, 2006), we study techniques for predicting ERel judgments based on implicit feedback information in a search session. More specifically, we study two ERel prediction settings. The two settings use different information in a search session for prediction and target to help search result ranking and search engine evaluation.

- **[Setting1]** – Predicting the ERel of a result based on current search query and past search history in the current session.
- **[Setting2]** – Predicting the ERel of a result based on the whole session's search log.

Setting1 uses the current query and past search history in a session to predict ERel. We can easily apply techniques of setting 1 to rank search results because all the information for prediction are available while we rank search results for a query. One can either rank results by the predicted ERel or combine the ERel prediction with other features to train ranking models. Setting2 addresses the high cost and selection bias of ERel judgments by inferring ERel from implicit feedback signals in a search session. In addition, RQ2 also examines ERel prediction using both implicit and explicit feedback information.

1.2.3 Understanding User Activities in a Search Session

Many previous studies examined the influence of search task on user behavior (J. Liu et al., 2010; Cole et al., 2011, 2014; Jiang, He, & Allan, 2014). However, it is difficult to personalize

IR systems based on these search tasks due to the lack of effective automatic methods for recognizing these search tasks at the beginning of a search session. For example, C. Liu (2012) reported an accuracy of only 67.2% for recognizing task types using a whole session's search history, and it remains unclear how well one can predict task type at the beginning of a search session (e.g., using only the first one or two queries of a session).

Many other examined factors have the same issue—we can control and examine the influence of these factors in laboratory studies, but the findings provide limited insights on personalization techniques due to the difficulty of recognizing the attributes of these variables on the fly. For example, in a study, C. Liu, Liu, and Belkin (2014) also reported limited accuracies of predicting search task difficulty (another widely studied influencing factor of search behavior) at different stages of a session without knowing task type (54.5%–60.7%). Other factors, such as searchers' domain knowledge and demographic information, may be readily available for registered search engine users, but search engines need to provide both non-anonymous and anonymous services.

In this study, we focus on the influence of session search history (past user activities and results in a search session) on future search activities in a session. We explore the following form of question: are users more likely to do X if Y happened in the session? Studying the influence of session search history on future search activities has a clear advantage—it does not rely on any automatic techniques to recognize variables in session history.

We study the influence of session search history on future search activities, with other factors such as task type and user's topic familiarity as controls. We focus on two primary search activities—query reformulation and clicking. We study the following two research questions:

- **RQ3** – What affects users' choices of word changes in a query reformulation during a search session? Can we predict such choices of word changes?
- **RQ4** – What affects users' decisions on whether or not to click on a search result's link after viewing its summary displayed on a SERP? Can we predict such decisions?

In order to answer RQ3 and RQ4, we conducted another user study (User Study 2), where we assigned searchers to work on tasks of different types and recorded their search behavior. A major difference between User Study 1 and User Study 2 lies in that in User Study 2, participants were not interrupted for ERel judgments during a session to ensure that the collected search activities are natural. Another difference is that User Study 2 recorded users' eye-movement data in order to examine user's click decisions. This is necessary because observed click activities are biased. For example, sometimes users did not click on a search result simply because they did not view it.

Studies on RQ3 and RQ4 enhance current understandings on the two important types of user activities in a session.

1.2.4 Predicting Future User Activities in a Search Session

In addition to studying the influence of session search history on future user activities, RQ3 and RQ4 also develop techniques for predicting future user activities based on session search history. More specifically, RQ3 examines two subtasks:

- predicting whether searchers would remove or retain a particular word in the next query;
- predicting whether or not searchers would add a particular word in the next query.

These two subtasks, altogether, predict all vocabulary differences in a query reformulation. They have wide potential applications such as query auto-completion and term-level query suggestions. In addition, the technique may also help simulate user interaction in a search session.

RQ4 predicts whether or not searchers would click on a search result provided that they viewed its summary. The prediction only relies on information in a search session. This technique can also be applied to different applications, for example, to reduce click errors in search result ranking (Jiang & Allan, 2016c). It also offers suggestions to improve click models (Chuklin, Markov, & de Rijke, 2015).

1.3 SCOPE

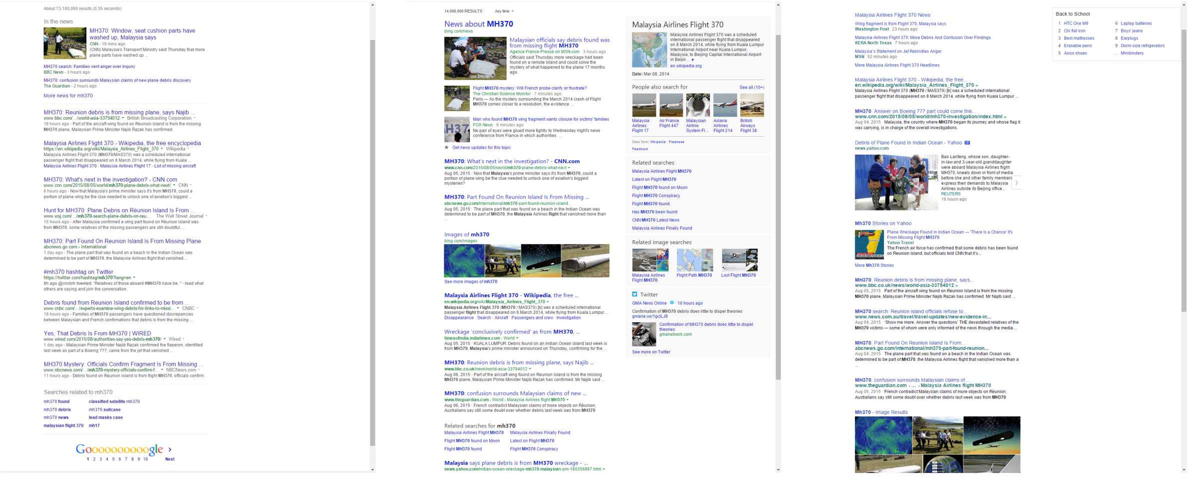
Search is full of variability. It is difficult to cover all possibilities in one study. This section clarifies the scope of the study from the following aspects: user, system, and task.

1.3.1 User

This study targets individual users of general web search engines. Yet it is difficult to know the actual population or sample from it. The user studies used convenience sampling. The scope was restricted to English-language web search engines due to the location of the researcher. The actual participants were mostly representative of users from 20 to 35 years old with college or higher education. About half of the participants were studying information related disciplines (e.g., library & information science, computer science). To exclude the influence of language efficiency on relevance judgment and search interaction, we restrict participants of both User Study 1 and User Study 2 to English native speakers. We expect differences may exist when generalizing to non-native speakers, but this is beyond the scope of the study.

1.3.2 System

We study the proposed research questions when the target users search textual information in general web search engines. However, for logging purposes, the user studies used experimental search systems instead of real web search engines. The experimental search systems redirected user queries to Google and returned the "10-blue links" and query suggestions (if any). The systems displayed search results and query suggestions in the same way they would appear on Google (e.g., query terms were highlighted). We removed other types of results such as verticals, ads, instant answers, related entities, and so on. We excluded other types of results either because they are subsidiary (e.g., ads) or because current web search engines did not all agree on how to display them on a SERP (e.g., the types of verticals and their positions on SERPs vary a lot in Google, Bing, and Yahoo at the time of the study). For example, Figure 1 shows SERPs returned by Google, Bing, and Yahoo for the same query "mh370". Both Google and Bing put news verticals at the top of the SERPs, while Yahoo put news verticals at the middle of the SERP. In addition, Bing provides related entities on the right of the 10-blue links, while both Google and Yahoo did not. Some studies showed that different elements on the SERP can affect user behavior (Z. Liu, Liu, Zhou, Zhang, & Ma, 2015), but this dissertation excludes such effects in the experimental design.



About 73,100,000 results (0.35 seconds)

In the news

MH370: Window, seat cushion parts have washed up, Malaysia says
CNN - 19 mins ago
(CNN) Malaysia's Transport Ministry said Thursday that more plane parts have washed up ...

MH370 search: Families vent anger over inquiry
BBC News - 3 hours ago

MH370: confusion surrounds Malaysian claims of new plane debris discovery
The Guardian - 2 hours ago

More news for mh370

MH370: Reunion debris is from missing plane, says Najib ...
www.bbc.com/.../world-asia-33794012 ▾ British Broadcasting Corporation ▾
18 hours ago - Part of the aircraft wing found on Reunion Island is from the missing **MH370** plane, Malaysian Prime Minister Najib Razak has confirmed.

Malaysia Airlines Flight 370 - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Malaysia_Airlines_Flight_370 ▾ Wikipedia ▾
Malaysia Airlines Flight 370 (**MH370**/MAS370) was a scheduled international passenger flight that disappeared on 8 March 2014, while flying from Kuala ...
Malaysia Airlines Flight 370 - Malaysia Airlines Flight 17 - List of missing aircraft

MH370: What's next in the investigation? - CNN.com
www.cnn.com/2015/08/05/world/**mh370**-plane-debris-what-next/ ▾ CNN ▾
6 hours ago - Now that Malaysia's prime minister says it's from **MH370**, could a portion of plane wing be the clue needed to unlock one of aviation's biggest ...

Hunt for MH370: Plane Debris on Réunion Island Is From ...
www.wsj.com/.../**mh370**-search-plane-debris-on-reu... The Wall Street Journal ▾
15 hours ago - After Malaysia confirmed a wing part found on Réunion Island was from **MH370**, some relatives of the missing passengers are still doubtful, ...

MH370: Part Found On Reunion Island Is From Missing Plane
abcnews.go.com › International
1 day ago - The plane part that was found on a beach in the Indian Ocean was determined to be part of **MH370**, the Malaysia Airlines flight that vanished ...

#mh370 hashtag on Twitter
https://twitter.com/hashtag/**mh370**?lang=en ▾
8h ago @cnnbrk tweeted: "Relatives of those aboard **#MH370** have be.." - read what others are saying and join the conversation.

Debris found from Reunion Island confirmed to be from ...
www.cnbc.com/.../experts-examine-wing-debris-for-links-to-missi... ▾ CNBC ▾
18 hours ago - Families of **MH370** passengers have questioned discrepancies between Malaysian and French confirmations that debris is from the missing ...

Yes, That Debris Is From MH370 | WIRED
www.wired.com/2015/08/authorities-say-yes-debris-**mh370**/ ▾ Wired ▾
1 day ago - Malaysian Prime Minister Najib Razak confirmed the flaperon, identified last week as part of a Boeing 777, came from the jet that vanished ...

MH370 Mystery: Officials Confirm Fragment Is From Missing ...
www.nbcnews.com/.../**mh370**-mystery-officials-confirm-f... ▾ NBCNews.com ▾
11 hours ago - Debris found on Reunion Island is from flight **MH370**, officials confirm.

Searches related to mh370

mh370 **found**	**classified satellite** mh370
mh370 **debris**	mh370 **suitcase**
mh370 **news**	**lead masks case**
malaysian flight 370	**mh17**

(a) Google

14,000,000 RESULTS Any time ▾

News about **MH370**
bing.com/news

Malaysian officials say debris found was from missing flight **MH370**
Agence France-Presse on MSN.com · 3 hours ago
Officials said Thursday more wreckage had been found on a remote island and could solve the mystery of what happened to the plane 17 months ago.

Flight **MH370** mystery: Will French probe clarify or frustrate?
The Christian Science Monitor · 7 minutes ago
Paris — As the mystery surrounding the March 2014 crash of Flight **MH370** comes closer to a resolution, the existence ...

Man who found **MH370** wing fragment wants closure for victims' families
FOX News · 6 minutes ago
No pair of eyes were glued more tightly to Wednesday night's news conference from France in which authorities ...

★ Get news updates for this topic

MH370: What's next in the investigation? - **CNN.com**
www.cnn.com/2015/08/05/world/**mh370**-plane-debris-what-next ▾
Aug 05, 2015 · Now that **Malaysia**'s prime minister says it's from **MH370**, could a portion of plane wing be the clue needed to unlock one of aviation's biggest mysteries?

MH370: Part Found On Reunion Island Is From Missing ...
abcnews.go.com/International/**mh370**-part-found-reunion-island...
Aug 05, 2015 · The plane part that was found on a beach in the Indian Ocean was determined to be part of **MH370**, the **Malaysia** Airlines **flight** that vanished more than ...

Images of **mh370**
bing.com/images
See more images of **mh370**

Malaysia Airlines Flight 370 - **Wikipedia**, the free ...
en.wikipedia.org/wiki/**Malaysia_Airlines_Flight_370** ▾
Malaysia Airlines Flight 370 (**MH370** / MAS370) [b] was a scheduled international passenger **flight** that disappeared on 8 March 2014, while flying from Kuala Lumpur ...
Disappearance · Search · Aircraft · Passengers and crew · Investigation

Wreckage 'conclusively confirmed' as from **MH370**, ...
timesofindia.indiatimes.com › World ▾
Aug 05, 2015 · KUALA LUMPUR: Debris found on an Indian Ocean island last week is from **MH370**, **Malaysia**'s prime minister announced on Thursday, confirming for the ...

MH370: Reunion debris is from missing plane, says Najib ...
www.bbc.co.uk/news/world-asia-33794012 ▾
Aug 06, 2015 · Part of the aircraft wing found on Reunion Island is from the missing **MH370** plane, Malaysian Prime Minister Najib Razak has confirmed. Mr Najib said ...

MH370: confusion surrounds Malaysian claims of new ...
www.theguardian.com › World › Malaysia Airlines flight **MH370** ▾
Aug 06, 2015 · French contradict Malaysian claims of more objects on Réunion; Australians say still some doubt over whether debris last week was from **MH370**

Related searches for **mh370**

Malaysia Airlines Flight **MH370**	Malaysia Airlines Finally Found
Flight **MH370** found on Moon	Latest on Flight **MH370**
Flight **MH370** found	Flight **MH370** Conspiracy

Malaysia says plane debris is from **MH370** wreckage - ...
news.yahoo.com/indian-ocean-wreckage-**mh370**-**malaysia**n-pm-180356887.html ▾

Malaysia Airlines Flight 370
Malaysia Airlines Flight 370 was a scheduled international passenger flight that disappeared on 8 March 2014, while flying from Kuala Lumpur International Airport near Kuala Lumpur, Malaysia, to Beijing Capital International Airport in Beijin... +
en.wikipedia.org
Date: Mar 08, 2014

People also search for See all (10+)

Malaysia Airlines Flight 17	Air France Flight 447	Malaysian Airline System Fl...	Asiana Airlines Flight 214	British Airways Flight 38

Data from: Wikipedia · Freebase
Feedback

Related searches
Malaysia Airlines Flight **MH370**
Latest on Flight **MH370**
Flight **MH370** found on Moon
Flight **MH370** Conspiracy
Flight **MH370** found
Has **MH370** been found
CNN **MH370** Latest News
Malaysia Airlines Finally Found

Related image searches

Malaysia Airlines Flight **MH370**	Flight Path **MH370**	Lost Flight **MH370**

Twitter
GMA News Online · 10 hours ago
Confirmation of **MH370** debris does little to dispel theories
gmane.ws/1gclLJ8
Confirmation of MH370 debris does little to dispel theories
gmanetwork.com
See more on Twitter

(b) Bing

Malaysia Airlines Flight 370 News
Wing fragment is from Flight 370, Malaysia says
Washington Post 23 hours ago
Malaysia Airlines Flight 370: More Debris And Confusion Over Findings
KERA North Texas 7 hours ago
Malaysia's Statement on Jet Rekindles Anger
MSN 52 minutes ago
More Malaysia Airlines Flight 370 Headlines

Malaysia Airlines Flight 370 - Wikipedia, the free...
en.wikipedia.org/wiki/Malaysia_Airlines_Flight_370 ⌄
Malaysia Airlines Flight 370 (**MH370** / MAS370) [b] was a scheduled international passenger flight that disappeared on 8 March 2014, while flying from Kuala Lumpur ...

MH370: Answer on Boeing 777 part could come this...
www.cnn.com/2015/08/05/world/mh370-investigation/index.html ⌄
Aug 04, 2015 · Malaysia, the country where **MH370** began its journey and whose flag it was carrying, is in charge of the overall investigations.

Debris of Plane Found In Indian Ocean - Yahoo
news.yahoo.com
Bao Lanfang, whose son, daughter-in-law and 3-year-old granddaughter were aboard Malaysia Airlines flight MH370, kneels down in front of media before she and other family members express their demands to Malaysia Airlines outside its Beijing office...
REUTERS
16 hours ago

Mh370 Stories on Yahoo
Plane Wreckage Found in Indian Ocean — 'There Is a Chance' It's From Missing Flight **MH370**
Yahoo Travel
The French air force has confirmed that some debris has been found on Reunion Island, but officials told CNN that it's...
More **Mh370** Stories

MH370: Reunion debris is from missing plane, says...
www.bbc.co.uk/news/world-asia-33794012 ⌄
Aug 05, 2015 · Part of the aircraft wing found on Reunion Island is from the missing **MH370** plane, Malaysian Prime Minister Najib Razak has confirmed. Mr Najib said ...

MH370 search: Reunion island officials refuse to...
www.news.com.au/travel/travel-updates/new-evidence-in...
Aug 04, 2015 · 'Show me more. Answer the questions' THE devastated relatives of the **MH370** victims — some of whom were only informed of the news through the media ...

MH370: Part Found On Reunion Island Is From...
abcnews.go.com/International/mh370-part-found-reunion...
Aug 04, 2015 · The plane part that was found on a beach in the Indian Ocean was determined to be part of **MH370**, the Malaysia Airlines flight that vanished more than a ...

MH370: confusion surrounds Malaysian claims of...
www.theguardian.com › ... › Malaysia Airlines flight MH370
Aug 05, 2015 · French contradict Malaysian claims of more objects on Réunion; Australians say still some doubt over whether debris last week was from **MH370**

Mh370 - Image Results

Back to School

1	HTC One M9	6	Laptop batteries
2	Chi flat iron	7	Boys' jeans
3	Best mattresses	8	Earplugs
4	Erasable pens	9	Dorm-size refrigerators
5	Asics shoes	...	Minibinders

(c) Yahoo

Figure 1: SERPs returned by Google, Bing, and Yahoo for a query "mh370".

In addition, this study does not consider searches on mobile devices, or users using input devices other than keyboard and mouse (e.g., touch screen, voice, and etc.). All the user studies were conducted on normal desktop computers.

1.3.3 Task

Our user studies assigned participants to work on four types of search tasks following Y. Li and Belkin's (2008) faceted task classification scheme. These tasks differ in two dimensions: search goal and targeted product (what types of information the task is looking for). Search goal is either *specific* or *amorphous*. Targeted product is either *factual* or *intellectual*. Different combinations of search goal and targeted product define four types of search tasks (2 × 2). § 3.3 introduces search task settings in detail.

We note the adopted search task setting is only one option. Many other studies (J. Liu & Belkin, 2010; J. Liu et al., 2010; C. Liu et al., 2012; J. Liu, Belkin, Zhang, & Yuan, 2013; C. Liu et al., 2014) also followed Y. Li and Belkin's (2008) faceted task classification scheme, although with slight differences in the settings. But some previous studies also used other types of tasks (Freund, Toms, & Clarke, 2005; Wu, Kelly, Edwards, & Arguello, 2012; Kelly, Arguello, Edwards, & Wu, 2015). We do not intend to claim superiority of any settings.

1.4 OUTLINE

This dissertation is organized as follows: Chapter 2 reviews related work; Chapter 3 introduces the purpose, framework, and methods of the study; Chapter 4 summarizes the collected data; Chapter 5, 6, 7, and 8 addresses RQ1, RQ2, RQ3, and RQ4, respectively; finally, we draw conclusions in Chapter 9.

2.0 RELATED WORK

This chapter reviews related work in three areas. § 2.1 discusses studies related to a search session. § 2.2 reviews studies in relevance judgment. § 2.3 summarizes previous work on user behavior modeling and, in particular, those related to query reformulation and click.

2.1 SEARCH SESSION

2.1.1 Definition

Search session, or shortly session, is a key construct discussed in this dissertation. Here a *search session* refers to a period that involves one or multiple consecutive rounds of searches from a user dedicated to the same problem. This definition is similar to those adopted in most previous studies using data collected from laboratory user studies (J. Liu et al., 2010; C. Liu et al., 2012, 2014; Cole et al., 2014; Jiang et al., 2014). However, the term is also used differently elsewhere, mostly in studies of web search using query logs (Catledge & Pitkow, 1995; White & Drucker, 2007; Jones & Klinkner, 2008; Kotov, Bennett, White, Dumais, & Teevan, 2011). For example, the operational setting of a search session extracted from web search logs usually involves consecutive requests from the same client identification (e.g., IP address and cookies) where the duration between adjacent queries does not exceed certain thresholds (e.g., 30 minutes). In such a case, searches within a session may not necessarily come from the same user nor target the same problem.

Despite the inconsistencies in its definition, while being used, the term "search session" usually implies more than one rounds of user requests and interaction, where a short-term search history is available. This is opposite to studies focusing on individual search requests, for example, the ad hoc search problem (Harman, 1992a; Hawking et al., 1999; Clarke et al., 2009). Some used “multi-query session” (Kanoulas, Carterette, Clough, & Sanderson, 2011) or “multiple-query session” (Järvelin, Price, Delcambre, & Nielsen, 2008) to emphasize this distinction. But conceptually it is inappropriate to divide search by the number of issued queries. The number of requests is not an intrinsic attribute of search but an outcome depending on problem complexity, system effectiveness, user’s domain knowledge, search skill, and other factors. The definition adopted here emphasizes grouping related searches on the same problem, or say, belonging to the same search task. When the problem is simple, a session can be as short as one request, e.g., looking for the home page of a website (navigational search). “Multi-query session” happens when information need is not fulfilled by the first request and the searcher decides to continue searching.

The use of the term “search session” in research also emphasizes a short-term, problem-related search history, which is different from studies focusing on long-term search history (usually spans at least several days in research) (Gauch et al., 2003; Teevan et al., 2005; Tan, Shen, & Zhai, 2006; Dou, Song, & Wen, 2007). The difference lies is that long-term search history mainly indicates user’s general interests, expertise, and preferences, but not necessarily relates to the problem at hand. A search session provides a short-term search history related to the ongoing task, which may or may not relate to the user’s general interests, expertise, and preferences. For example, a software engineer with general interests and expertise on information technology may search topics related to root canel before a dentist appointment.

It should be clarified that the notion and setting of a search session being adopted in this study is an ideal case. This helps simplify study and analysis. It is straightforward to collect data of search sessions pertaining to this definition from laboratory user studies, where researchers can control user identity and search task. But in practical scenarios (e.g., using web search logs), it requires extra effort to obtain data of such sessions. The reasons mainly include:

First, user identification may be inaccurate or unavailable. For example, most search engines at the time have to rely on IP address and browser cookies to distinguish different users when they did not log in. This may fail to differentiate multiple searchers from the same device. It may also fail to recognize the same user searching on different devices, which becomes increasingly popular these years. For example, Montanez, White, and Huang (2014) estimated that at least 5% of searchers use multiple devices and they conduct about 16% of all searches in their query logs. Users also do not necessarily adopt the same search providers' services on different devices. These issues increase the difficulty to group searches from the same user.

Second, consecutive requests do not necessarily target the same problem. Unlike in a controlled environment, real search engine users do not always focus on a single task during a search period. For example, both self-reported surveys (Spink, Bateman, & Jansen, 1998) and query logs (Spink, Ozmutlu, & Ozmutlu, 2002; Spink, Park, Jansen, & Pedersen, 2006) suggest switching of topics and multi-tasking is common in web search.

Third, sometimes searches related to the same problem are temporally interleaved. Multitasking (Spink, Ozmutlu, & Ozmutlu, 2002; Spink et al., 2006) is a part of the reason. Besides, users may stop searching (e.g., interrupted by other more urgent tasks), but come back to the problem after a relatively long period of time. Teevan, Adar, Jones, and Potts (2007) found in a search log that 38% of repeating queries involved at least one new click (results that users did not click on in their previous requests with identical queries), suggesting these searchers were probably resuming previous tasks to look for new information.

Many techniques can deal with these issues and help extract "ideal" sessions from web search logs. He, Göker, and Harper (2002) studied session boundary criteria for grouping consecutive searches belonging to the same topic. Jones and Klinkner (2008) developed techniques for identifying interleaved queries for the same task. Kotov et al. (2011) and H. Wang et al. (2013) extracted "cross-session search tasks". White and Awadallah (2015) developed personalization techniques for multiple searchers on shared devices.

Whereas results and findings based on an "ideal" session setting, as presented in this dissertation, need to be further examined before being generalized and employed to practical scenarios. This is not only because the process of extracting such ideal session data may be inaccurate in real search engines. A more subtle issue is that various factors (e.g., switching devices, multi-tasking, interrupting and resuming search tasks) may impact user behavior and search process in a practical environment. Yet we do not target to examine these factors, mainly because the limited availability of resource (e.g., unanonymized web search logs).

2.1.2 Models

There are many models in information behavior that provide helpful guidance for studying search systems. But their scopes vary greatly and are often too large to be directly applied to problems within a search session. For example, Wilson's models (Wilson, 1981, 1997) provide guidance on studying contextual factors in search, but these factors are mostly external to and indirectly related to a search session.

Among the many models, the scope of the anomalous state of knowledge (ASK) model (Belkin, 1980; Belkin, Oddy, & Brooks, 1982) is closest to the setting of a search session. More specifically, the ASK model provides the following important suggestions for studying problems in a search session:

First, the ASK model links search to user's problem. Belkin et al. (1982) maintained that search is motivated from user's awareness of insufficient knowledge to solve a problem, e.g., "... user, faced with a problem, recognizes that her/his state of knowledge is inadequate for resolving that problem, and decides that obtaining information about the problem area and its circumstances is an appropriate means towards its resolution." Although not necessarily all searches are for problem-solving, the ASK model suggests a search session should be studied under a larger context to consider the reason that leads to the search session.

Second, the ASK model suggests to evaluate a search system by its effectiveness in solving the user's problem and situational need, e.g., "Whether an anomaly is resolved or not is evaluated in terms of the problem. It should be noticed that the problem determines not only the conceptual requirements of an appropriate response, but also the situational requirements ... " (Belkin et al., 1982). More recently, Belkin, Cole, and Liu (2009) proposed a detailed evaluation framework for this purpose. We follow the ASK model and Belkin et al.'s (2009) framework to collect ephemeral relevance, a contextual judgment regarding the usefulness of a clicked result, or, equivalently, the effectiveness of a click interaction.

Third, the ASK model points out that, due to limited knowledge, a user's query can be ineffective and not descriptive of what the user really wants, e.g., "…it is unrealistic (in general) to ask the user of an IR system to say exactly what it is that she/he needs to know, since it is just the lack of that knowledge which has brought her/him to the system in the first place." (Belkin et al., 1982). Belkin et al. (1982) also suggested systems should "... try to represent them (anomalies) in terms of the user's larger-scale intentions and goals without asking the user to specify the information needed to resolve the anomaly." The prediction of ephemeral relevance can also be considered as a type of representation for the user's intention and goal.

Moreover, the ASK model also suggests that a search process may update the user's anomalous state of knowledge and cognitive mind. This requires the search system to be able to cope with such changes during a search session.

In addition to ASK, several other models also provide guidance to study problems in a search session. Kuhlthau's (1991) information search process (ISP) model breaks down a search process into six stages, with different affective feelings, cognitive states, actions, and goals ("tasks") at different stages. A similar study is Ellis's (1989) behavioral model. Although the scope of an information search process is much larger than that of a search session, the ISP model and Ellis's (1989) behavioral model both suggest that user's affective and cognitive states, search activities, and intentions may undergo change during a search session.

2.1.3 Search Task

Search task is a widely studied influencing factor of user interaction in a search session. Its complex nature and definition is beyond our scope. Here a task follows the information search task discussed by Y. Li and Belkin (2008). They consider (information) search task as motivated by work task that "people perform in order to fulfill their responsibility for their job". We agree with Y. Li and Belkin (2008) on that a search task is motivated by a higher-level problem of the searcher, but we believe the problem is not restricted to job-related work tasks. Instead, our scope of a task is closer to the "problem" discussed in the ASK model (Belkin et al., 1982), where the user's problem makes the user aware of an ASK and is "the driving force of the IR situation." The actual tasks being used and considered in our study mainly stand for a cognitive view of information search. But as Wilson (1981) discussed, search (and information seeking at large) can be used to satisfy not only cognitive needs, but also affective (e.g., search for fun) and even physiological ones. Spink, Jansen, Wolfram, and Saracevic (2002) reported that a substantial proportion of web search queries submitted to the Excite search engine belongs to the categories "Entertainment or recreation" and "Sex and pornography". However, such needs are beyond the scope of our study.

Since we focus on information search and retrieval, in following discussions we uses *search task* and *task* interchangeably.

Tasks are of different types. For example, Marchionini (1989) studied two types of tasks: an open task refers to one that can be satisfied by not only a specific piece of information (an example used in his study is to collect facts about women space travelers), while a closed task looks for a specific fact (e.g., "identify the year in which the speed skating event was introduced into the modern Olympics"). Reid (2000) categorized tasks into internally generated and externally generated ones, where the distinction is whether the task doer or someone else set the task. Similarly, Gross (2001) divided tasks by self-generated or imposed ones. K.-S. Kim and Allen (2002) used known-item search task and subject search task in their studies, where the former looks for a specific and objective fact (e.g., find general requirements of admission to the University of Missouri-Columbia) and the latter is subjective and depends on the searcher's contexts (e.g., "find any information that you think will be useful for getting a job and for planning your future career"). J. Kim (2009) studied factual tasks, interpretive tasks, and exploratory tasks. A factual task only locates facts, while an interpretive task seeks to enhance the searcher's intellectual understanding of a topic. The difference between an interpretive task and an exploratory one is that user has a specific goal in the former but a vague one in the latter. Both Jansen, Booth, and Smith (2009) and Wu et al. (2012) created tasks following a taxonomy of learning objectives (Anderson, Krathwohl, & Bloom, 2001). It includes six levels—remember, understand, apply, analyze, evaluate, and create—each requires an increasing level of cognition and effort than the previous one. However, these work mostly look into task type from one or two dimensions.

Y. Li and Belkin (2008) proposed a faceted classification framework of tasks (both applicable to search task and other tasks). They classify tasks from six facets: source of task (internal generated, external generated, or collaboration); task doer (individual, individual in a group, or group); time (e.g., the frequency of conducting the task, how long the task lasts, and the stage of the task); products (facts, intellectual, image, or mixed product); process (e.g., one-time or multiple-time); goal (e.g., single goal or multiple goal, specific goal or amorphous goal, etc.). Y. Li and Belkin (2008) also summarized the common attributes of tasks, including: objective task complexity, subjective task complexity, difficulty, interdependence (to what degree people other than the task doer are involved), salience, urgency, knowledge of task topic and task procedure. These attributes are mostly subjective in nature, suggesting a strong connection between task and user. J. Liu et al. (2010) later added another task attribute "level" to Y. Li and Belkin's (2008) scheme, which refers to whether the task looks up information at a document or segment level. C. Liu et al. (2012) further added an attribute "naming", where a named task has a specific identified target, and an unnamed one does not.

The definition of a search session, as discussed in § 2.1.1, is mainly observational—it just groups consecutive related searches and activities targeting the same problem. But it requires further clarification on the problem, or say, the task of a search session. Following Y. Li and Belkin's (2008) framework, Table 1 summarizes faceted characteristics of tasks that are reasonable for a search session.

- The source of the task in a search session can be internally generated, externally generated, or collaborative. Either type can lead to individual's consecutive searches on the same problem, which satisfies the definition of a search session here.
- The definition of a search session in this study restricts the task doer to an individual.
- A search session typically deals with short-term tasks. This is because by our definition, a search session involves only consecutive searches for the same problem. It seems unlikely that one can continuously searching for a long time.
- The task conducted in a search session is not restricted to look for any specific types of search product.
- A search session can deal with either a one-time task or a specific execution of a multiple-time task.

Table 1: Faceted characteristics of tasks typically conducted in a search session.

Facets	**Tasks typically conducted in a search session**
Source of task	internally generated, externally generated, or collaboration
Task doer	individual
Time	short-term
Product	factual, intellectual, image, or mixed product
Process	one-time or multiple-time
Goal	specific, amorphous, or mixed goal

- The definition of a search session is not restricted to any types of goal.

Many later studies followed Y. Li and Belkin (2008) and classify search tasks by their classification scheme. For example, the TREC session tracks 2012–2014 (Kanoulas, Carterette, Hall, Clough, & Sanderson, 2012; Carterette, Kanoulas, Hall, Bah, & Clough, 2013; Carterette, Kanoulas, Hall, & Clough, 2014) used four types of tasks that differ in product (factual or intellectual) and goal (specific or amorphous); J. Liu et al. (2010), Cole et al. (2011), Cole et al. (2014), and C. Liu et al. (2014) all used four tasks with different products (mixed or factual), levels (document or segment), goals (specific, mixed, or amorphous), and objective complexity (high or low); C. Liu et al.'s (2012) tasks vary in naming (named or unnamed), product (mixed or factual), level (document or segment), and goal (specific, mixed, or amorphous). We also follow Y. Li and Belkin's (2008) framework when designing experiments. We introduce the detailed choices of tasks in § 3.3.

2.1.4 Contexts

Our study is related to *personalized search* (Pitkow et al., 2002) and *contextual search* (Bennett et al., 2015) at large, where most studies used some context information to improve search systems and/or examined the impact of some context factors on user behavior and search process. We focus on two types of contexts: task attributes and past search history in a session. The setting is similar to some previous work such as X. Shen, Tan, and Zhai (2005), C. Liu et al. (2012), and D. Guan et al. (2013). This is in contrast to studies focusing on other types of context, for example, searcher's long-term interests and activities (Gauch et al., 2003; Teevan et al., 2005), domain expertise (White, Dumais, & Teevan, 2009), reading levels (Collins-Thompson et al., 2011), locations (Bennett et al., 2011), social connections (Carmel et al., 2009), gender (Lorigo et al., 2006), and etc.

Previous work on this topic mainly focuses on two directions. The first thread of work makes use of session search history to improve search result ranking or to assist users. This can be dated back to as early as relevance feedback using explicit user judgments (S. E. Robertson & Spärck Jones, 1976; Salton & Buckley, 1990; Harman, 1992b), although few involved real users at that time. Yet later studies heavily rely on implicit feedback (Kelly & Teevan, 2003). In the context of information retrieval, implicit feedback refers to the user activities that indicate users' preferences of search results regarding their relevance or usefulness. Within a search session, implicit feedback usually includes past query and click (X. Shen et al., 2005), search result dwell time (Kelly & Belkin, 2004; White & Kelly, 2006), query reformulation (D. Guan et al., 2013; Luo, Zhang, & Yang, 2014), and etc. Implicit feedback is preferred over explicit feedback in many situations because the latter requires more effort, and many implicit feedback-based approaches can perform similarly effective (White, Ruthven, & Jose, 2005). TREC also ran evaluation campaigns from 2011 to 2014 (Kanoulas, Carterette, Hall, Clough, & Sanderson, 2011; Kanoulas et al., 2012; Carterette et al., 2013, 2014) dedicated to tasks with a similar setting. A common focus in these studies is to predict or infer result relevance from within-session user activities in order to improve search performance. In addition to search result ranking, within-session context information can also be applied to assist searchers, e.g., providing query suggestions (Cao et al., 2008), predicting browsing interests (White, Bailey, & Chen, 2009), and etc. Besides, many also compared and combined within-session context and long-term historical search activities (Bennett et al., 2012).

Another line of work focuses on the influence of different context factors on user behavior and search process within a session. Search task is a widely studied influencing factor.

Previous studies reported different search behavior in sessions with different types of tasks, levels of complexity, goals, types of search product, etc. (White et al., 2005; White & Kelly, 2006; Kelly & Belkin, 2004; J. Liu & Belkin, 2010; J. Liu et al., 2010; Cole et al., 2011, 2014; Jiang et al., 2014). In addition to task type and attribute, many also reported variation of search behavior in different stages of a search session, for users with different domain knowledge and topic familiarity, etc. (J. Liu & Belkin, 2014, 2015; X. Zhang, Liu, Cole, & Belkin, 2015). The usually examined user behavior includes: click, query, click dwell time, search result page (SERP) dwell time, query reformulation, SERP browsing, result web page browsing, and etc. The two lines of work sometimes overlap with each other. For example: Kelly and Belkin (2004) and White and Kelly (2006) discussed the influence of task and personalization on the effectiveness of using search result dwell time as implicit feedback measures; C. Liu et al. (2012) applied task-specific implicit feedback models.

2.2 RELEVANCE JUDGMENT

Relevance is a key construct of information retrieval and information science (Saracevic, 1996, 2007). An in depth discussion of its complex notion is beyond the scope of our study. Here we summarized five different manifestations of relevance by (Saracevic, 1996, 2007):

- *System relevance* only distinguishes between retrieved results and those that failed to be retrieved.
- *Topical relevance* considers the closeness of a document and a query in topic and content.
- *Cognitive relevance* further considers the relation between a result and a user's knowledge and cognitive state.
- *Situational relevance* considers the relation between a result and "the situation, task, or problem at hand", which is also the focus of our study. We derive our notion of ephemeral relevance from situational relevance. § 3.2.3 discusses ephemeral relevance in detail.
- *Motivational relevance* considers the relationship between a result and "the intents, goals, and motivations of a user."

We focus on discussing the practical measurement of relevance in information retrieval (relevance judgment) in the rest of this section.

2.2.1 The TREC-style Relevance Judgment

Test collection based IR evaluation methods are heavily built upon the Cranfield paradigm and the evaluation practice conducted in TREC evaluations, especially in the TREC ad hoc search tasks (Harman, 1992a). We refer to the relevance judgment approach adopted in the TREC ad hoc search tasks as *the TREC-style relevance judgment*.

The TREC-style relevance judgment uses topic statement as surrogates for a user's information need. A topic statement includes an ID, a title, a description, and a narrative. The title field is usually considered as a search query, although the process of developing a topic statement does not necessarily include any real search. The description field explains the information need in a better detail. The narrative field provides detailed description of what documents are relevant. For example, the following shows the narrative field of TREC topic No. 51. The judgment criteria indicated from the narrative are mainly topical.

> A relevant document will cite or discuss assistance to Airbus Industrie by the French, German, British or Spanish government(s), or will discuss a trade dispute between Airbus or the European governments and a U.S. aircraft producer, most likely Boeing Co. or McDonnell Douglas Corp., or the U.S. government, over federal subsidies to Airbus.

A concern of relevance judgment is on the discrepancies of different judgers in assessments. Voorhees (1998) conducted an experiment where judgers re-assessed old TREC topics, such that their assessments can be compared with the original TREC relevance judgments. Voorhees (1998) reported that the mean overlap of relevant results between two groups relevance judgments is as low as around 40%–50%, and the overlap among three groups of relevance judgments can be as low as just 30%. However, it seems such differences in relevance judgments does not affect the evaluation of IR systems, where the main goal is to generate a ranking of IR systems by performance. In this scenario, the absolute value of the evaluation score is less meaningful. Voorhees (1998) reported that the ranking correlation of IR systems using different sets of relevance judgments are as high as around 80%–90%.

2.2.2 Beyond Topical Relevance

After the TREC ad hoc tracks, lots of effort were made on bringing relevance other than topical relevance into IR evaluation and practice. Yet these effort are not systematic and usually only covers a few specific aspects of relevance other than topical relevance. For example, the TREC web track after 2009 (Clarke et al., 2009) set a few new relevance judgments criteria that are beyond the topical level. They judged whether results are the targeted home pages for a navigational query (queries that explicitly looking for the home page of a website such as "facebook"). This section reviews three efforts of the community in relevance beyond the topic level, including; and the TREC interactive tracks (Over, 2001); the TREC novelty tracks (Harman, 2002; Soboroff & Harman, 2003; Soboroff, 2004); the TREC diversity tracks (Clarke et al., 2009; Clarke, Craswell, & Soboroff, 2010, 2011, 2012, 2013; Collins-Thompson et al., 2014).

The TREC novelty tracks was conducted during 2002–2004 (Harman, 2002; Soboroff & Harman, 2003; Soboroff, 2004). These tracks focuses on solving tasks of the following form: given a query and a list of retrieved results, find *sentences* that are novel to the searcher. The task shares similarity with passage retrieval, and a context is provided (a previous query and results). The form of the task also copes with a fact that search and aggregation of relevant information is usually incremental—after retrieving and seeing a query's results, the user may only want to see new information that are not covered in previous results.

The novelty track task is evaluated by sentence-level F-measure. For each topic, judgers first annotate a set of relevant sentences according to the context query and retrieved results. Then, systems are evaluated by comparing to this groundtruth set of relevant sentences. The first year of this track (Harman, 2002) does not distinguish different topics (tasks). But in the later two years (Soboroff & Harman, 2003; Soboroff, 2004), topics are divided into two types of tasks–events and opinions. The two types of tasks share similarity with Y. Li and Belkin's (2008) faceted classification in task product. In the TREC novelty tracks, although the relevance judgments considered the novelty of results regarding its context, the redundancy and novelty among retrieved results are ignored in these tracks.

In contrast to the novelty tracks, the TREC interactive tracks (Over, 2001) considered a different type of novelty requirement among search results. The TREC interactive tracks (Over, 2001) judged results at a subtopic level, which is referred to as "instances". In evaluation, one of the metrics adopted is "instance recall". This metric measures the coverage of results on different instances. In such case, systems failing to retrieve results covering multiple instances are penalized. This evaluation approach in nature favors systems retrieving a set of results where each result contains unique information. Therefore, the TREC interactive tracks actually considered redundancy and novelty among retrieved results.

Both the novelty tracks and interactive tracks are examples of performing relevance judgments and evaluation at cognitive relevance level, where novelty and redundancy of information is the main focus. However, novelty and redundancy requirements differ along among users. A more salient factor is user's knowledge. Such context, however, is hardly considered in the scope of these tracks.

Compared with the novelty tracks, the TREC diversity tracks (Clarke et al., 2009, 2010, 2011, 2012, 2013; Collins-Thompson et al., 2014) have a related, but different goal. The diversity track track is motivated by the fact that web search queries are usually ambiguous in nature and have subtopics. It is usually difficult to detect the user's real intent to issue the query. Retrieved results (prior to that time) have the risks of biasing towards a certain explanation of the query or a subtopic of the query, while failing to satisfy other possible intents. Therefore, the goal of the diversity track is to rank search results in a way that maximizes the coverage of all possible intents, or subtopics, of that query. The task shares many similarities with the interactive tracks, although the motivation is not exactly the same.

In evaluation, the TREC diversity tracks also adopt metrics that explicitly consider novelty among results (Agrawal, Gollapudi, Halverson, & Ieong, 2009; Clarke et al., 2008). The metrics in the TREC diversity tracks share similarity with instance recall in that both favor systems retrieving results with good coverage of subtopics. However, the metrics used in the diversity tracks also further discount the utility of results if previous results in the list already covered the same subtopic. Therefore, these metrics explicitly require results to be novelty and diverse in nature. Similar ideas are also applied to recommender systems (Vargas & Castells, 2011).

These effort, however, mainly made progress in bringing cognitive relevance into relevance judgments, but those related to situational and motivational are rare. Another related effort is the TREC session tracks (Kanoulas, Carterette, Hall, et al., 2011; Kanoulas et al., 2012; Carterette et al., 2013, 2014). They judged relevance of results as regarding to a task-level topic. However, the relevance judgments criteria is similar to those adopted in the TREC web tracks. Jiang, He, Han, Yue, and Ni (2012) and Jiang, He, and Han (2012) discussed the concerns of the duplicate results in the TREC session tracks.

2.2.3 Relevance Criteria

This dissertation differs from many previous studies in that it focuses on situational relevance in a search session. A major challenge is how to measure situational relevance accurately and study factors related to situational relevance. This relates to previous work on user criteria in relevance judgments. More specifically, previous work as reviewed in this section help this study formulate hypotheses and design experiments accordingly. The main instruments of User Study 1 are also adapted from work (Xu & Chen, 2006) in this line of research.

Xu and Chen (2006) verified users' relevance judgments criteria for problem-solving tasks. Their work was motivated by the fact that previous theoretical work proposed many factors/dimensions for relevance, but few verified on these criteria and measured the effects of different criteria. Thus, they adopted a psychometric approach. They first identified several common aspects of relevance criteria from Grice's theory of communication (Grice, 1991), and then designed multiple-item instruments regarding each criteria. The identified and measured dimensions include: scope; novelty; reliability; topically; understandability. All the five dimensions are hypothesized positively correlated with relevance. Each criteria dimension is measured using 3–5 questions. In this study, the instruments of situational and task-level relevance, and those for changes in user's relevance judgment criteria, are all adapted from Xu and Chen's (2006) instruments.

In the experiment process, they asked subjects to pick one from four search topics that interested them the most. The four topics are "inappropriate intake of vitamins", "outlook of the job market related to the subject's main area of study", "reasons for the .com bubble burst in 2000", and "a three-day trip plan to Tibet". And subjects were allowed to define their own topic if none interested them. The four tasks also covered several different facets in Y. Li and Belkin's (2008) framework. For example, the "a three-day trip plan to Tibet" task looks like a search for informational products with an amorphous goal, while the "inappropriate intake of vitamins" looks like a search looking for factual products with a clear goal. This is also one of the reason that this study adopted Xu and Chen's (2006) instruments. Before judging results, the subjects first searched the topic on Internet The subjects are allowed to select documents for relevance judgments. The relevance judgments are made regarding to the topic as a whole, rather than a specific situation. Results indicate high validity of the instruments and significant effects of the measured factors (except scope) on measured relevance. Y. Zhang et al. (2014) recently argued an issue in Xu and Chen's (2006) experiments—users are allowed to select documents for relevance judgments, which may affect the strength of different factors.

Xu (2007) further applied a similar approach to a non-problem solving context. The tasks are divided into two types, including "epistemic information search", where users are searching to satisfy their desire for knowledge (but not related to a task), and "hedonic information search", where users search for fun or affective stimulation. Both types of tasks are beyond the scope of this study. During relevance judgments, participants were asked to identify a topic they are intrinsically interested in (similar to Y. Li and Belkin's (2008) internally generated tasks). Results, again, suggest high validity of the instruments, and significant effects of the measured factors on measured relevance, although the strength of the criteria's influence on relevance differ in a few places (e.g., novelty seems to be the most salient factor in non-problem solving tasks). Again, among the five criteria, scope has the weakest influence on relevance.

Y. Zhang et al. (2014) recently applied Xu and Chen's (2006) framework and adapted approaches to relevance judgments in crowd-sourcing Y. Zhang et al.'s (2014) work differs from Xu and Chen (2006) in that they only measured the five relevance criteria, and applied confirmatory factor analysis to construct relevance. However, it remains unclear to which extent the five criteria cover all factors of relevance judgments (and what the remaining factors are), which affect the constructed relevance. This also makes it difficult to compare the results with Xu and Chen (2006). The tasks included in relevance judgments are also problem-solving tasks.

Different from Xu and Chen's (2006) approach (where hypothetical criteria are tested and verified), Freund and Berzowska (2010) examined users' relevance judgments criterion in task-centered search from users' comments at the time of relevance judgments. The tasks vary in five types: fact-finding, deciding, doing, learning, and problem-solving. A content analysis approach is applied to analyze these comments. They analyzed 833 comments from 22 participants and identified six relevance criteria codes and subcodes, including: topic (topic coverage, level of detail), situation (audience, location/context), purpose, presentation (design, links/references, images, information findability), quantity (length of amount of information), and quality (currency, credibility, readability, importance/interest).

In addition to the studies on confirming user's relevance judgments criteria, several previous work reported variation of user's relevance judgment criteria in different time of a search process.

Taylor et al. (2007) analyzed users' self-reported relevance judgments along with the time the results were considered in information searching and the reasons for the judgments. They found that in contrast to later stages in an information searching process, at early stages users are more likely to judge results by criteria such as their own interests ("Interest"), understandability ("Ability to understand"), and whether they have learned from the document ("Teaches"). Criteria such as "Specificity", "Recency", and "Depth/Scope" are reported less frequently at early stages. This suggests that at different time of a search session, and situation of a task, users' criteria of relevance judgments may change.

Xu and Wang (2008) studied changes in the process of relevance judgments (order effects (Eisenberg & Barry, 1988)). They hypothesized and tested three effects that may lead to variation of result relevance: a *learning effect* that leads to declined relevance of results over time due to more stringent standards of searchers on topicality and novelty; a *subneed scheduling effect* where premature access to a relevant document may reduce its usefulness; a *cursoriness effect* that searchers overuse their effort at early stages and leave low cognitive capacity for results accessed at later stages. This type of relevance judgments change does not happen during a nature search session, but it also provides clues for hypothesizing reasons that may lead to variation of result relevance during a search session.

Most recently, Shokouhi, White, and Yilmaz (2015) studied anchoring effects in relevance judgments, which suggests that the list of judged results may also affect users' relevance judgments. This effects, however, is ignored in this study due to the difficulty of controlling such factors.

2.3 SEARCH BEHAVIOR AND MODELING

2.3.1 Query Reformulation

Query reformulation is one of the two activities we examined in this study. It is important because it stands for a user's self-motivated move to make progress in a search session. Query reformulation refers to the activity of submitting a query where there was a previous one in the session. Studies of query reformulation mostly focus on the relationship and difference of the two queries involved in a query reformulation.

Anick (2003) summarized some patterns of changes in query reformulation, including: head (adding a linguistic head term to an existing term in previous query), modifier (adding a linguistic modifier), elaboration (adding context words related to the query), location (adding a location), alternative (using an alternative way to express the same meaning), hyponym (replacing a query term with a more specific one), morphological (using a morphological variant), syntactic variant (syntactically rephrasing the query), acronym (using or expanding an acronym), spelling correlation, and change (the new query is about a new topic that is not related to the previous query). Anick's (2003) patterns focus on linguistic variations between the two queries and do not include clear classification criteria. For example, some of the patterns focus on semantic differences, while some others focus on syntactic ones. Anick (2003) also reported that different types of query reformulation patterns vary by popularity. Some patterns (e.g., modifier, elaboration, head, alternative) are more popular than others. In addition, Anick's (2003) patterns heavily rely on human annotation, and it remains unclear how accurately the patterns can be recognized in an automatic manner.

Rieh and Xie (2006) proposed a more principled framework for classifying query reformulation patterns. Their framework includes three facets and several sub-facets for each facet. The first facet is content change, including: specification (using more specific term), generalization (using more general terms), replacement with synonyms, and parallel movement (subtopic change). The second facet is format change, including: term variation (using or expanding acronyms), operator usage (putting an search query operator), error correction. The third facet involves change in resource, including: changing resource type (e.g., requesting a difference source of information, image, news article), domain suffix (adding domain suffix to a term). Compared with Anick's (2003) patterns, Rieh and Xie's (2006) framework makes it possible to connection user's query reformulation patterns to their possible intents in information search. For example, specification may indicate users make progress from a general topic to a specific subtopic (yet it may also indicate the previous query is too general, such that the user tries to correct it by making the query more specific). Similar to Anick's (2003) patterns, Rieh and Xie's (2006) patterns also rely on human labeling, and it remains unclear how accurate these patterns can be recognized automatically.

Huang and Efthimiadis (2009) presented an taxonomy of query reformulation patterns. The patterns include: word reorder, whitespace and punctuation, remove words, add words, URL stripping, stemming, acronyms, substring, abbreviation, word substitution, and spelling correction. An advantage of the taxonomy lies in that the majority of the classes can be identified automatically in a straightforward way (simply based on patterns without the need of training any supervised models). However, this also misses semantics of query reformulation. For example, specification and generalization are not explicitly included into Huang and Efthimiadis's (2009) patterns, although adding and removing words are closely connected with specification and generalization.

These studies summarized patterns of changes in real users' query reformulations, which provide basis for many later studies. Many query suggestion approaches are built on the basis of linguistically changing the content of an existing query. Query suggestions generated using this approach are also referred to as synthetic query suggestions (because the generated queries do not necessarily exist in a search log and thus may not come from real users' queries). In contrast, most other query suggestion approaches are based on statistics of human formulated queries in search logs. For example, X. Wang and Zhai (2008) developed an approach of generating query suggestions using query term addition and query term modification patterns. They used a translation model to select candidate words related to the existing context words of a query (e.g., the left and right 2–3 words). Then, the candidate words are selected based on their connections to the context words. X. Wang and Zhai's (2008) used query log as the corpus for training translation models. Therefore, the selected words are representative of the use of query terms by the search logs' population in the query's linguistic context. Jones, Rey, Madani, and Greiner (2006) also proposed an approach of generate query substitutions from query logs. Their approach is based on words and phrases from existing queries in logs with strong connections with the query. Both word features and word contexts features are used.

One of the tasks studied in this dissertation (prediction of query term addition and removal) shares many similarities with synthesized query suggestion. However, the goals and motivations are completely different, and so are the criteria of evaluation. The purpose of query suggestion is to provide help and support to searchers (e.g., users may not have a good query in mind). For this purpose, the suggested queries are evaluated by their effectiveness in search. Yet the focus of the task in this dissertation is to faithfully predict what users would do in query reformulation. Effectiveness of queries in search is not the goal of evaluation. Instead, the goal of the task explored here even includes: if users are going to add an ineffective query term or remove an important query term (which lead to inefficient queries), such actions still need to be predicted. Comparing to previous work on query suggestion, this task heavily relies on users' task and within-session context—because what users would do in query reformulation are for the purposes of completing the task. Jones and Fain (2003) presented a work of predicting query term deletion, which is motivated by the needs of detecting inefficient queries (e.g., deletion usually happens because previous query are over-specified). However, research in this direction is rare, probably due to the lack of direct motivation to solve this task. But the recent development of simulation-based evaluation techniques makes this task valuable. Some recent work simulated query reformulation (Baskaya, Keskustalo, & Järvelin, 2012) or query modification strategies (Verberne, Sappelli, Järvelin, & Kraaij, 2015), yet none verified the accuracy of the simulation (in terms of fitting user's behavior).

In contrast to previous work, we look into changes in query reformulation at word level—the unit of analysis in our study is a specific word, and whether users will remove, retain, or add the word in query reformulation. This is also relevant to previous work on choices of words in query reformulation and interactive relevance feedback.

Spink and Saracevic (1997) studied five sources of query terms in mediated online searching. Among the sources they examined, question statement is similar to task description in our study, and we also consider relevance feedback as a source for new terms. In addition, the content of search results is also an important source of knowledge for query reformulation (Koenemann & Belkin, 1996). Z. Yue, Han, He, and Jiang (2014) examined possible sources of query words in collaborative search. Some of them may also be applied to other types of searches, including users' past queries and viewed search results. Another source we examined is query suggestions displayed on the SERP. Kelly, Gyllstrom, and Bailey (2009) compared term and query suggestions, where users reported that query suggestions provide ideas for manually formulating queries; Jiang et al. (2014) reported that before query reformulation, searchers viewed task description and query suggestions more frequently.

The word changes we examined in this paper were rarely studied in previous work from the user's perspective. However, some work built technical solutions for contextual search and query suggestion based on these word changes. D. Guan et al. (2013) separately considered added, retained, and removed words in relevance feedback; Dang and Croft (2010) generated synthesized query suggestions by considering similar patterns.

2.3.2 Click

Click is an important search activity. Searchers usually need to click on a search result and read its content in order to acquire relevant information, although this is not always the case. For example, users may directly obtain relevant information from a SERP without clicking on any results, which is usually called "good abandonment" (J. Li, Huffman, & Tokuda, 2009). Another exception is that recently many search engines provide direct answers on a SERP (Chilton & Teevan, 2011; Bernstein, Teevan, Dumais, Liebling, & Horvitz, 2012), e.g., showing stock price for a query "MSFT". However, for (probably) the majority of the cases, users still rely on click to obtain relevant information from a search engine.

Click is an important implicit feedback information (Joachims, 2002; Joachims et al., 2005)—clicking on a search result usually denotes user's (weak) judgment regarding the relevance of the result based on its snippet displayed on a SERP. Although not always correct, such implicit judgments provide valuable resources for search engines. One can extract relevance labels from a click log to train search systems (Joachims, 2002). One can also aggregate click statistics from a search log as features for ranking results (Agichtein, Brill, & Dumais, 2006; Agichtein, Brill, Dumais, & Ragno, 2006).

However, click also suffers from several biases. Joachims (2003) and Joachims et al. (2005) discussed the presentation bias of clicks—user click behavior depends on the presentation of the result on the SERP as well as other results on the SERP. More specifically, this includes several different biases. A widely discussed one is position bias—results ranked at higher positions are more likely to be clicked. Although many verified this phenomenon, the cause of position bias is explained and modeled in different ways. Another types of bias is summary bias (Y. Yue, Patel, & Roehrig, 2010)—results with more attractive summaries get more clicks. Joachims et al. (2005) also discussed trust bias, which refers to the bias that users are more willing to believe top ranked results are worthy of clicking. Yet not all these biases receive an equal amount of attention. For example, the majority of the work focuses on position bias. In addition, Joachims et al. (2005) proposed that click should be explained as preferential feedbacks among results, rather than absolute feedbacks of results' quality.

2.3.2.1 Position Bias

One explanation of position bias is that users' decreased visual attention on results at lower ranks makes them less likely to click on results at the bottom of a SERP. Although the decay of visual attention has been verified consistently by many using eye-tracking devices (Joachims et al., 2005, 2007; Cutrell & Guan, 2007; Z. Guan & Cutrell, 2007), Craswell et al. (2008) argued this cannot explain why click rate drops faster than users' visual attention. As such, Craswell et al. (2008) explains position bias using a cascade model. The meaning of a cascade model today is not only restricted to Craswell et al.'s (2008) model, but a series of models with the trait that a result's click probability depends on click probabilities of results ranked at higher positions. The main idea is that if users clicked on a result at higher ranks and solved their information need, they do not need to get back to the SERP and examine others, which causes less clicks for results ranked lower than the clicked one. Instead, if the top ranked results are not relevant, a relevant result at relatively lower positions may still attract users' attention.

Following these ideas, many work proposed models accordingly to explain observed user clicks. This is important for search engine because, as long as position bias are solved, one can get an unbiased (at least less biased) estimate of click probabilities, which may serve as a better groundtruth for training ranking models. Earlier studies are mostly based on manually created rules. For example, Joachims (2002) applied a rule to generate preferential relevance judgments—clicked results should be more relevant than unclicked ones ranked at higher positions. Later, Joachims et al. (2005), Radlinski and Joachims (2005), and Joachims et al. (2007) generalized rules to other preferential relations on the same SERP, as well as those across multiple SERPs. Joachims et al. (2005) and Joachims et al. (2007) also verified the accuracy of the rules with human preferential relevance judgments (over result summaries). Some rules can be as accurate as over 85%.

Agichtein, Brill, Dumais, and Ragno (2006) further used a richer sets of user interaction features to generate preferential relevance labels. To counterbalance position bias, they subtracted results' click probabilities for a query by a background model—click probabilities of results at the same rank from all other queries. They showed that the differences correlate closely with actual result relevance as explicitly rated by human assessors. The same idea was applied to other features, such as browsing and query features. These features were used to train a RankNet model to rank results and generate preferential relevance judgments. Results show better accuracy of predicting preferential relevance judgments comparing to a baseline approach adapted from Joachims's (2002) strategies. However, the sets of features are also widely used in search result ranking models (Agichtein, Brill, & Dumais, 2006). Thus it remains unclear how reliable such automatically generated preferential judgments can be applied to train ranking models using the same sets of features, or one with large overlap of features.

In contrast, most later work only rely on click to generate relevance labels. These work share one similarity—they all aim at estimating unbiased click probabilities of results from observed biased click evidence. The estimated click probabilities are sometimes referred to as attractiveness, as they intend to measure the condition click probabilities of results provided that the results have been viewed. One of the approach adopted for this purpose is to compare observed click probabilities of results to those of viewing results at the same position. For example, Richardson, Dominowska, and Ragno (2007) applied such an approach. This approach follows the explanation that position bias is caused by decreasing visual attention of users. Later, Craswell et al. (2008) first proposed a cascade model for explaining position bias. Craswell et al.'s (2008) model assumes that: users would sequentially examine results from top to bottom; they click on results by their attractiveness (the unbiased click probabilities); once they click on a result, they can be satisfied and never come back to examine following results. Although some of the assumptions look overly strong (e.g., following this model, users would at most have only one click per SERP), Craswell et al.'s (2008) model is the first to consider dependency between a result's click probability and those at higher ranks. This idea is applied to many later work in click models, as well as IR evaluation metrics (Chapelle, Metlzer, Zhang, & Grinspan, 2009).

Many later work refined the simple cascade model proposed by Craswell et al. (2008):

G. E. Dupret and Piwowarski (2008) proposed a model. This model assumes that the user do not always look at a result summary and click on it based on attractiveness. Instead, it allows a probability that users would decide whether or not to look at a result summary, and click event depends on this viewing decision. In addition, the chances of examine a result summary is assumed to be dependent on both the rank of the result and the distance of the result to the last clicked rank. G. E. Dupret and Piwowarski (2008) discussed an intuition—users are more likely to stop viewing a rank list if they did not see an attractive result for a long time, which is measured by the distance of the result to the last clicked (attractive) result.

Chapelle and Zhang (2009) proposed a dynamic Bayesian network model for clicks. Comparing to Craswell's simple cascade model, it makes the following improvements. First, click depends on both attractiveness and examination, and failure to fulfill either leads to no click. This is also similar to G. E. Dupret and Piwowarski's (2008) model. Second, the model introduces a hidden variable for post-click user satisfaction. It is assumed that users continue to view the SERP and click on other results because they are not satisfied with previous clicks. Therefore, the estimated post-click user satisfaction can also serve as a surrogate for relevance. This takes into consideration users' post-click activities as evidence to infer relevance of results. In addition, examination of a result depends on both examination of the previous result and whether users are satisfied in the previous result. If users are satisfied, they will not examine the next position. Results show strong performance of this model. In addition, the estimated post-click user satisfaction is proved to be useful. Their results showed that, after using both predicted result attractiveness and post-click satisfaction for ranking, nDCG is improved comparing to using only attractiveness for ranking. Guo, Liu, and Wang (2009) and Guo, Liu, Kannan, et al. (2009) also proposed similar models.

The idea of introducing post-click activities also shares similarity to studies of using click dwell time as implicit feedback (Kelly & Belkin, 2004). A later work by Zhong et al. (2010) modeled the post-click satisfaction variable depending on many other feature variables as well, including click dwell time, click dwell time form web pages in the same domain, the interval time between query reformulations, whether user has further clicks on the clicked web page, and whether the user switched to other search engines. Results show slight improvements comparing to Chapelle and Zhang's (2009) approach.

G. Dupret and Liao (2010) proposed a model for estimating "intrinsic document relevance". The idea is also similar to the post-click satisfaction variable in Chapelle and Zhang's (2009) model. G. Dupret and Liao's (2010) model focuses on inferring whether users would stop after clicking and viewing a document and its previous documents. It is assumed that clicked document has certain utility that is additive (which is missing in Chapelle and Zhang's (2009) model), and searchers would stop searching by a probability that depends on the total utility of results being clicked and viewed so far. Therefore, it models the likelihood that searchers, after clicking one or several documents, would stop searching. The model can be used to infer document utility after viewing. Comparing to Chapelle and Zhang's (2009) model and Guo, Liu, Kannan, et al. (2009), a major difference of this model is that the stopping event is modeled based on all previous clicks in the session, rather than only the current clicked document.

Most click models for position bias before 2010 only considered the "10-blue links", which is the main component of a SERP. However, in recent five years, more and more SERP elements other than 10-blue links appear, e.g., vertical results. Many later work studied and modeled the relation between "10-blue links" and other SERP elements. For example: Danescu-Niculescu-Mizil, Broder, Gabrilovich, Josifovski, and Pang (2010) considered the relationship between organic results ("10-blue links") and sponsored results (ads) on clicks; C. Wang et al. (2013) modeled clicks for SERPs providing vertical results. In addition, user variation may cause differences in individual's click pattern, which leaves room for personalized click models (S. Shen, Hu, Chen, & Yang, 2012). Another limitation of most existing click models lies in the sequential examination hypothesis—users look at the SERP from top to bottom and made their decisions. Recently, Chao et al. (2015) relaxed this assumption in modeling clicks.

2.3.2.2 Attractiveness Bias Click models for position bias mostly aim at estimating unbiased click probabilities of results. This probability is considered a cheap surrogate of relevance. The underlying assumption is that click probability of results correlate with results' relevance, which, however, is not always the case. For example, another important bias of click observations is attractive bias—attractive results get more clicks, yet not necessarily more relevant. Similar issues are also discussed by Chapelle and Zhang (2009). However, comparing to the amount of work for position bias, only limited effort were made on summary bias. The core issue of studying summary bias lies on the influence of result summary characteristics on its click probability after viewing.

Whether searchers click on or skip a result summary depends on many factors. Tombros and Sanderson (1998) first applied query-biased summaries to search systems. They found increased accuracy and speed of users in result clicking comparing to systems using query-independent summaries. White, Jose, and Ruthven (2003) came to a similar conclusion. Their studies suggest user click decisions may relate to the relation between summaries and search queries. Cutrell and Guan (2007) further experimented using different lengths of summaries in two types of search tasks. They observed that searchers have best clicking accuracy on short summaries in navigational search, but on long summaries in informational search tasks. This further suggests click and skip behavior may also relate to textual characteristics of summaries and search tasks. Clarke, Agichtein, Dumais, and White (2007) predicted searcher clicks using result summary features. Shokouhi, White, Bennett, and Radlinski (2013) studied how repeated results in different queries were clicked by searchers, which suggests click decision is contextual.

Y. Yue et al. (2010) focuses on result attractiveness as a bias to click behavior. They conducted an online experiment on the Google web search engine with a result swapping setting. The experiments examine the influence of title and abstract query term highlighting on click probabilities Results suggest that clicks are substantially biased towards results with more attractive titles. In comparison, this study does not involve an online experiment with similar setting, but rely on the combination of click and eye-tracking devices to examine such influence (and other factors on click).

Since click is involved in many online experiments as the criteria to evaluate and compare systems, the accuracy of these experiments are affected by summary bias—users prefer to click on results from one system not because they look more relevant, but due to their attractiveness in summary. The two correlate with each other, but also differ from each other. The differences between relevance and click, therefore, leads to inaccuracies in online experiments. In light of this issue, Hofmann, Behr, and Radlinski (2012) used result summary features to correct result summary attractiveness bias in interleaved experiment.

The influence of result summary characteristics is also verified by White and Horvitz (2013) in search related to health topics. They found that users are significantly more likely to examine and click on captions containing potentially alarming medical terminology such as "heart attack" or "medical emergency". Such bias exists in a way independent of position bias. They also adjusted click models according to this bias.

Despite the fact that summary bias leads to many negative effects in web search applications, they are very useful in online advertisement. In pay-per-click mode online advertisement, user clickthrough rates have direct economic incentives. Therefore, many applied models to predict click probability of ads (sponsored results). Search engines can rank ads by predicted click probabilities in order to maximize their revenue. Result summary features play an important role in these applications (Richardson et al., 2007; Srikant, Basu, Wang, & Pregibon, 2010). This is because for many new ads, one may not have enough past click to estimate reliable click models.

In contrast, studies of summary bias—and more generally studies on factors influencing user's click behavior after viewing results—are limited in the domain. This study focuses on this challenge, because this technique may provide a secondary basis for search result ranking. For example, search engine can rank results by not only relevance, but also optimized click sequence.

3.0 METHODOLOGY

This chapter introduces the purpose, design, and methods of this study. § 3.1 describes the framework of the study. § 3.2 defines and clarifies key constructs. § 3.3 discusses the search tasks adopted in the user studies. We introduce the designs of the two user studies in § 3.4 and § 3.5, respectively.

3.1 FRAMEWORK

We study search result relevance and user activities in a search session. More specifically, we focus on ephemeral relevance (ERel)—a contextual measurement of search result usefulness in a session—and two types of user activities—query reformulation and click. Our purpose is both explanatory and practical. First, we examine the influencing factors of ERel and the two types of user activities in a session. Second, we use past search history in a session to predict ERel and future user activities in the same session.

We review the four research questions introduced in Chapter 1:

- **RQ1** – On what criteria do users assess ephemeral relevance (ERel)? How does ephemeral relevance judgment differ from topical relevance judgment and context-independent usefulness judgment?
- **RQ2** – What information other than users' explicit judgment tells the ERel of a search result? How well do these information predict ERel judgments?
- **RQ3** – What affects users' choices of word changes in a query reformulation during a search session? Can we predict such choices of word changes?

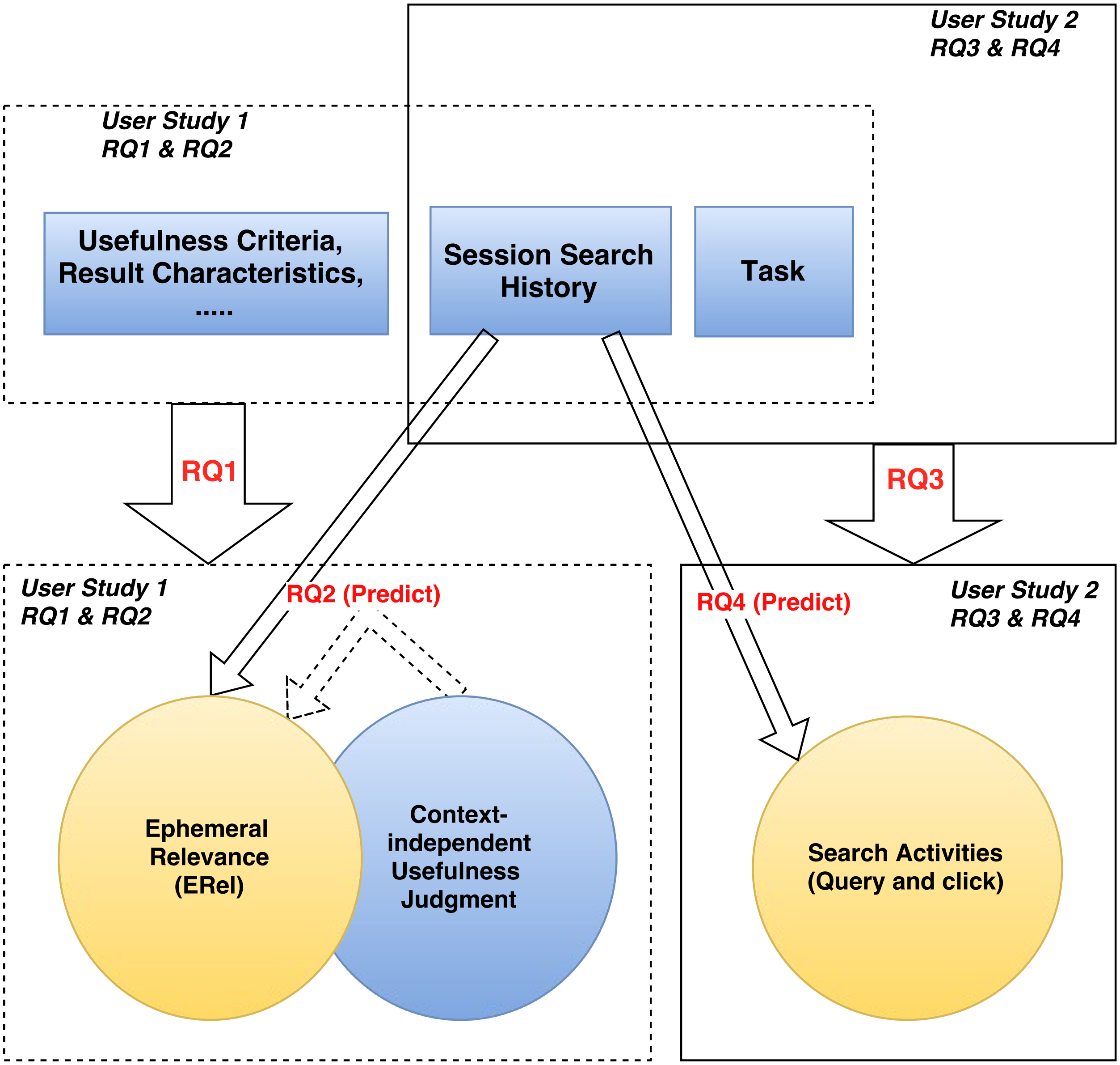

User Study 2
RQ3 & RQ4
User Study 1
RQ1 & RQ2
Usefulness Criteria, Result Characteristics, ...
Session Search History
Task
RQ1
RQ3
User Study 1
RQ1 & RQ2
RQ2 (Predict)
RQ4 (Predict)
User Study 2
RQ3 & RQ4
Ephemeral Relevance (ERel)
Context-independent Usefulness Judgment
Search Activities (Query and click)

Figure 2: A conceptual framework of the study.

- **RQ4** – What affects users' decisions on whether or not to click on a search result's link after viewing its summary displayed on a SERP? Can we predict such decisions?

Figure 2 shows the framework of the study. We examine ERel and user activities (yellow color). We hypothesize different factors (blue color) influence ERel and user activities. RQ1 examines the influence of the factors on ERel. The factors include usefulness judgment criteria, search result characteristics, session search history, and task attributes. RQ3 examines the influence of session search history and task attributes on future search activities in the same session.

We designed two laboratory user studies to examine the influence of the factors on ERel (RQ1) and search activities (RQ3). User Study 1 collected users' search result judgments in contextual and context-independent settings. Study 1 also controlled task goal (whether the goal is specific or amorphous) and product (what type of information the task is looking for) in order to examine the effects of tasks on ERel (introduced in § 3.3). § 3.4 introduces the design of User Study 1.

User Study 2 collected users' search behavior in different types of tasks. User Study 2 also controlled task goal and product. Unlike User Study 1, User Study 2 collected users' eye-movement information using an eye-tracking device in order to examine click decisions (defined in § 3.2.6). The collected data were used to study RQ3 and RQ4. § 3.5 introduces the design of User Study 2.

Another purpose of this study is to develop and evaluate techniques for predicting ERel and future user activities in a search session. The prediction techniques focus on how to leverage implicit feedback signals from past search history to predict ERel (RQ2) and future user activities (RQ4) in a search session.

The user study and corresponding chapter addressing each research question are:

- RQ1 – User Study 1 (§ 3.4) and Chapter 5
- RQ2 – User Study 1 (§ 3.4) and Chapter 6
- RQ3 – User Study 2 (§ 3.5) and Chapter 7
- RQ4 – User Study 2 (§ 3.5) and Chapter 8

Table 2: Differences between search sessions in an ideal case (e.g., collected in a laboratory user study) and those in a practical one (e.g., extracted from search logs).

	Ideal (Laboratory Study)	**Practical (Search Log)**
Task	continuously working on the same task	multi-tasking; interruption
User	the same user	the same "user" identifier
Duration	controlled	algorithmic

3.2 DEFINITIONS

3.2.1 Search Session

In this study, we define *a search session* as a period that involves one or multiple consecutive rounds of searches from a user dedicated to the same problem (task). This definition is ideal. It is representative of the search sessions collected through laboratory user studies (J. Liu et al., 2010; C. Liu et al., 2012, 2014; Cole et al., 2014; Jiang et al., 2014). However, we note that the practical meaning of a search session is usually different in studies of web search and query log analysis. Table 2 summarizes a few major differences:

- First, we restrict our scope to search sessions where a user continuously worked on the same problem. Practically, a user may work on different tasks and switch back and forth during the same period (multi-tasking) (Spink, Ozmutlu, & Ozmutlu, 2002; Spink et al., 2006). In addition, a user may be interrupted during a search session and may or may not resume the search process afterwards.
- Second, user identification may be inaccurate in a practical search engine—user activities in a search session extracted from a search log may not necessarily come from the same user.
- Third, the duration of a search session is usually controlled by researchers in laboratory user studies. Our study has the same issue. In contrast, in studies of search log analysis, researchers usually determine the length of a session in an algorithmic way or based on

heuristics (e.g., a 30-minute cutoff).

Despite these differences, many techniques can address these issues and extract "ideal" sessions from web search logs. § 2.1.1 reviewed related techniques. However, we note that findings based on an "ideal" search session setting need to be further examined before generalizing to practical scenarios.

3.2.2 Topical Relevance (TRel)

Topical relevance is a key construct discussed in this study, mainly for the purpose of comparing with ephemeral relevance (ERel). Saracevic (1996) gave a definition of topical relevance, which focuses on topicality:

> "Topical or subject relevance: relation between the subject or topic expressed in a query, and topic or subject covered by retrieved texts, or more broadly, by texts in the systems file, or even in existence. It is assumed that both queries and texts can be identified as being about a topic or subject. Aboutness is the criterion by which topicality is inferred." (Saracevic, 2007)

In this study, the working definition of topical relevance is the relevance judgment collected using the approach and criteria of the TREC 2014 web track (Collins-Thompson et al., 2014). The TREC approach is the de facto standard of relevance judgment in the IR community due to its large influence. The 2014 web track is TREC's latest evaluation of a general-purpose ad hoc search task. The TREC approach asks assessors to judge search result relevance without an actual search context. The TREC relevance judgment criteria focus on topicality.

We collect topical relevance judgments (TRel) in User Study 1 after each search session using a question adapted from the TREC 2014 web track's criteria. § 3.4 introduces details.

3.2.3 Ephemeral Relevance (ERel)

Without loss of generality, we define the *ephemeral relevance* (ERel) of an information object (such as a search result or a piece of text) as the amount of useful information a user would acquire from the object by interacting with it under a natural condition at a particular time of a search process. More specifically, its working definition in this study, for the purpose of measurement, is the perceived amount of useful information a user acquired from a clicked result right after the user finished examining the search result during a search session.

We consider ERel as a "snapshot" of the *situational relevance* by Saracevic:

> "Situational relevance or utility: Relation between the situation, task, or problem at hand, and information objects (retrieved or in the systems file, or even in existence). Usefulness in decision making, appropriateness of information in resolution of a problem, reduction of uncertainty, and the like are criteria by which situational relevance is inferred. This may be extended to involve general social and cultural factors as well." (Saracevic, 2007)

ERel can also be considered as a particular implementation of Belkin et al.'s (2009) evaluation model, which assesses an IR system at three levels: "1. the usefulness of the entire information seeking episode with respect to accomplishment of the leading task; 2. the usefulness of each interaction with respect to its contribution to the accomplishment of the leading task; 3. the usefulness of system support toward the goal(s) of each interaction, and of each ISS" (information seeking strategy). The second level of the model measures the usefulness of "each interaction". We can also consider ERel as a specific case of the second level where we restrict the interaction to a click.

ERel has the following characteristics:

- Usefulness to the problem (task) is the primary criterion for assessing ERel. *Relevance* and *usefulness* are interchangeable in the context of ERel.
- ERel depends on the result, the user, the search task, and the time of a search process (a search session). Not only ERel but also the states of the user and the search task are time-dependent. Users may have both cognitive and affective changes during a search process (Kuhlthau, 1991), and the search task at hand may also evolve during a search session. Thus, ERel captures not only the dynamics of search result relevance, but also the states of the user and the search task in a search session.
- ERel measures the effectiveness of the interaction for acquiring relevant/useful information from a result. It is not an attribute of the result. For example, users do not necessarily read the whole result document in order to obtain all the information. They

may skip reading a result if it costs too much effort to locate or understand the information they need. Thus, ERel measures not only how much useful information the result contains, but also how much the users are willing to acquire from the result in the particular search context.

- As its name denotes, ERel slips away soon and cannot be restored, because both the states of the user and the problem at hand are changing. ERel needs to be assessed by the user just-in-time in a search process.

We call ERel a *contextual* relevance/usefulness judgment measure because users assess ERel in a genuine search context. ERel takes into account factors such as the status and background of the user, the search task, the time of a session, previously viewed results, the easiness to understand the content of the result, etc. In contrast, we use *context-independent* relevance judgments to refer to relevance judgments that do not involve a search context (such as TRel). ERel has many theoretical advantages, which may make it a more accurate measure than static relevance judgments.

3.2.4 Context-independent Usefulness Judgment (Usef)

§ 3.2.2 and § 3.2.3 introduced topical relevance (TRel) and ephemeral relevance (ERel) judgment. They differ from two aspects: judgment criteria and context. First, TRel uses topicality as the primary criterion but ERel focuses on usefulness. Second, TRel judgment does not involve context but ERel judgment does.

To separately examine the influence of judgment criteria and context, we collect a third type of judgment, context-independent usefulness judgment (Usef). We collect Usef judgment in a context-independent setting, after users finished their search sessions. But we use a judgment question similar to ERel and focus on usefulness as the judgment criterion. Usef is a user's judgment regarding the usefulness of a search result to the task after the search session.

Table 3 summarizes the difference between TRel, ERel, and Usef. § 3.4 introduces details of the experiment for collecting these judgments.

Table 3: Differences between topical relevance (TRel), ephemeral relevance (ERel), and context-independent usefulness judgments (Usef).

	Judgment Criteria	**Judgment Setting**
ERel	Usefulness	Contextual
TRel	Topicality	Context-independent
Usef	Usefulness	Context-independent

3.2.5 Query Reformulation

In a search session, *query reformulation* refers to the activity of formulating a new search query when there was an old one. Table 42 shows an example search session with four queries. The user reformulated three times in the session: from the first query to the second, from the second to the third, and from the third to the fourth.

When analyzing a query reformulation, we focus on the difference between the old query and the new one. We assume that the difference indicates the user's decisions in query reformulation, such as whether to remove or to retain a word, and whether or not to add a word to the new query. Chapter 7 analyzes these word changes in details.

3.2.6 Click Decision

In addition to query reformulation, we also examine users' click behavior in a search session. We focus on *click decision*, which is defined here as a user's decision on whether or not to click on a search result's link after viewing its summary displayed on a search result page (SERP). Users need to make such click decisions when they browse results on a SERP. The observed clicks in search logs are the results of the click decisions users made during a search session. In User Study 2, we collect users' eye-movement behavior using an eye-tracking device to examine click decisions.

Table 4: An example of query reformulations and word changes in a search session.

	Query	Word Changes
1	depression symptoms	-
2	depression definition	retain: depression remove: symptoms add: definition
3	depression treatment	retain: depression remove: definition add: treatment
4	depression treatment cost	retain: depression retain: treatment add: cost

3.3 SEARCH TASKS

In § 2.1.3, we reviewed the characteristics of search tasks typically conducted in a search session based on Y. Li and Belkin's (2008) faceted task classification scheme. Table 5 further summarizes the characteristics of the tasks considered in our user studies and compares with the possible options in a search session.

- **Source of task** – in order to control task product and goal, our user studies did not allow participants to generate their own tasks. All search tasks were externally generated and assigned to the participants.
- **Time** – all participants performed the tasks during a short duration, typically lasting about 10 minutes.
- **Product** – our user studies only considered tasks targeting either factual or intellectual products. We did not consider tasks targeting mixed types of products.
- **Goal** – our user studies only considered tasks with either a specific or amorphous goal.

Table 5: Faceted characteristics for tasks typically conducted in a search session and those considered in this study.

Facets	**Tasks typically conducted in a search session**	**Tasks considered in this study**
Source of task	internally generated, externally generated, or collaboration	externally generated
Task doer	individual	individual
Time	short-term	short-term
Product	factual, intellectual, image, or mixed product	factual or intellectual
Process	one-time or multiple-time	one-time
Goal	specific, amorphous, or mixed goal	specific or amorphous

Table 6: Task type in our study and the corresponding name in TREC session tracks.

Task Type in Our Study	**Corresponding Name in TREC**
Factual+Specific (F+S)	Known Item
Factual+Amorphous (F+A)	Known Subject
Intellectual+Specific (I+S)	Interpretive
Intellectual+Amorphous (I+A)	Exploratory

We did not consider tasks with mixed goals.

Our user studies controlled task goal and targeted product. The goal of a search task is either *specific* (well-defined and fully developed) or *amorphous* (ill-defined or unclear goals that may evolve along with the user's exploration). The targeted product of a task is either *factual* (to locate facts) or *intellectual* (to enhance the user's understanding of a problem).

The tasks adopted in our user studies come from the 2012 and 2013 TREC session tracks (Kanoulas et al., 2012; Carterette et al., 2013). The TREC session tracks named the tasks as: *known item* (factual search with a specific goal), *known subject* (factual search with an amorphous goal), *interpretive* (intellectual search with a specific goal), and *exploratory* (intellectual search with an amorphous goal). The naming shares similarities with J. Kim's (2009) work, but we feel the names of *known subject* and *interpretive* look less intuitive. We use the naming convention in Table 6 in this study. Appendix A lists all the tasks and task descriptions.

3.4 USER STUDY 1: EPHEMERAL RELEVANCE

3.4.1 Purpose

User Study 1 is a laboratory user study. We use User Study 1 to examine the two research questions related to ephemeral relevance (RQ1 and RQ2). User Study 1 asked participants to work on preassigned tasks of different types and collected their perceptions of the search results and the search sessions.

Our main purpose is to collect ephemeral relevance (ERel), topical relevance (TRel), and context-independent usefulness (Usef) judgments. In addition, we also collected other information that may help explain ERel and the differences between ERel, TRel, and Usef (RQ1), including: the criteria of judging ERel and Usef, the effort spent on the clicked results, and users' judgments regarding some properties of the clicked results (such as novelty, reliability, understandability, and specificity). In order to construct prediction models for ERel using implicit feedback (RQ2), we also recorded users' search behavior such as search queries and clicks in a search session.

3.4.2 Experiment Conditions

User Study 1 controlled the target product (either *factual* or *intellectual*) and goal (either *specific* or *amorphous*) of the assigned search tasks. This included $2 \times 2 = 4$ combinations in total. The experiment used a 2×2 within-subject design. We rotated task sequence using a reduced Latin square, where it requires a group of four participants to cover all conditions.

We divided participants into groups of four. We assigned the participants within each group exactly the same four tasks, with different task sequence. We assigned different groups of participants to work on different four tasks. The purpose was to reduce the concern of task parity in IR experimental design (Kelly et al., 2015). We used the search tasks developed by the TREC session tracks (Carterette et al., 2016). Although we can control task types, there is no guarantee that the tasks are comparable from other aspects (such as task complexity and difficulty). Assigning different tasks to different groups of participants reduces the influence of other factors on experiment results.

User Study 1 employed 28 participants and assigned 7 groups of tasks to them (28 unique tasks in total). The tasks come from the TREC 2012 and 2013 session tracks (Kanoulas et al., 2012; Carterette et al., 2013). Table 71 (Appendix A) includes the TREC topic numbers for the tasks assigned to each group.

3.4.3 Task Workflow

We assigned each participant to work on four search tasks. Each participant worked on each task for about 20 minutes following the workflow in Figure 3.

3.4.3.1 Pre-task Survey

In Stage A, the participants read the task description and completed a pre-task survey for their familiarity with the topic of the search task (topic familiarity) and four other questions (not examined here). We measured topic familiarity using the following question **Fami**. Figure 10 (Appendix C) shows a screenshot of the system for this stage.

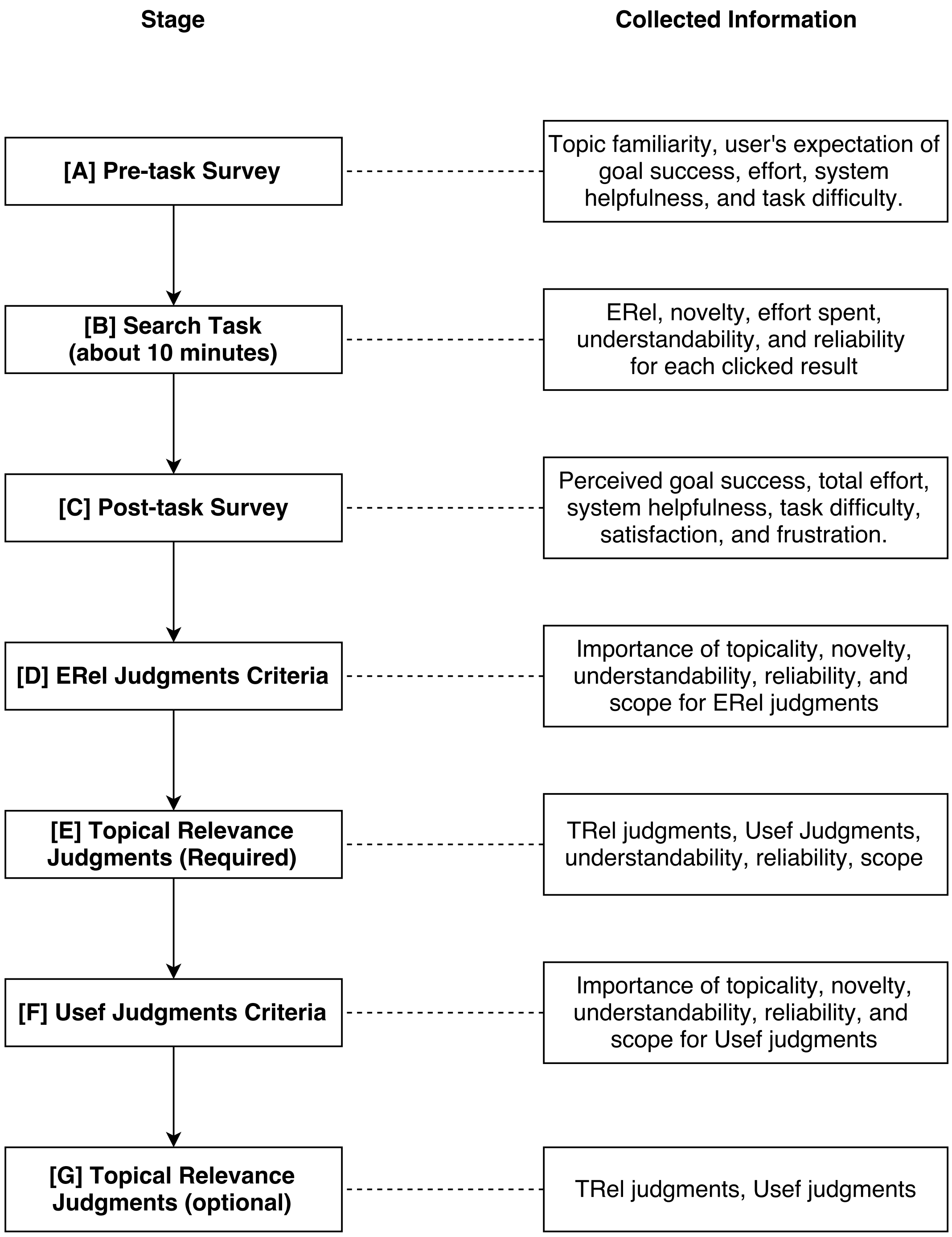

Stage
Collected Information
[A] Pre-task Survey
Topic familiarity, user's expectation of goal success, effort, system helpfulness, and task difficulty.
[B] Search Task (about 10 minutes)
ERel, novelty, effort spent, understandability, and reliability for each clicked result
[C] Post-task Survey
Perceived goal success, total effort, system helpfulness, task difficulty, satisfaction, and frustration.
[D] ERel Judgments Criteria
Importance of topicality, novelty, understandability, reliability, and scope for ERel judgments
[E] Topical Relevance Judgments (Required)
TRel judgments, Usef Judgments, understandability, reliability, scope
[F] Usef Judgments Criteria
Importance of topicality, novelty, understandability, reliability, and scope for Usef judgments
[G] Topical Relevance Judgments (optional)
TRel judgments, Usef judgments

Figure 3: User Study 1: Ephemeral Relevance – task workflow.

Table 7: User Study 1: Ephemeral Relevance – questions being asked during contextual judgments.

[**ERel**] How much useful information did you get from this web page?

none						a lot of
1	2	3	4	5	6	7

[**Novel**] How much new information did you get from this web page?

none						a lot of
1	2	3	4	5	6	7

[**Effort**] How much effort did you spend on this web page?

none						a lot of
1	2	3	4	5	6	7

[**Relia**] How trustworthy is the information in this web page?

not at all trustworthy						very trustworthy
1	2	3	4	5	6	7

[**Under**] How difficult was it for you to follow the content of this web page?

very difficult						very easy
1	2	3	4	5	6	7

Table 8: User Study 1: Ephemeral Relevance – questions being asked in the post-task survey.

[**Diff**] How difficult was this task?

very easy						very difficult
1	2	3	4	5	6	7

[**Succ**] How well did you fulfill the goal of this task?

very badly						very well
1	2	3	4	5	6	7

[**SEffort**] How much effort did this task take?

minimum						a lot of
1	2	3	4	5	6	7

[**Help**] How well did the system help you in this task?

very badly						very well
1	2	3	4	5	6	7

[**Sat**] How satisfied was your search experience?

very unsatisfied						very satisfied
1	2	3	4	5	6	7

[**Frus**] How frustrated were you with this task?

not frustrated						very frustrated
1	2	3	4	5	6	7

Table 9: User Study 1: Ephemeral Relevance – questions being asked during post-session judgments.

[**TRel**] How relevant is this web page (please select one from the following)?

- **Key**: this page or site is dedicated to the topic; authoritative and comprehensive, it is worthy of being a top result in a web search engine.
- **Highly Relevant**: the content of this page provides substantial information on the topic.
- **Relevant**: the content of this page provides some information on the topic, which may be minimal.
- **Not Relevant**

[**Usef**] How much useful information does this web page provide for the task?

none						a lot of
1	2	3	4	5	6	7

[**Spec-Q**] This web page is specifically related to my query "XXX"* rather than the task in general.

strongly disagree						strongly agree
1	2	3	4	5	6	7

* Here XXX was automatically replaced with the query that retrieved the result.

[**Spec-S**] This web page is specifically related to a sub-problem rather than the task as a whole.

strongly disagree						strongly agree
1	2	3	4	5	6	7

[**Fami**] How familiar are you with the topic of this task?

very unfamiliar						very familiar
1	2	3	4	5	6	7

3.4.3.2 Search and Contextual Judgments In Stage B, participants worked on the assigned search task for about 10 minutes using an experimental search system (introduced in Section 3.4.4). They were instructed that they could issue any search queries and check any web pages, but they were not allowed to use other search engines. Figure 11 (Appendix C) shows a screenshot of the search interface.

The participants were instructed to try their best to work on the search task (e.g., find out answers to all the questions in the task description, or explore the stated problem as much as possible). They were told that the experiment coordinator would notify them to stop when the time was up, but they were also told that they could ask the coordinator to finish this stage if they felt they could perfectly answer all questions or solve the problem stated in the task description. In 19 out of the 112 collected sessions, searchers asked the coordinator to terminate Stage B by themselves.

The participants were asked to perform contextual judgments for every search result they clicked on. After they clicked on a search result's link, the system froze and popped up a contextual judgment survey. When participants switched back from the result web page to the search system, they needed to finish the contextual judgment questions in order to resume the search session. Figure 12 (Appendix C) shows a screenshot of the contextual judgment interface, where a contextual judgment survey popped up and the search system in the background froze.

The contextual judgments for a click result included five questions as in Table 7.

- The **ERel** question measures the ephemeral relevance of the clicked result. The wording of the question intentionally put an emphasis on how much useful information users *got from the web page*, rather than how much *the web page provides*. This takes into account that searchers do not always fully examine a result document. For example, they may abandon a web page if it costs too much effort to locate relevance information. The **ERel** question intended to cover such issues. This is also a hypothetical difference between contextual (ERel) and context-independent usefulness judgments (Usef).

- The **Novel** question measures the novelty of the clicked result in terms of how much new information users acquired from the clicked web page.
- The **Effort** question measures the amount of effort users spent on the clicked result.
- The **Relia** question measures the reliability of the information in the clicked result.
- The **Under** question measures the understandability of the clicked result in terms of the easiness to understand the content of the clicked result.

Participants were instructed to spend normal effort to examine the search result web pages as they would when using search engines in a regular condition. In addition, they were specifically instructed that they did not need to go back to the clicked result to double check its content in order to answer the contextual judgment questions. When they answered the contextual judgment questions, the system also did not provide them with a link to the clicked result (to discourage switching back to the clicked result). This is because in a pilot study, we observed that some participants switched back and forth between the clicked result and the contextual judgment survey to verify the content of the clicked web page. Some participants in the pilot study also reported that they assumed they were expected to check the web pages in detail such that they could answer the questions correctly. We made changes to the instructions and settings as mentioned above in the formal study to address these issues.

3.4.3.3 Post-task Survey In Stage C, participants finished a post-task survey measuring their search experience in the session. Figure 13 (Appendix C) shows a screenshot of this stage. Table 8 shows the questions of the post-task survey. The survey included six questions for measuring users' perceptions of task difficulty (**Diff**), goal success (**Succ**), session effort (**SEffort**), the helpfulness of the system to the task (**Help**), user satisfaction (**Sat**), and frustration (**Frus**). These measured constructs are widely used for measuring user experience in web search engines (Jiang, Hassan Awadallah, Shi, & White, 2015; Jiang, Hassan Awadallah, Jones, et al., 2015).

3.4.3.4 ERel Judgment Criteria In Stage D, participants weighed the importance of five factors when they answered the **ERel** question (Table 7) during the task search session. The five factors are topicality, novelty, reliability, understandability, and scope. Previous work (Xu & Chen, 2006; Y. Zhang et al., 2014) verified that the five factors are important for relevance judgments. User Study 1 measured the importance of these five factors for ERel judgments.

For each factor, we constructed three statements confirming or disconfirming a result satisfies the factor. We showed the participants these statements and asked them to rate the importance of the statement when they answered the **ERel** question. The following table shows an example. **NOV1** is a statement that the result web page is novel. We asked participants to rate the importance of **NOV1** for their ERel judgment question using a 7-point Likert scale from *not at all important* (1) to *very important* (7).

> [**Instruction**] You may still remember—each time you visited a web page and switched back to our system, we asked you the question "*How much useful information did you get from this web page?*" Please weigh the importance of the following factors when you answered this question.
>
> [**NOV1**] The web page provides new information to me.

not at all important						very important
1	2	3	4	5	6	7

In total, 15 statements were constructed for these five factors. The 15 statements were adapted from Xu and Chen's (2006) survey for relevance judgments criteria. Participates rated the importance of all 15 statements. The 15 statements were displayed in a random sequence. Appendix B includes the 15 statements. The system also showed participants an instruction and a screenshot for the ERel judgments they accomplished to clarify and help them recall the ERel judgment question. Figure 14 (Appendix C) shows a screenshot of the system at this stage.

3.4.3.5 Required Context-independent Judgments In Stage E, participants were asked to perform context-independent judgments regarding all the clicked results in a session.

The clicked results were displayed to the searchers in a random order. For each clicked result, participants answered six questions as showed in Table 9:

- The **TRel** question measured the topical relevance of the result using the standard adopted in the TREC web tracks (Clarke et al., 2009; Collins-Thompson et al., 2014). The relevance levels (from the highest to the lowest) and the corresponding numerical values are: *key* (3), *highly relevant* (2), *relevant* (1), and *not relevant* (0). The TREC web tracks also included a level *Nav* – the correct homepage for a navigational search query (e.g., "facebook"). Here User Study 1 did not use this level in **TRel** because the adopted search tasks did not include navigational search tasks.
- The **Usef** question measured the usefulness of the result to the task in general. It is considered as the major topical relevance measurement in this study. **Usef** differs from **TRel** in that it used a similar wording and scale as **ERel** did, which reduces the differences caused by inconsistencies in wording (such as those between TRel and ERel). Both **ERel** and **Usef** used the wording "how much useful information ...", but differed in that **ERel** put an emphasis on "...did you get from this web page", while **Usef** asked "...does this web page provide for the task". In addition, both **ERel** and **Usef** used a 7-point scale from *none* (1) to *a lot of* (7).
- The **Spec-Q** question intended to measure the specificity of the result to the search query retrieved the result compared with the task in general. The **Spec-S** question intended to measure the specificity of the result to a sub-problem of the task compared with the task in general. However, we do not examine responses to these two questions in this study because many participants reported confusions regarding the two questions.
- The **Relia.ps** and **Under.ps** (not included in Table 9) measured the reliability and understandability of the result web page. The two questions are exactly the same as **Relia** and **Under** used in ERel judgments. These questions intended to measure whether or not participants' perceptions regarding the reliability and understandability changed in the session.

Participants were specifically instructed to re-examine the clicked web pages in a better detail and answer the questions. The system also forced them to revisit these web pages and examine for at least 30 seconds before they could answer the questions. Figure 15 (Appendix C) shows a screenshot for the system in this stage.

3.4.3.6 Topical Relevance Judgments Criteria In Stage F, participants were asked to weigh the importance of the five criteria when they answered the question **Usef** in topical relevance judgments (Stage E). Stage F used the same five criteria and 15 statements as Stage D did for ERel judgments criteria. Figure 16 (Appendix C) shows a screenshot of the system for this stage.

3.4.3.7 Optional Topical Relevance Judgments In Stage G, participants were asked to assess topical relevance for the retrieved but unvisited results in the search session. This stage is optional. Only participants who completed all the previous stages in less than 20 minutes were requested to work on this stage. However, different from Stage F (required topical relevance judgments), participants were only asked to answer **TRel** and **Usef** for each result. Figure 15 (Appendix C) shows a screenshot for this stage.

3.4.4 Experimental Search System

In User Study 1, participants worked on the search tasks using an experimental search system. The system redirected user queries to Google and returned modified Google search results to the users. The system only showed the ordinary "10-blue links" and query suggestions (if any) to users. Other SERP elements were removed, either because that are subsidiary (e.g., ads) or due to the lack of consensus in current search engines on how to display these elements on SERPs (e.g., vertical results, related entities). The experimental system displayed results in the same way they would appear on Google—font size, weight, color, and other formatting were maintained. The major difference between the experimental system and Google in SERP design was that the system showed task descriptions on the top of the SERP. This was to help searchers remember the requirements of their tasks.

3.4.5 Experiment Procedure

The experiment for a participant took between 100 to 120 minutes. Each participant worked on four search tasks. The expected time for each task was 20 minutes, but the actual time spent was sometimes longer to ensure that the participants could finish all the required stages (Stage A to Stage F).

The general procedure of the experiment is as follows. First, participants were introduced the purpose of the experiment and its risks. They were requested to sign a consent form and answer an intro-survey for their background information. Then, they were introduced to work on a training task (Appendix A), where they went through the general workflow of a search task and the questions. The training session took about 10 to 15 minutes, where participants were requested to answer all the questions from Stage A to Stage F to ensure they understood the workflow. They were encouraged to ask the coordinator for any confusions they had on the process and the questions. Third, participants worked on the four formal tasks in a regular setting. They were required to take a 5-minute break after finishing two tasks.

Participants were recruited through fliers posted to the campuses of two universities in the United States. Participants were restricted to English native speakers. Their participation was voluntary. They were reimbursed for $15 per hour.

In total, User Study 1 recruited 35 participants. 5 participants were included in a pilot study for testing the experiment workflow and the clarity of the research questions. 2 participants quited during the experiment. The rest 28 participants completed the whole experiments. Their responses were used for the analysis.

3.5 USER STUDY 2: SEARCH BEHAVIOR

3.5.1 Purpose

User Study 2 collected participants' search behavior in sessions of different types of tasks to answer RQ3 and RQ4. User Study 1 cannot provide *natural* search behavior data because we interrupted the participants for contextual judgments. In addition, User Study 2 also collected users' eye-movement data using an eye-tracking device. RQ4 relies on these eye-movement data to examine users' click decisions.

3.5.2 Experiment Conditions

Similar to User Study 1, User Study 2 controlled task product (two levels) and goal (two levels). User Study 2 used a 2 × 2 within-subject design, where each participant worked on all four tasks (conditions). We rotated task sequence using a reduced Latin square, where a group of four subjects can cover all rotations. Similar to User Study 1, participants within each group worked on the same four tasks, and we assigned different tasks to different groups of participants to increase task diversity. Table 72 (Appendix A) shows the TREC topic numbers of tasks assigned to each group.

3.5.3 Task Workflow

In each task, the participants finished the following three stages. We only use data from the first stage (Search) here. We used data collected in User Study 2 in an earlier study (Jiang et al., 2014), which examined research questions different from this study.

3.5.3.1 Search In the search stage, participants used the experimental system to find information in order to solve the task. We instructed them to use the experimental search system as if they were using public search engines such as Google and Bing. We instructed them that they could search using any textual queries, browse search result pages, click on and view results, and use query suggestions, etc. However, we instructed them that they were not allowed to use other search engines. We set a time limit of 10 minutes to each task. After 10 minutes, the system showed a highlighted link notifying the participants to terminate the search stage. However, we also instructed the participants that they could finish the task before 10 minutes if they believed they had already learned enough to solve the task perfectly.

3.5.3.2 Report In the report stage, the participants rated their familiarity with the topic of the task before they worked on the task, the difficulty of the task, and their search performance using a 5-point Likert scale. Then, we asked participants to write a paragraph reporting their outcomes of the search session. During this stage, the system showed a countdown of 5 minutes to help the participants to finish in about that time. However, the system would not freeze after 5 minutes. We instructed participants to make full use of the time instead of finishing as soon as possible during the report stage.

Among the information collected in the report stage, we only use participants' responses to their familiarity with the task before they worked on the task. We asked the participants the following question:

How familiar were you with the topic before you worked on this task?

not at all familiar				**very familiar**
1	2	3	4	5

3.5.3.3 Relevance Judgment In this stage, we asked the participants to judge 25 selected search results regarding their relevance to the search task. We selected the 25 results by the following priority: visited (clicked) results $\succ$ unvisited results ranked higher than at least one clicked result $\succ$ other retrieved results. We also asked an external annotator to judge other retrieved results using the same criteria. We asked participants the following question:

How relevant is this web page to the task?

- **Highly relevant** (2): the web page provides useful information and is dedicated to the topic.
- **Relevant** (1): the web page provides some useful information, but only a part of the web page is related to the task.
- **Not relevant** (0).

3.5.4 Experimental System

User study 2 used an experimental search system providing modified Google search results. First, the system removed results other than the "10-blue links" and query suggestions. Second, the system only showed nine results each page (rather than the usual ten) to make sure that users do not need to scroll down to see all of the result items. This change made it much simpler to analyze the eye-tracking data. However, Joachims et al. (2005) also reported that scrolling down affects browsing patterns on results displayed below the screen cutoff of a SERP. User Study 2 did not consider such issue. Third, if Google provided query suggestions ("related searches") for a query (usually showed at the bottom of a SERP), the system moved the query suggestions to the right side of the 10-blue links, again, to eliminate scrolling pages.

The system looked very similar to Google except a few places specifically designed for this experiment. It showed the task description at the top of a SERP. This is because a pilot study of User Study 2 found that, without showing the task description, users might constantly switch between SERPs and the page showing the task description, because it was difficult for them to remember all details of a task. We believe this would cause greater issues to the collected data (e.g., more constantly switching of web pages) than showing task description on the SERP. In addition, the system showed a highlighted "finish task" button when the session exceeded the time limit (but not before the limit was reached).

3.5.5 Apparatus

User Study 2 used a Tobii x50 eye-tracker to collect eye movement data. Among the various types of eye movement observations, we only analyze fixation (stably gazing at an area of the screen). Previous studies found that fixation on an area of the screen usually indicates that users were reading information displayed on the area of interest (AOI) (Rayner, 1998). The AOIs in this study include each search result summary, query suggestion, and task description. It is assumed here that fixation on these AOIs indicates that the participant viewed the corresponding result abstract, query suggestion, and task description. We used ClearView, a software accompanying the eye-tracker, to analyze fixations on the defined AOIs. We set the minimum duration of fixation to 100ms, a common value adopted in many previous studies of web search behaviors using the same series of eye-tracker (Cutrell & Guan, 2007; Z. Guan & Cutrell, 2007). In following chapters, we report "user examined a result summary" if we observed any fixations on the AOI of the result.

3.5.6 Experiment Procedure

The total experiment duration for a participant was about 2 hours. We reimbursed the participants by the rate $15 per hour. At the beginning of the experiment, we introduced participants to the system and a training task (with all the three stages but shorter time limits). Then the participants worked on four formal tasks. After two formal tasks, they took a 10-minute break. We interviewed the participants at the end of the experiment.

We recruited the participants through fliers posted to the campus of the University of Pittsburgh. The flier required the participants 1) to be English native speakers and students, and 2) to have a perfect eye-sight without glasses or lens. All participants claimed they met the criteria.

4.0 DATA

This chapter summarizes overview statistics of the data collected in User Study 1 (§ 4.2) and User Study 2 (§ 4.3) and other datasets § 4.4. § 4.1 also introduces some technical details of processing the datasets.

4.1 DATA PROCESSING

4.1.1 Text Processing

Some analysis conducted in the study are based on textual data, e.g., counting the number of words included in a user query, measuring the text similarity between a query and a web page, etc. Unless particularly noted, this dissertation processed textual data (such as user queries and the contents of web pages) using the following settings:

- Lucene's `StandardTokenizer` (Version 5.4) was applied to tokenize texts;
- Words were normalized using the Krovetz stemmer (Krovetz, 1993);
- Letter case differences were ignored;
- No stop words were removed;
- If the text source is an HTML document, HTML tags and auxiliary components (such as CSS and JavaScript) were removed.

4.1.2 Click Dwell Time Estimation

Click dwell time is a widely used measure in web search. As the name denotes, click dwell time intends to measure how long a user has stayed on a clicked search result. It is widely used as an indicator of search result relevance and for other purposes.

The actual click dwell time estimated by a search engine is usually inaccurate due to the limited ways of logging user activities. The most common approach used in current web search engines is called "server-side" dwell time (Y. Kim, Hassan, White, & Zitouni, 2014a). The server-side dwell time of a clicked result is simply measured as the duration from the time of clicking on the result to the next activity of the same user observed by the search engine. Since most search engines only record searchers' query and click activities, the next observed user activity is either another search result click or a new SERP request (either the submission of a new query or a request to return a different page of results for the current query).

Server-side click dwell time estimation has a few limitations:

1. It does not exclude the duration from the searcher got back to a SERP and the next observed activity.
2. It is difficult to determine the dwell time of the last click of a session because there is no "next" activity.
3. Users may have clicked on new links in a clicked search result. Therefore, even excluding the two issues discussed above, the dwell time estimation does not necessarily represent the dwell time spent on the clicked search result only.

Despite these issues, server-side click dwell time is widely used in current search engines and was proved useful for many purposes in previous studies (White et al., 2005; White & Kelly, 2006; Agichtein, Brill, Dumais, & Ragno, 2006; Agichtein, Brill, & Dumais, 2006).

User Study 1 had logged the timestamps that the users started and finished the in situ judgments. This makes it possible to address the first and the second issue of server-side click dwell time estimation.

Let t_{click} denote the time stamp of clicking on a result and t_{j1} be the time that the users started the in situ judgments of the clicked result. User Study 1 estimated the dwell time of a clicked result as the duration from the time of clicking on the result to the start of the in situ judgments (Equation 4.1). This estimation is expected to be more accurate than regular server-side click dwell time estimation, as it excludes the time of viewing the SERP (although this cannot solve the third limitation). The reported results for User Study 1 are based on the click dwell time estimation in Equation 4.1 unless particularly noted.

$$t_{\text{dwell}} = t_{\text{j1}} - t_{\text{click}} \tag{4.1}$$

Some of the analysis related to User Study 1 also reported results based on server-side click dwell time estimation for comparison. In such a case, the estimation had excluded the time spent on the in situ judgments. Let t_{next} denote the time stamp of the next observed activity by the server and t_{j2} be the time of finishing the in situ judgments of the clicked result. The server-side click dwell time in User Study 1 is estimated as Equation 4.2, where $t_{\text{next}} - t_{\text{click}}$ is the regular estimation of server-side click dwell time and $t_{\text{j2}} - t_{\text{j1}}$ is the time spent on the in situ judgments.

$$t_{\text{dwell}} = t_{\text{next}} - t_{\text{click}} - (t_{\text{j2}} - t_{\text{j1}}) \tag{4.2}$$

User Study 2 did not include in situ judgments. Therefore, all reported results related to User Study 2 are based on server-side click dwell time ($t_{\text{next}} - t_{\text{click}}$).

4.2 DATA COLLECTED IN USER STUDY 1: EPHEMERAL RELEVANCE

4.2.1 Background of the Participants

User Study 1 recruited participants through fliers posted to two universities in the United States—Simmons College (Boston, MA) and the University of Massachusetts Amherst[1]. In total, User Study 1 recruited 33 participants. Five of them participated in a pilot study for testing the workflow of the experiment and the clarity of the questions. 28 subjects finished the formal study (21 from Simmons College and the rest from the University of Massachusetts Amherst). User Study 1 required the participants to be English native speakers to exclude the influence of language proficiency on search result judgments (Hansen & Karlgren, 2005). The participants were reimbursed for $15 per hour.

[1] This was only due to that the author lived in Massachusetts while completing User Study 1.

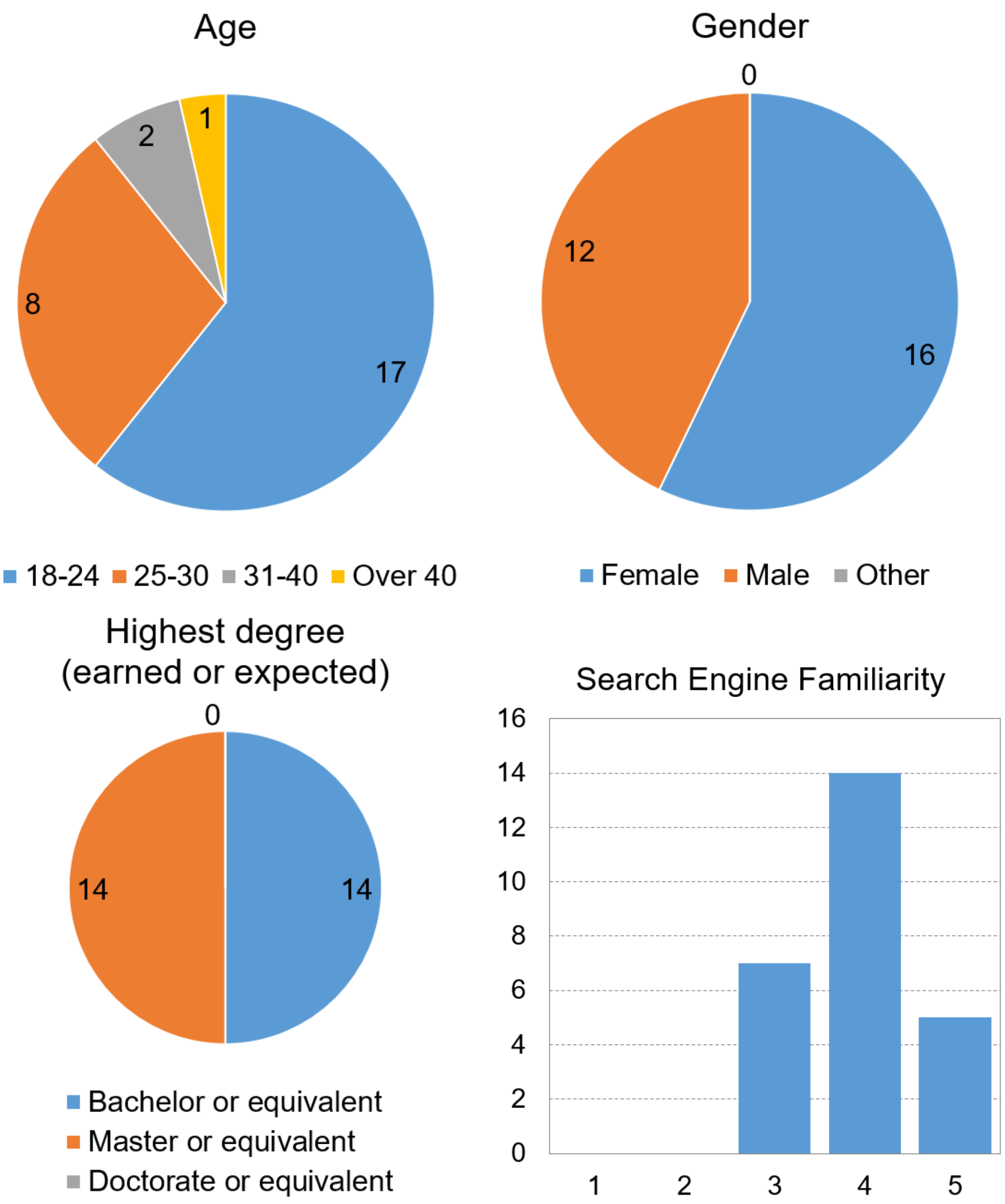


Figure 4: 28 formal participants' responses of the background survey in User Study 1: Ephemeral Relevance.

Table 10: 28 formal participants' responses of their areas of study in User Study 1.

Children's Literature, library science	English and Literary Arts
Library and Information Science	Library and Information Science
Library Science	Accounting
Art History	MLIS, concentration in school libraries
Management	Children's Literature
Library and Information Science	Library and Information Science
Library Science and Info Management	Library and Information Science
Library Sciences (Archives)	English, Computer Science
Marketing/English	Writing, Literature and Publishing
MSW/MBA	English
Nursing	nursing
Computer Science & Health Informatics	Neuroscience & Behavior
Psychology	Library and Information Science
Public Health, Biology Track	Biology

Participants of the formal study answered a background survey before the experiment. The background survey included five questions:

- Gender (multiple choice) – *Male*, *Female*, or *Other*.
- Age (multiple choice) – *18–24*, *25–30*, *31–40*, or *over 40*.
- Highest degree earned or expected (multiple choice) – *Bachelor or equivalent*, *Master or equivalent*, or *Doctorate or equivalent*.
- Area of study (short text question).
- Proficiency of using web search engines – the participants answered using a 5-point Likert scale from 1 (*very badly*) to 5 (*very well*). The question was *How well do you rate your proficiency of using web search engines (e.g., Google, Bing, Yahoo, etc.)?*

Figure 4 and Table 10 summarize the responses. The participants are representative of college and graduate students of 18–30 year old. All of them rated their proficiency of using web search engines as at least neutral (3). According to the responses in Table 10, about half of them were studying information-related disciplines such as Library & Information Science and Computer Science[2]. The others' background covered a diverse range of areas.

[2] This was probably due to that the user study was conducted at the School of Library and Information

Table 11: Overall statistics of the data collected in User Study 1: Ephemeral Relevance.

Measure	**Value**
# sessions	112
# unique queries per session	4.03
# viewed SERPs per session	5.12
# clicks per session	6.57
# unique visited results per session	6.49
Average query length	4.35
Average # unique words in a query	4.32
Average click dwell time	57.08 s
Average click dwell time (server-side estimation)	71.23 s
Average time spent on in situ judgments	12.10 s

4.2.2 Overall Statistics

User Study 1 collected 112 search sessions from 28 unique participants on 28 different tasks (seven tasks for each combination of task product and goal). Each participant worked on four distinct tasks and each task was performed by four different subjects. Table 11 summarizes some overall statistics of the data collected in User Study 1. On average, the participants spent 12.1 seconds to complete the in situ judgments (including all five questions).

Table 12 further reports the participants' ratings of topic familiarity and search experience in different types of sessions, with the effects of task product and goal tested using a two-way repeated measures ANOVA analysis. The results suggest no significant differences of each task type regarding users' familiarity with the task topic and their search experience, except that the interaction of task product and goal shows a significant effect on users' perceptions of system helpfulness at 0.05 level ($p = 0.03$). Table 12 confirms that our particular selections of tasks in the four conditions are comparable regarding attributes other than task product and goal (such as topic familiarity and task difficulty).

Science at the Simmons College and the School of Computer Science at the University of Massachusetts Amherst.

Table 12: Users' ratings of topic familiarity and search experience in each task condition (Product×Goal) of User Study 1: Ephemeral Relevance.

	Group Mean ($N = 28$)				Effects		
Measures	**F+C**	**I+C**	**F+A**	**I+A**	**Product (F or I)**	**Goal (C or A)**	**Interaction**
Topic Familiarity	2.18	2.07	2.21	1.79	$p = 0.35$	$p = 0.67$	$p = 0.60$
Task Difficulty	3.32	3.68	3.82	3.68	$p = 0.68$	$p = 0.48$	$p = 0.46$
Satisfaction	5.61	5.39	5.07	5.36	$p = 0.89$	$p = 0.22$	$p = 0.27$
Frustration	1.86	2.00	2.32	2.25	$p = 0.87$	$p = 0.13$	$p = 0.68$
Goal Success	5.68	5.64	5.25	5.36	$p = 0.88$	$p = 0.15$	$p = 0.75$
System Helpfulness	6.04	5.64	5.25	5.79	$p = 0.78$	$p = 0.11$	$*$
Session Effort	3.50	3.75	3.61	4.04	$p = 0.22$	$p = 0.51$	$p = 0.75$

Task product is either *Factual* (F) or *Intellectual* (I).
Task goal is either *Clear* (C) or *Amorphous* (A).
Effects of product and goal are tested using two-way ANOVA.
*, **, and *** indicate $p < 0.05$, 0.01, and 0.001, respectively.

4.3 DATA COLLECTED IN USER STUDY 2: SEARCH BEHAVIOR

4.3.1 Background of the Participants

User Study 2 recruited participants through fliers posted to the campus of the University of Pittsburgh[3]. The fliers required the participants 1) to be English native speakers, 2) to be college or graduate students, and 3) to have a perfect eye-sight (20/25) without glasses or lens. All participants claimed they met the criteria. In total, User Study 2 recruited 26 participants, including six participated in a pilot study for testing the workflow of the experiment and the eye-tracking device. 20 subjects finished the formal study. The participants were reimbursed for $15 per hour.

Participants of the formal study answered a background survey before the experiment. The background survey asked the same questions as those being asked in User Study 1. Figure 5 and Table 13 summarize the responses. The participants are representative of

[3] This was only due to that the author lived in Pittsburgh while completing User Study 2.

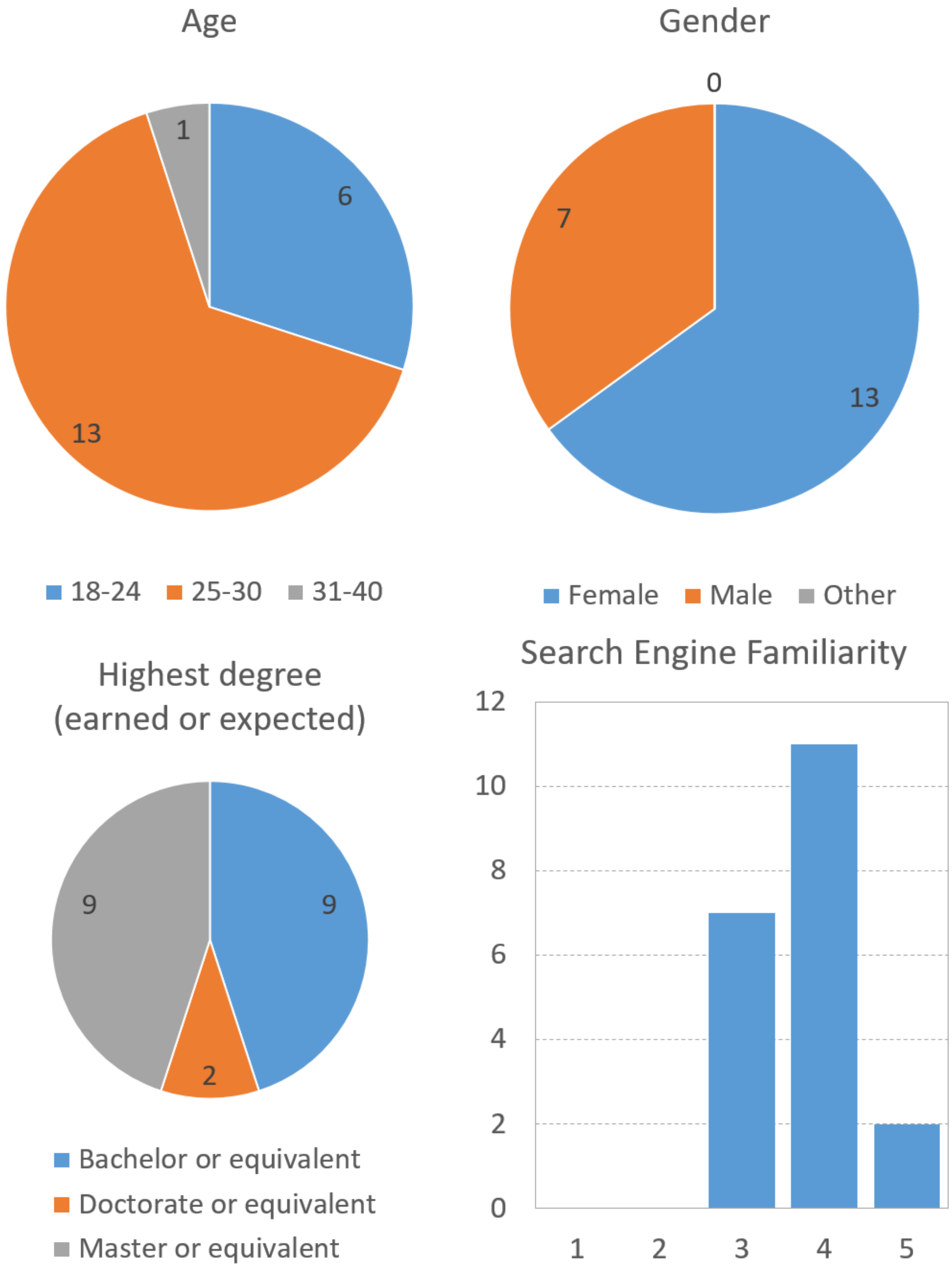


Figure 5: 20 formal participants' responses of the background survey in User Study 2: Search Behavior.

Table 13: 20 formal participants' responses of their areas of study in User Study 2: Search Behavior.

library and information science	Library and Information Science
Epidemiology	MLIS
Infectious Diseases and Microbiology	Business
Creative Writing	English
Library and Information Technology	Library Science
Spanish & Portuguese Language	Database and Web Systems
Nursing	Psychology
anthropology	Library and Information Science
Social Work	Mechanical Engineering
Psychology	Library & Information Science

college and graduate students of 18–30 year old. All of them rated their proficiency of using web search engines as at least neutral (3). According to the responses in Table 13, about half of them were studying information-related disciplines such as Library & Information Science and Database and Web Systems[4]. The others' educational background covered a diverse range of areas.

4.3.2 Overall Statistics

User Study 2 collected 80 search sessions from 20 unique participants on 20 different tasks (five tasks for each combination of task product and goal). Each participant worked on four distinct tasks and each task was performed by four different subjects. Table 14 summarizes some overall statistics of the data collected in User Study 2. It is worth noting that the participants in User Study 2 clicked on more results and spent shorter time on each clicked results compared with those in User Study 1. This suggests that the in situ judgments did have certain influence on users' search behavior, because the two user studies are mostly similar in settings except that User Study 1 included in situ judgments after each click.

Table 15 further reports the participants' ratings of topic familiarity and search experi-

[4] This was probably due to that the user study was conducted at the School of Information Sciences at the University of Pittsburgh.

Table 14: Overall statistics of the data collected in User Study 2: Search Behavior.

Measure	**Value**
# sessions	80
# unique queries per session	4.11
# viewed SERPs per session	4.85
# clicks per session	10.48
# unique visited results per session	9.30
Average query length	4.02
Average # unique words in a query	3.99
Average click dwell time (server-side estimation)	53.49 s

Table 15: Users' responses of search experience in each task condition (Product×Goal) of User Study 2: Search Behavior.

	Group Mean (N = 20)				**Effects**		
Measures	**F+C**	**I+C**	**F+A**	**I+A**	**Product (F or I)**	**Goal (C or A)**	**Interaction**
Topic Familiarity	1.85	1.85	2.10	2.00	$p = 0.87$	$p = 0.55$	$p = 0.83$
Task Difficulty	2.55	2.30	2.45	2.60	$p = 0.84$	$p = 0.74$	$p = 0.37$
User Performance	3.80	3.80	3.90	3.50	$p = 0.41$	$p = 0.65$	$p = 0.40$

Task product is either *Factual* (F) or *Intellectual* (I).
Task goal is either *Clear* (C) or *Amorphous* (A).
Effects of product and goal are tested using two-way ANOVA.
*, **, and *** indicate $p < 0.05$, 0.01, and 0.001, respectively.

ence in different types of sessions, with the effects of task product and goal examined using a two-way repeated measures ANOVA analysis. Similar to those reported for User Study 1, the results suggest no significant differences of each task type regarding users' familiarity with the task topic and their search experience. Table 15 also confirms that our particular selections of tasks in the four conditions are comparable regarding attributes other than task product and goal.

4.3.3 Verification of Eye-tracking Data

Eye-tracking devices cannot guarantee perfect accuracy. Study 2 required all participants to have a perfect eye-sight (20/25) without glasses or lens. Although this is not required, the manual of the eye-tracker noted that imperfect eyesight may reduce the accuracy of the collected data. All participants claimed that they met the criteria (but the experiment coordinator did not verify whether this was true). Before experiments, all participants were requested to go through an initial adjustment, as instructed by the manual of the eye-tracker. Three subjects who failed to pass this stage were excluded from the study. All reported participants passed the initial adjustment.

In addition, the recorded data was verified by calculating the percentage of clicked results with an observed eye fixation. This is based on the assumption that users should have viewed the summary of a result before clicking on it. Under this assumption, this percentage measures the recall of user's attention on the set of clicked results. The same method was adopted by Joachims et al. (2005). In the collected data, this percentage is as high as 87%, similar to the values reported in previous studies (Joachims et al., 2005). Whereas it is difficult to verify the precision of the collected eye fixations.

4.4 OTHER DATASETS AND RESOURCES

Linguistic statistics of words were computed using the ClueWeb09[5] corpus (the English subset). ClueWeb09 is a standard web search corpus including 503,903,810 English language web pages. This study used ClueWeb09 to compute word statistics (such as IDF).

[5] http://www.lemurproject.org/clueweb09.php/

5.0 UNDERSTANDING EPHEMERAL RELEVANCE

This chapter studies ephemeral relevance (ERel) and its differences compared with topical relevance (TRel) and context-independent usefulness judgments (Usef) based on the participants' responses collected in User Study 1. This chapter proposes seven hypotheses related to RQ1:

- **RQ1** – On what criteria do users assess ephemeral relevance (ERel)? How does ephemeral relevance judgment differ from topical relevance judgment and context-independent usefulness judgment?
- **H1.1** – Searchers may have different criteria for contextual (ERel) and context-independent usefulness judgments (Usef).
- **H1.2** – ERel is related to users' perceptions on the topicality, novelty, understandability, and reliability of the search result at the time of assessing ERel.
- **H1.3** – ERel is related to the effort spent on the result.
- **H1.4** – ERel is related to the search task.
- **H1.5** – ERel differs from TRel and Usef.
- **H1.6** – The difference between ERel and Usef judgments is influenced by the difference of users' perceptions on the results in the two judgments.
- **H1.7** – ERel correlates with user experience better than TRel and Usef judgments.

We examine these hypotheses and answer RQ1 in following sections: § 5.1 is dedicated to H1.1; § 5.2 examines H1.2–H1.5 by looking into the relationships between ERel and each factor individually; § 5.2 further studies H1.2–H1.5 by looking into the relationships between ERel and other factors altogether; H1.6 is examined in each section by discussing the influence of task product and goal on the findings; § 5.6 draws a conclusion to RQ1.

5.1 USEFULNESS CRITERIA

Many previous studies examined the criteria of relevance judgment. For example, Xu and Chen (2006) hypothesized that relevance judgment relates to five factors—the topicality, novelty, understandability, reliability, and scope of a search result. They verified based on a laboratory user study that relevance judgment is significantly linked to the former four factors but not scope. In contrast to the abundance of knowledge about relevance criteria, we know little about on which basis searchers assess the usefulness of a search result. Also, it remains unclear whether or not the usefulness criteria vary in contextual and contextual-independent settings.

Following Xu and Chen's (2006) study, we hypothesis usefulness judgment also relates to the five factors. User Study 1 explicitly collected participants' perceptions of the importance of the five factors in contextual and context-independent usefulness judgments. As introduced in § 3.4.3.4, we measured the importance of each factor using three items—topicality (TOP1, TOP2, and TOP3), novelty (NOV1, NOV2, and NOV3), understandability (UND1, UND2, and UND3), reliability (REL1, REL2, and REL3), and scope (SCP1, SCP2, and SCP3). Users answered each question by a 7-point Likert scale, where the responses range from 1 (not at all important) to 7 (very important). We separately measure the importance of each factor for the two different judgment settings (contextual and context-independent). Appendix B includes the questions. Figure 14 and 16 (Appendix C) shows the interface for answering these questions. In total, we collected 28 participants' responses in 112 search sessions.

Our main hypothesis is that usefulness judgment relates to the five factors by different extents. In addition, we suspect that searchers may weigh these factors differently in the two judgment settings, which may contribute to the difference of contextual and context-independent usefulness judgments. We are also interested in whether or not search task product and goal have any effects on the importance of the factors in usefulness judgments.

This section starts with an analysis of the internal consistency of the responses in § 5.1.1. Further, we particularly seek answers to the following questions:

- **RQ1.1** – How much do users weigh the five factors in contextual and context-independent

Table 16: Internal consistency of users' responses to the importance of each factor in contextual and context-independent usefulness judgment settings ($N = 112$). The importance of each factor in each setting was measured using three questions (items).

	Cronbach's α	
Factors	**Contextual**	**Context-independent**
Topicality (TOP)	0.768	0.817
Novelty (NOV)	0.912	0.964
Understandability (UND)	0.932	0.953
Reliability (REL)	0.856	0.907
Scope (SCP)	0.828	0.903

usefulness judgments? (§ 5.1.2)

- **RQ1.2** – Do users weigh the five factors differently in contextual and context-independent usefulness judgments? (§ 5.1.3)
- **RQ1.3** – Do task product and goal affect the importance ratings of the factors? (§ 5.1.4)

5.1.1 Internal Consistency

We study usefulness judgment criteria using a multi-item questionnaire, where the importance of each factor in each judgment setting was measured using three different questions. This section examines the internal consistency of the responses using Cronbach's α. Table 16 reports the results.

Users' responses to the three questions regarding the importance of novelty, understandability, reliability, and scope are highly consistent ($\alpha > 0.8$) in both contextual and context-independent usefulness judgments. Their answers to the three questions for the importance of topicality (TOP1, TOP2, and TOP3) also have acceptable internal consistencies ($\alpha = 0.7$) in the two judgment settings. This suggests that the design of the questions for measuring the importance of novelty, understandability, reliability, and scope are generally successful, while there seems still room to improve the questions related to topicality. We encourage future studies develop more consistent measurements for the importance of topicality.

To sum up, participants' responses to the multi-item questionnaire are mostly consistent. In the following sections, we measure the importance of each factor in each judgment setting using the average rating of the three items (questions). We use TOP, NOV, UND, REL, and SCP for the importance of topicality, novelty, understandability, reliability, and scope, respectively. For example, $\text{TOP} = (\text{TOP1} + \text{TOP2} + \text{TOP3})/3$.

5.1.2 Importance of Different Factors

Participants' responses agree that all the five factors are generally important for usefulness judgments, but their extents of importance vary greatly. In both contextual and context-independent judgments, participants' ratings suggest the following preferential ranking of the factors regarding their importance in usefulness judgment:

$$\textbf{topicality, reliability} \succ \textbf{understandability} \succ \textbf{novelty, scope}$$

Figure 6 and Figure 7 show the box plots of the importance ratings of the five factors in contextual and context-independent usefulness judgments. As the figures show, despite a few exceptions, the majority of the importance ratings for the five factors exceed the threshold of neutral (4 in a 7-point Likert scale). This suggests that participants generally believe all the five factors are important for their usefulness judgments, regardless of in a contextual (contextual) or context-independent (context-independent) setting.

Table 17 and Table 18 further report the mean values of the importance ratings, with the differences of the ratings among the five factors examined using one-way ANOVA and Tukey's HSD post-hoc test. One-way ANOVA analysis suggest that the importance ratings of the five factors are significantly different regardless of in contextual ($F(4, 111) = 53.23$, $p < 0.001$) or context-independent ($F(4, 111) = 57.66$, $p < 0.001$) usefulness judgment settings. Tukey's HSD post-hoc tests further disclose the following differences among the five factors regarding their importance.

- Users rated topicality (TOP) and reliability (REL) as the most important two factors in both contextual and context-independent usefulness judgments. TOP and REL received significantly higher importance ratings than other three factors in both judgment settings

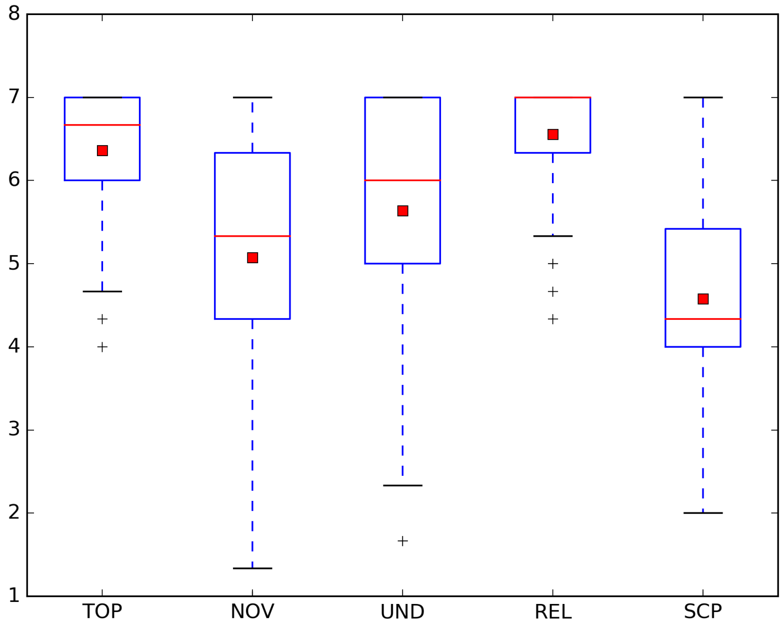


Figure 6: A box plot of the importance ratings of the five factors in contextual usefulness judgments ($N = 112$).

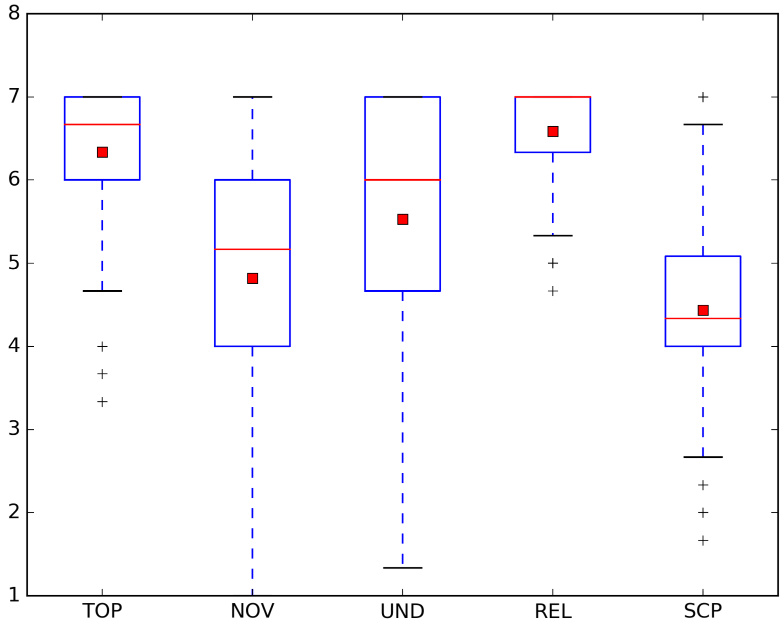


Figure 7: A box plot of the importance ratings of the five factors in context-independent usefulness judgments ($N = 112$).

Table 17: Mean values of the importance ratings of the five factors in contextual judgments ($N = 112$).

Factor	**Mean**	**Significant differences (by Tukey's HSD)**
TOP	6.36	>NOV ***, >UND ***, >SCP ***
NOV	5.07	<TOP ***, <UND **, <REL ***, >SCP *
UND	5.64	<TOP ***, >NOV **, <REL ***, >SCP ***
REL	6.56	>NOV ***, >UND ***, >SCP ***
SCP	4.58	<TOP ***, <NOV *, <UND ***, <REL ***

One-way repeated measure ANOVA suggests the importance ratings of the five factors are significantly different at 0.001 level; $F(4, 111) = 53.23$, $p < 0.001$.

Table 18: Mean values of the importance ratings of the five factors in context-independent judgments ($N = 112$).

Factor	**Mean**	**Significant differences (by Tukey's HSD)**
TOP	6.34	>NOV ***, >UND ***, >SCP ***
NOV	4.82	<TOP ***, <UND ***, <REL ***
UND	5.53	<TOP ***, >NOV ***, <REL ***, >SCP ***
REL	6.59	>NOV ***, >UND ***, >SCP ***
SCP	4.44	<TOP ***, <UND ***, <REL ***

One-way repeated measure ANOVA suggests the importance ratings of the five factors are significantly different at 0.001 level; $F(4, 111) = 57.66$, $p < 0.001$.

(the differences are significant at $p < 0.001$). But the ratings for TOP and REL do not differ significantly, suggesting they are comparably important factors in usefulness judgments.

- Participants rated novelty (NOV) and scope (SCP) as the least important two factors among the five for their usefulness judgments. The importance ratings of NOV and SCP are significantly lower than the other three factors at $p < 0.001$ in both contextual and context-independent usefulness judgment settings. We also observed a significant difference between NOV and SCP at 0.05 level in contextual judgments, but we found none in the context-independent setting. Comparing to the clear differences of NOV and SCP with the other three factors, the difference between NOV and SCP seems limited. It requires further investigation to determine whether users weigh NOV and SCP substantially different.
- In both judgment settings, participants rated understandability (UND) as significantly less important than TOP and REL ($p < 0.001$) but significantly more important than NOV and SCP ($p < 0.001$).

Summing up the results, participants' responses suggest all the five factors are generally important for usefulness judgments, but their extents of importance vary clearly—topicality (TOP) and reliability (REL) are the most important two, followed by understandability (UND), and novelty (NOV) and scope (SCP) are the least important among the five.

5.1.3 Contextual vs. Context-independent Judgments

Participants' importance ratings in the two usefulness judgment settings (contextual and context-independent) are mostly consistent, except that they rated novelty (NOV) as slightly more important in contextual judgments than in a context-independent setting (a significant difference at 0.05 level was observed).

Table 19 compares the importance ratings of each factor in contextual and context-independent judgments. We test the differences of ratings in two judgment settings using Wilcoxon signed-rank test due to the skewed distribution of the responses. We also report paired t-test for reference, which are highly consistent with those using signed-rank test.

Table 19: Mean importance ratings of the five factors: contextual vs. context-independent judgments.

Factor	**Contextual**	**Context-independent**	**Wilcoxon**		**Paired *t*-test**	
TOP	6.36	6.34	$p = 0.885$		$p = 0.748$	
NOV	5.07	4.82	$p = 0.028$	*	$p = 0.010$	*
UND	5.64	5.53	$p = 0.055$		$p = 0.090$	
REL	6.56	6.59	$p = 0.399$		$p = 0.432$	
SCP	4.58	4.44	$p = 0.071$		$p = 0.062$	

The results show that users rated novelty (NOV) as slightly and significantly more important in contextual judgments than in context-independent ones (5.07 vs. 4.82, $p = 0.028$ by Wilcoxon test). Despite being statistically significant, the actual difference (0.25) seems very small in a 7-point Likert scale. For the other four factors, their differences of importance ratings in the two judgment settings are not statistically significant at 0.05 level, regardless of using Wilcoxon test or paired *t*-test.

5.1.4 Task Influence

Our results suggest that task product and goal have some influence on the importance of topicality (TOP) and reliability (REL) in context-independent usefulness judgments. However, neither task product nor goal seems to affect the importance of any factors in an contextual judgment setting.

Table 20 reports the two-way ANOVA analysis of task product (*factual* or *intellectual*) and goal (*clear* or *amorphous*) on the importance ratings of the five factors in contextual usefulness judgments. According to the table, neither task product, goal, nor their interaction show any significant effects on the importance ratings of the five factors in contextual judgments.

Table 21 further reports the results for context-independent usefulness judgments. Task product shows a significant effect on the importance of reliability (REL) at 0.05 level—

Table 20: Importance ratings of the five criteria in contextual judgments under different task conditions (Product×Goal).

	Group Mean (N = 28)				**Effects**				
Factor	**F+C**	**I+C**	**F+A**	**I+A**	**Product (F or I)**		**Goal (C or A)**		**Interaction**
TOP	6.45	6.33	6.39	6.26	$p = 0.28$		$p = 0.45$		$p = 0.95$
NOV	4.82	4.57	4.57	5.11	$p = 0.55$		$p = 0.43$		$p = 0.08$
UND	5.46	5.43	5.50	5.36	$p = 0.52$		$p = 0.86$		$p = 0.70$
REL	6.64	6.29	6.36	6.39	$p = 0.10$		$p = 0.39$		$p = 0.15$
SCP	4.25	4.32	4.46	4.25	$p = 0.70$		$p = 0.68$		$p = 0.41$

Task product is either *Factual* (F) or *Intellectual* (I).
Task goal is either *Clear* (C) or *Amorphous* (A).
Effects of product and goal are tested using two-way ANOVA.
*, **, and *** indicate $p < 0.05$, 0.01, and 0.001, respectively.

Table 21: Importance ratings of the five criteria in context-independent judgments under different task conditions (Product×Goal).

	Group Mean (N = 28)				**Effects**				
Factor	**F+C**	**I+C**	**F+A**	**I+A**	**Product (F or I)**		**Goal (C or A)**		**Interaction**
TOP	6.49	6.39	6.17	6.31	$p = 0.79$		$p = 0.01$	*	$p = 0.29$
NOV	4.39	4.43	4.54	4.96	$p = 0.30$		$p = 0.16$		$p = 0.40$
UND	5.32	5.39	5.25	5.43	$p = 0.37$		$p = 0.88$		$p = 0.62$
REL	6.71	6.43	6.39	6.39	$p = 0.04$	*	$p = 0.02$	*	$p = 0.16$
SCP	4.11	4.36	4.11	4.29	$p = 0.23$		$p = 0.82$		$p = 0.85$

Task product is either *Factual* (F) or *Intellectual* (I).
Task goal is either *Clear* (C) or *Amorphous* (A).
Effects of product and goal are tested using two-way ANOVA.
*, **, and *** indicate $p < 0.05$, 0.01, and 0.001, respectively.

users rated REL as significantly more important in tasks looking for factual products (F) than those for intellectual understanding (I). Task goal also shows significant effects on the importance of topicality and (TOP) reliability (REL) at 0.05 level—users rated both TOP and REL as significantly more important in tasks with clear goals (C) than in those with amorphous ones (A).

Despite a few observed significant effects, overall task product and goal have very limited impact on the relative importance of the five factors in usefulness judgments, regardless of in an contextual judgment setting or a context-independent one—for any particular task type (Product×Goal), the five factors keep the same preferential ranking as we reported in § 5.1.2 regarding their importance in usefulness judgments.

5.1.5 Summary of Findings

In this section, we examined participants' responses to a multi-item questionnaire regarding the importance of five factors (topicality, novelty, understandability, reliability, and scope) in contextual and context-independent usefulness judgments, respectively. These responses mostly show high internal consistency. The results disclose answers to the following research questions:

- **RQ1.1** – How much do users weigh the five factors in contextual and context-independent usefulness judgments? (§ 5.1.2)

  Users generally believe all the five factors are important for usefulness judgments regardless of in an contextual or context-independent judgment setting. However, their responses suggest a preferential ranking of the five factors regarding their importance in usefulness judgments: topicality, reliability $\succ$ understandability $\succ$ novelty, scope. Our results disclose that topicality and reliability are two fundamental criteria of usefulness judgment, while understandability, novelty, and scope are secondary.

- **RQ1.2** – Do users weigh the five factors differently in contextual and context-independent usefulness judgments? (§ 5.1.3)

  Users' ratings regarding the importance of the five factors in two usefulness judgment settings are mostly consistent. They did rate novelty as significantly more important in

Table 22: Pearson's correlation matrix of variables.

	ERel	**Nov**	**Effort**	**Under**	**Relia**	**TRel**	**Usef**	**Under.ps**
Nov	0.70							
Effort	0.25	0.27						
Under	0.26	0.20	−0.36					
Relia	0.47	0.43	0.11	0.28				
TRel	0.65	0.49	0.18	0.19	0.45			
Usef	0.75	0.56	0.18	0.24	0.47	0.83		
Under.ps	0.27	0.23	−0.30	0.72	0.28	0.25	0.31	
Relia.ps	0.45	0.42	0.09	0.23	0.82	0.51	0.54	0.30

Light , dark , and darker shadings indicate the correlation is significant at 0.05, 0.01, and 0.001 levels.

contextual usefulness judgments than context-independent ones, but the magnitude of difference is very small and does not influence the preferential ranking of the five factors regarding their importance in usefulness judgments.

- **RQ1.3** – Do task product and goal affect the importance ratings of the factors? (§ 5.1.4) Task product and goal have very limited influence on the importance of the five factors. Task product and goal show a few statistically significant effects on the importance of topicality and reliability in context-independent usefulness judgments, but exhibit none in an contextual judgment setting. The observed differences do not influence the preferential ranking of the five factors regarding their importance in usefulness judgments.

5.2 WHAT AFFECTS EPHEMERAL RELEVANCE

This section examines factors related to ERel judgments using regression analysis. We also discuss factors related to the context-independent usefulness judgments (Usef) and compare with those for ERel. We examine three regression models:

- **M1** – The dependent variable (DV) is ERel. The independent variables (IVs) include other search result judgments (except Usef), the product and goal of the search task (both are binary variables), and users' familiarity with the topic of the task (assessed

Table 23: Regression analysis: ERel judgments as dependent variable (M1 and M2) and context-independent usefulness judgments (Usef) as dependent variable (M3).

	DV: ERel		**DV: Usef.**
Independent Variable	**M1**	**M2**	**M3**
(Intercept)	−0.95	−0.91	−0.09
Product: *Factual*	0.32	0.30	0.05
Goal: *Specific*	0.24	0.17	0.15
Familiarity	0.10	0.09	0.01
Novelty	0.48	0.40	0.17
Effort	0.13	0.11	0.06
Understandability	0.15	0.13	0.04
Reliability	0.12	0.13	−0.04
Topical Relevance	0.70	0.01	1.47
Usefulness	-	0.47	-
Understandability (post-session)	0.00	−0.03	0.07
Reliability (post-session)	−0.04	−0.11	0.15
Adjusted R^2	**0.640**	**0.702**	**0.735**

Light, dark, and darker shadings indicate the coefficient is significant at 0.05, 0.01, and 0.001 levels, respectively.

before the start of a session using a 7-point Likert scale).

- **M2** – The dependent variable is also ERel. In addition to M1's IVs, M2 further includes Usef as an IV.
- **M3** – The dependent variable is the context-independent usefulness judgments (Usef). M3 and M1 share the same list of IVs. We report M3 to make a comparison with M1.

We discuss the influence of the factors on ERel judgments mainly based on M1. We compare M2 with M1 to determine to what extent the context-independent usefulness judgments (Usef) help explain the contextual ones (ERel). We do not interpret the influence of variables based on M2 because the inclusion of Usef as an IV makes it difficult to explain the model (§ 5.2.7 discusses this issue in details).

We examine multicollinearity between variables using variance inflation factor (VIF). The IVs of each model satisfy VIF < 4. The VIF values are below the commonly suggested threshold (4–10) for concerns on multicollinearity issues (Menard, 2002).

Table 24: Changes in adjusted R^2 by excluding variables from M1 and M3 (variables are sorted by Δ adjusted R^2 in M1).

	Δ Adjusted R^2			
Removed Variables	**M1**		**M3**	
Novelty	−0.139	***	−0.017	***
Topical Relevance	−0.067	***	−0.276	***
Under. & Under.ps	−0.009	***	−0.004	**
Product: *Factual*	−0.006	***	0.000	
Effort	−0.006	***	−0.001	
Familiarity	−0.005	***	0.000	
Goal: *Specific*	−0.003	**	−0.001	
Relia. & Relia.ps	−0.003	*	−0.006	***

*, **, and *** indicate the difference between the reduced model and the original one is significant at 0.05, 0.01, and 0.001 levels.

Table 23 reports the results of the regression analysis. To determine the contribution of IVs, we examine the changes in adjusted R^2 caused by eliminating variables from the regression models and report the results in Table 24. Table 22 reports the correlation matrix between variables.

5.2.1 Topical Relevance

ERel and topical relevance judgments (TRel) are closely related, confirming that topicality is a salient factor for ERel judgments. The close relationship also indicates that current TREC-style relevance judgments (as assessed by the users themselves) are reasonable surrogates for ERel judgments.

Multiple evidence reveals a close relationship between ERel and topical relevance. M1 suggests that topical relevance has a significant positive effect on ERel judgments ($b = 0.70$, $p < 0.001$). The coefficient indicates that with other variables being equal, a higher level of topical relevance (by the TREC web track's criteria) increases the ERel rating by 0.70 unit in a 7-point scale. ERel and topical relevance also have a moderate-to-strong correlation ($r = 0.65$, $p < 0.001$). Moreover, excluding topical relevance from M1 leads to a significantly worse model ($p < 0.001$), reducing adjusted R^2 by 0.067 (the second largest magnitude).

5.2.2 Novelty

Novelty is the most important factor among the examined variables for ERel judgments, indicating that searchers hope to find novel useful information in a search session.

Novelty has a significant positive effect on ERel judgments ($b = 0.48$, $p < 0.001$). The coefficient of M1 indicates that with other conditions being equal, one-unit increase in the novelty judgment raises the ERel rating by 0.4 unit (both in a 7-point scale). Novelty also has a strong linear correlation with ERel ($r = 0.70$, $p < 0.001$). Moreover, excluding novelty from M1 yields a significantly worse model ($p < 0.001$) and a decline in adjusted R^2 by 0.139 (the largest magnitude of change in Table 24).

5.2.3 Effort

ERel judgments also relate to the effort spent on the results in a positive way, confirming that ERel captures users' interaction for acquiring useful information from the results. Effort exhibits a significant positive effect on ERel ($b = 0.13$, $p < 0.001$), suggesting that one-unit greater effort spend on a result increases the ERel judgment by 0.13 unit (with other conditions being equal). We also find a weak positive linear correlation between ERel and effort ($r = 0.25$, $p < 0.001$).

However, despite its statistically significant effect in M1, effort only seems to have a small practical contribution for explaining the variance of the ERel judgments. Excluding effort from M1 only brings down the adjusted R^2 by 0.006, although the reduced model is significantly different at 0.001 level. We suspect a possible reason is that the connection between ERel and effort is more complex than simply a linear relationship. § 5.5 analyzes this issue in detail.

5.2.4 Understandability

We collected users' understandability judgments twice in the experiment (post-click and post-session). The two judgments have a strong correlation ($r = 0.72$, $p < 0.001$), but they also have differences in 38% of the results. Their mean absolute difference is 0.63 (in a 7-point scale). This indicates that users' perceptions on the understandability of a result indeed undergo changes in a search session.

Model M1 shows that ERel only relates to the post-click understandability judgments but not the post-session ones. Although ERel has a weak positive linear correlation with both understandability judgments ($r = 0.26$ and 0.27, respectively), only the post-click judgments show a significant positive effect in M1 ($b = 0.15$, $p < 0.001$). The post-session judgments do not show any significant effect in M1, suggesting that it provides little value in addition to the post-click judgments for explaining the variance of the ERel judgments.

The relationship between ERel and the two understandability judgments discloses an advantage of ERel—it takes into account a user's ability to understand at a particular time of a search session. As a user's understanding varies over time, the user may prefer results with different understandability levels at different stages of a session, e.g., a user may expect to read easy-to-understand introductory texts such as a Wikipedia entry at the beginning of a session. Collecting ERel judgments potentially makes it possible to account for such issues in system design and evaluation.

5.2.5 Reliability

The post-click and post-session reliability judgments have a strong correlation ($r = 0.82$, $p < 0.001$). They are different in 43% of the results. Their mean absolute difference is 0.60 (in a 7-point scale). This suggests that users' perceptions on the reliability of a search result may also change in a session.

ERel also only relates to users' post-click reliability judgments but not the post-session ones. Although ERel has a moderate correlation with both reliability judgments ($r = 0.47$ and 0.45, respectively), only the post-click judgments have a significant positive effect on ERel ($b = 0.12$, $p < 0.05$).

ERel seems to account for users' perceptions on the reliability of search results at the time they examined the results. However, we believe this brings in a risk of performing ERel judgments. Unlike the subjective nature of understandability, the reliability of a result is a rather objective existence. It is reasonable to believe that after a search session's exploration, searchers may assess the reliability of results with better accuracy after a session (in post-session judgments). The observed difference between post-click and post-session reliability judgments may indicate the post-click reliability judgments are less accurate than the post-session ones, since a user may fail to accurately assess the reliability of a result during a search session due to the limited knowledge on the task. As a significant factor for ERel judgments, the possibly defective post-click reliability judgments may consequently reduce the quality of the ERel judgments as well.

5.2.6 Task Attributes

ERel judgments also relate to a few attributes of the search task. We only examined task product, goal, and users' familiarity with the task topic in this paper. M1 confirms that all three attributes have significant effects on ERel judgments.

Searching in a session targeting a *factual* product (compared with one for an *intellectual* product) has a significant positive effect on ERel judgments ($b = 0.32$, $p < 0.001$) in model M1. The coefficient suggests that with other variables being equal, ERel ratings in a *factual* task session are higher than those in an *intellectual* session by 0.32 unit.

Searching in a session with a *specific* goal (compared with an *amorphous* one) also shows a significant positive effect on ERel ($b = 0.24$, $p < 0.01$) in model M1. The coefficient suggests that with other variables being equal, ERel ratings in sessions with a *specific* goal are higher than those in sessions with an *amorphous* one by 0.24 unit.

In addition, M1 suggests that users' familiarity with the task topic also has a significant positive effect on ERel judgments ($b = 0.10$, $p < 0.001$)—with other conditions being constant, one-unit higher familiarity with the task topic increases the ERel ratings by 0.1 unit.

5.2.7 Context-independent Usefulness Judgments

M2 further includes the context-independent usefulness judgments (Usef) as an independent variable for ERel judgments. The inclusion of Usef improves the model significantly ($p < 0.001$), enhancing the adjusted R^2 by 0.062. This indicates that the context-independent usefulness judgments (Usef) can help other variables better explain the contextual usefulness judgments (ERel). The context-independent usefulness judgments are also effective surrogates for ERel judgments—they have a strong positive correlation ($r = 0.75$, $p < 0.001$), stronger than that between ERel and TRel ($r = 0.65$, $p < 0.001$).

M1 and M2 are mostly consistent on the influence of the variables except for a few cases. First, topical relevance does not show any significant effect in M2, probably because of its strong correlation with Usef. ($r = 0.83$, $p < 0.001$). As long as we include Usef. as an IV, topical relevance provides little additional value for explaining the variance of ERel. Second, the post-session reliability judgments (Relia.ps) have a negative coefficient in M2 ($b = -0.11$, $p < 0.05$), but we do not believe this indicates a negative relationship between ERel and Relia.ps. Instead, we suspect this is because Relia.ps has a significant positive effect on Usef., such that after including Usef. as an IV for ERel, M2 needs to estimate a negative coefficient on Relia.ps. to set off the implicit influence of Relia.ps brought by Usef.

M3 examines factors related to the post-session usefulness judgments. In contrast to ERel, Usef only relates to novelty, topical relevance, and the post-session understandability and reliability judgments. A comparison between M1 and M3 discloses many differences between ERel and Usef:

- ERel captures users' real time perceptions on the understandability and reliability of the results when they examined the results in a search session, while Usef only significantly relates to those after a search session while users performed the Usef judgments.
- ERel judgments are influenced by the search task while Usef judgments are not.
- Usef judgments do not account for the actual effort spent in a session for acquiring relevant/useful information from the result but ERel judgments do.
- Usef judgments relate to topical relevance by a greater extent than ERel judgments do (according to Table 24).

5.3 WHAT AFFECTS THE DIFFERENCE BETWEEN EREL AND USEF?

This section further examines factors related to the difference between ERel and Usef judgments. The purpose is to determine what differences of the two judgments we should expect

Table 25: Distribution of Diff = ERel − Usef ($N = 736$).

≤ −3	**−2**	**−1**	**0**	**1**	**2**	**≥ 3**
3.7%	8.8%	16.6%	39.9%	19.6%	7.2%	4.2%

in a certain condition. The two judgments have almost the same mean values (4.68 vs. 4.65), and their mean absolute difference is 0.97 in a 7-point scale. Table 25 reports the distribution of their difference. We perform a regression analysis, where the dependent variables are:

- **Diff = ERel − Usef** (signed difference) – to which extent a user rates the usefulness of a search result by a greater extent right after examining the result compared with after the search session.
- **|Diff| = |ERel − Usef|** (absolute difference) – to which extent the two usefulness judgments (ERel and Usef) vary (regardless of the sign).

We examine task attributes and other search result judgments as independent variables. The IVs also include the difference of the two understandability judgments and that of the two reliability judgments. We define ΔUnderstandability = Under − Under.ps and ΔReliability = Relia − Relia.ps. We include ΔUnderstandability and ΔReliability into Diff.'s IVs, but |ΔUnderstandability| and |ΔReliability| for |Diff|'s.

Table 26 shows that both models explain the variance of Diff and |Diff| significantly better than the null model at 0.001 level, but both models also have limited explanatory power (adjusted $R^2 = 0.262$ and 0.039). We suspect this is because 1) 39.9% of the results' ERel and Usef judgments do not vary, and 2) a large fraction of the difference may be caused by some random effects. However, we note a few independent variables do show significant effects.

5.3.1 Signed Difference

The signed difference between ERel and Usef judgments (Diff) are related to the following factors:

Table 26: Regression analysis: Diff and $\lvert$Diff$\rvert$ as dependent variable (Diff = ERel − Usef).

	Coefficients	
Independent Variable	**Diff**	**$\lvert$Diff$\rvert$**
(Intercept)	−0.86	0.34
Product: *Factual*	0.28	0.18
Goal: *Specific*	0.09	0.05
Familiarity	0.09	0.01
Topical Relevance	−0.77	−0.13
Novelty	0.31	0.02
Understandability	0.04	0.02
Reliability (post)	−0.03	0.04
Effort	0.08	0.03
ΔUnderstandability	0.07	-
ΔReliability	0.16	-
$\lvert\Delta$Understandability$\rvert$	-	0.14
$\lvert\Delta$Reliability$\rvert$	-	0.16
Adjusted R^2	**0.262**	**0.039**

Light , dark , and darker shadings indicate the coefficient is significant at 0.05, 0.01, and 0.001 levels, respectively.

Task Product – Working on a *factual* task session (compared with an *intellectual* one) shows a significant positive effect on Diff ($b = 0.28$, $p < 0.01$). With other variables being equal, we expect a higher signed difference between ERel and Usef judgments (such as a greater positive difference or a smaller negative one) by 0.28 unit in *factual* tasks than in *intellectual* ones. The actual mean signed difference we observed in *factual* tasks (Diff = 0.096) is also higher than that in *intellectual* ones (Diff = −0.048).

Task familiarity – Task familiarity shows a significant positive effect on Diff ($b = 0.09$, $p < 0.01$), suggesting that the signed difference of the two usefulness judgments raises as users' familiarity with the task increases.

Topical Relevance and **Novelty** – Topical relevance has a significant negative effect on Diff ($b = -0.77$, $p < 0.001$), while novelty has a significant positive one ($b = 0.31$, $p < 0.001$). We believe this is because users weigh the two factors differently in ERel and Usef judgments. Users weigh novelty by a greater extent in ERel judgments (as Table 24 shows), such that with other variables being equal, the difference of ERel and Usef judgments is positively correlated with the novelty of a search result. In contrast, users weigh topical relevance by a greater extent in Usef judgments than in ERel judgments, such that topical relevance is negatively correlated with the difference of ERel and Usef judgments.

Effort – The effort spent on the results shows a significant positive effect on Diff ($b = 0.08$, $p < 0.05$), suggesting the signed difference between ERel and Usef judgments is positively correlated with the effort spent on the result.

Overestimated reliability – It is reasonable to assume that users' context-independent reliability judgments are more accurate than their contextual ones because their knowledge increased during their exploration in a search session. Thus, ΔReliability indicates to which extent users overestimated the reliability of a result when they examined it. ΔReliability has a significant positive effect on the signed difference of ERel and Usef judgments ($b = 0.16$, $p < 0.01$), indicating that users' inaccurate perceptions on the reliability of a result is a possible reason for the difference between ERel and Usef judgments.

5.3.2 Absolute Difference

The absolute difference between ERel and Usef judgments (Diff.) are related to the following factors:

Task Product – Working on a *factual* task session (compared with an *intellectual* one) also has a significant positive effect on the absolute difference between ERel and Usef judgments ($b = 0.18$, $p < 0.05$), suggesting that with other variables being equal, we should expect a greater variation between the two usefulness judgments in *factual* tasks compared with in *intellectual* ones. The actual mean absolute difference we observed in *factual* tasks (Diff. = 1.04) is also higher than that in *intellectual* ones (Diff. = 0.89).

Topical Relevance – Topical relevance shows a significant negative effect on the absolute difference between ERel and Usef judgments ($b = -0.13$, $p < 0.05$). This suggests ERel and Usef judgments vary by a smaller extent for results with high topical relevance, or equivalently, ERel and Usef judgments vary by a greater extent for those with low topical relevance.

Changes in Understandability and Reliability – both ΔUnderstandability and

Table 27: Pearson's correlation between user experience measures and the mean value of the clicked results' judgments.

User Experience	mean ERel	mean Usef	mean TRel
Satisfaction	**0.50**	0.48	0.44
Frustration	−0.37	**−0.41**	−0.30
System Helpfulness	**0.40**	0.38	0.31
Goal Success	**0.51**	0.49	0.37
Session Effort	**−0.42**	**−0.42**	−0.33
Task Difficulty	**−0.46**	−0.43	−0.36

All correlations are significant at least at 0.01 level.

ΔReliability have significant positive effects on the absolute difference between ERel and Usef judgments. This suggests that a greater change in the users' perceptions on understandability and reliability is positively associated with a greater variation of the ERel and Usef judgments.

5.3.3 Summary

To summarize, this section shows that the difference between ERel and static usefulness judgments relates to a few factors. First, results confirm the hypothesis H5, suggesting that the difference of the two usefulness judgments relate to changes in users' perceptions on the understandability and reliability of the search results. Second, we show the difference also relates to a few task attributes, probably because ERel judgments are influenced by task attributes but Usef judgments are not.

5.4 EREL, TREL, USEF, AND USER EXPERIENCE

The main purpose of performing relevance judgment is to collect ground truth data to optimize and evaluate search systems. A good measure for relevance judgment should be able to identify high-quality results, such that presenting the results to users leads to a satisfactory search experience.

This section compares ERel, Usef, and TRel judgments for their abilities to correlate with users' search experience. We assume the quality of the clicked results in a session is a factor for the user's experience in that session. For each session, we use the mean ERel, Usef, and TRel of the clicked results as indicators for that session's search experience. We correlate the mean values of the judgments with users' perceptions on six search experience measures in the collected 112 search sessions. Table 27 reports the results.

Although ERel has many theoretical advantages, Table 27 shows that the collected ERel and Usef judgments have only slight differences in terms of correlating with the six user experience measures. Mean ERel of results has slightly stronger correlations with satisfaction, system helpfulness, goal success, and task difficulty, while mean Usef has a slightly stronger correlation with frustration. The differences in correlation values do not exceed 0.04, suggesting that whether to offer high ERel results or high Usef ones may not differ much in terms of enhancing user experience in a session.

The limited practical advantage of ERel compared with Usef judgments in terms of correlating with user experience is not unexplainable. First, the collected ERel and Usef judgments do not vary greatly ($r = 0.75$). Second, as we discussed in Section 5.2.5, users may not have enough knowledge to correctly assess the credibility of information during a search session, which may consequently reduce the quality of the collected ERel judgments.

Considering that it requires a more complex setting (and probably a higher cost) to collect ERel judgments, collecting static usefulness judgments seems a more practical choice. Table 27 shows a clear difference between TRel and both ERel and Usef in terms of correlating with the six user experience measures (0.04–0.14). This suggests that using usefulness as the criterion for assessing search results better correlate with users' search experience.

5.5 EREL, EFFORT, AND INTERACTION

Previous sections examined the relationship between ERel and other variables using linear models. However, a deeper analysis shows that the regression models concealed complex and non-linear relationships of the variables.

This section examines the ERel of and the effort spent on results with different understandability and reliability levels (post-click judgments). We group results into five levels to make the sample size of each group as close as possible (although group size still varies a lot due to the very skewed distribution of users' judgments). The five understandability levels are 1–2 ($N = 45$), 3–4 ($N = 69$), 5 ($N = 84$), 6 ($N = 137$), and 7 ($N = 401$). The five reliability levels are 1–3 ($N = 117$), 4 ($N = 112$), 5 ($N = 152$), 6 ($N = 154$), and 7 ($N = 201$). Figure 8 plots the results.

ERel, Effort, and Understandability – Users acquired a lot of useful information (mean ERel = 4.87) with only a small amount of effort (mean effort = 1.93) from the results with the highest level of understandability (7). While encountering results that are more difficult to understand (Under. = 6 and 5), users spent significantly greater effort (mean effort = 3.01 and 3.49), and they were able to acquire a similar amount of useful information (mean ERel = 5.04 and 4.92). When the results are even more difficult to understand (Under. =3–4), the trend of spending more effort stopped (mean effort = 3.54), and the acquired amount of useful information also declined significantly (mean ERel = 3.91). When the results are extremely difficult to understand (Under =1–2), users started to abandon examining results, spending fewer effort (mean effort = 3.16) and acquiring very limited amount of useful information (mean ERel = 2.67).

ERel, Effort, and Reliability – We also observe a similar pattern on results with different reliability levels. Users acquired a lot of useful information (mean ERel = 5.36 and 5.45) with a small amount of effort (mean effort = 2.57 and 2.42) from the results with the two highest reliability levels (Relia. =7 and 6). When the results provide less reliable information (Relia. = 5 and 4), they spent significantly greater effort (mean effort = 2.86 and 2.86), but started to acquire a significantly fewer amount of useful information (mean ERel = 5.01 and 4.04). They abandoned examining results when the reliability level is very low (1–3), spending the least amount of effort (mean effort = 1.90) and acquiring very limited useful information (mean ERel = 2.69).

Figure 8 discloses that the process of examining search results and acquiring useful information involves complex interaction, suggesting the dynamic nature of ephemeral relevance.

5.6 ANSWERS TO RQ1

To conclude, this section answered the research question RQ1, and verified the validity of the seven hypotheses.

- **RQ1** – On what criteria do users assess ephemeral relevance (ERel)? How does ephemeral relevance judgment differ from topical relevance judgment and context-independent usefulness judgment?

  Results show ephemeral relevance (ERel), topical relevance (TRel), and context-independent usefulness judgments (Usef) are highly related, but sufficiently different. Both ERel and the difference between ERel with Usef depend on multiple factors, including topical relevance (TRel), the novelty of the search result, the effort spent, the perceived understandability and reliability of the results, changes in searchers' perceptions of understandability and reliability, and task attributes. The usefulness criteria for assessing ERel and TRel are mostly consistent, both depending on topicality, reliability, understandability, scope, and novelty.

- **H1.1** – Searchers may have different criteria for contextual (ERel) and context-independent usefulness judgments (Usef).

  Results show searchers did rated novelty as slightly but significantly more important in contextual usefulness judgment (ERel) than in context-independent ones (Usef). However, the criteria for assessing usefulness in contextual and context-independent settings are mostly consistent.

- **H1.2** – ERel is related to users' perceptions on the topicality, novelty, understandability, and reliability of the search result at the time of assessing ERel.

  Similar to previous studies on factors for relevance judgments (Xu & Chen, 2006; Y. Zhang et al., 2014), we found that ERel judgments also significantly relate to topicality, novelty, understandability, and reliability. Particularly, ERel depends on users' real time perceptions on the understandability and reliability of the results at the time of examining the results, confirming that ERel indeed captures users' state of mind in a search session.

- **H1.3** – ERel is related to the effort spent on the result.

ERel judgments are significantly affected by the effort spent on the results, indicating that ERel depends on users' actual interaction with the results (such as how much effort to spend).

- **H1.4** – ERel is related to the search task.

  ERel judgments significantly relate to three search task attributes, suggesting that we may expect certain variation of ERel judgments in different types of tasks.

- **H1.5** – ERel differs from TRel and Usef.

  Results confirm that ERel and Usef. judgments are different in many aspects. Particularly, Usef. only significantly relates to topicality, novelty, and users' perceptions on understandability and reliability in post-session judgments.

- **H1.6** – The difference between ERel and Usef judgments is influenced by the difference of users' perceptions on the results in the two judgments.

  Our study confirms that users' perceptions regarding the understandability and reliability of a result undergo changes in a session. Although such changes had only happened in a limited fraction of results, they do contribute significantly to the difference between ERel and Usef.

- **H1.7** – ERel correlates with user experience better than TRel and Usef judgments.

  We found both ERel and Usef have significantly better correlations with user experience than TRel. However, ERel shows only limited advantages over Usef regarding correlating with user experience measures.

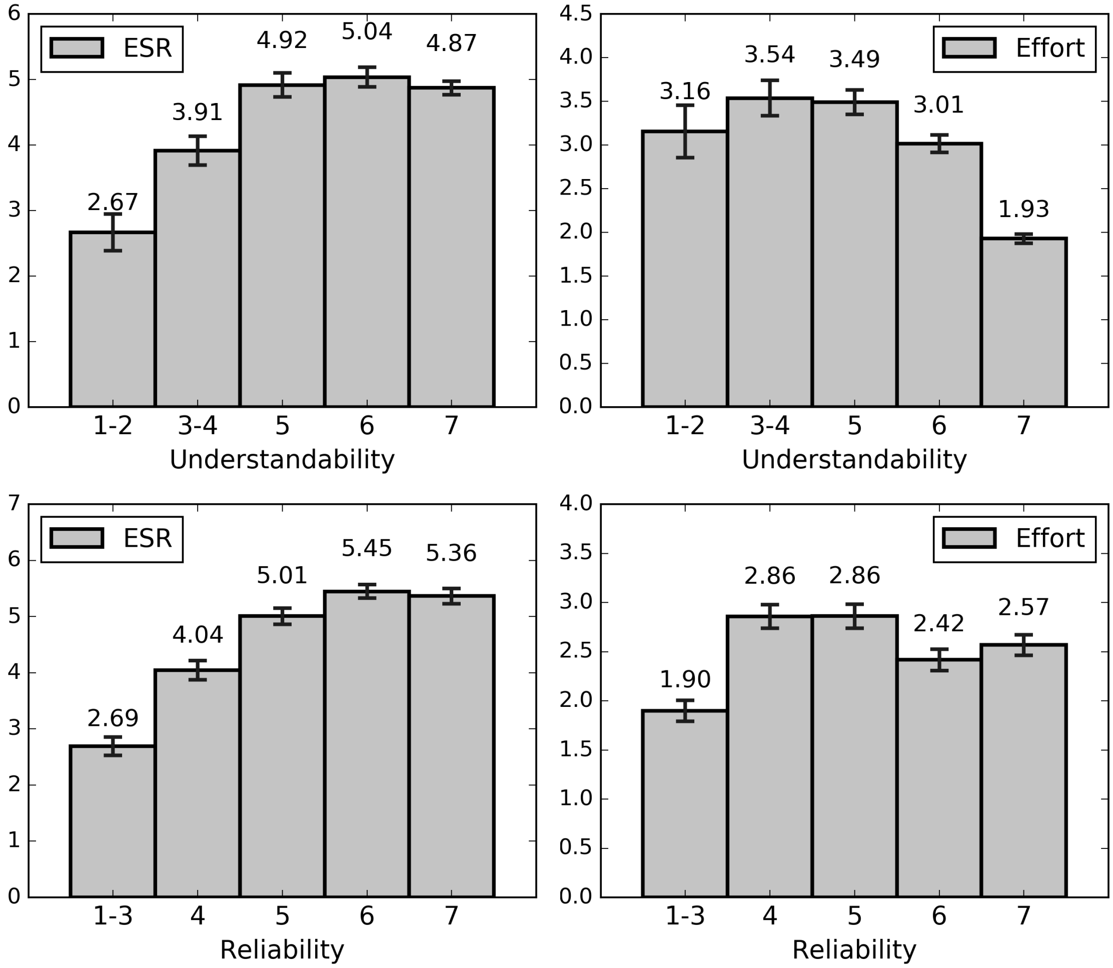


Figure 8: Mean ERel and effort (the error bars are standard error) for results with different novelty, understandability, and reliability levels (in post-examination judgments).

6.0 PREDICTING EPHEMERAL RELEVANCE

This chapter studies and evaluates techniques for predicting ephemeral relevance (ERel) in two different scenarios. We propose five hypotheses related to RQ2:

- **RQ2** – What information other than users' explicit judgment tells the ERel of a search result? How well do these information predict ERel judgments?
- **H2.1** – The ERel of a result is related to the characteristics of the result document.
- **H2.2** – We can predict the ERel of a result based on the text similarity between a query and the result.
- **H2.3** – We can predict the ERel of a result based on past search history in the session.
- **H2.4** – We can infer the ERel of a result afterwards based on how searchers interact with the result and the SERP showing the result.
- **H2.5** – We can infer the ERel of a result afterwards based on how searchers behave in follow-up searches in the session.

We examine these hypotheses and answer RQ2 in following sections: § 6.1 defines the tasks of predicting ERel and the scenarios of applying ERel prediction; § 6.2 introduces the prediction techniques, focusing on the prediction features; § 6.3 examines H2.1–H2.5 and answers RQ2 by comparing prediction performance using different information (features).

6.1 PREDICTION TASKS

User Study 1 collected ERel judgments of the clicked results. However, it relies on users' explicit feedback. The approach is also expensive and difficult to scale up. In addition, the process of collecting ERel judgments requires genuine search contexts, which has many restrictions compared with the TREC-style context-independent judgments, where hired assessors annotate preassigned sets of documents one by one. For example, one cannot preassign a set of documents for searchers to assess their ERel (because it requires search contexts to judge ERel). For these reasons, this chapter aims to develop techniques for predicting ERel of results based on implicit feedback.

ERel prediction can be applied to search engines in two different scenarios, where we use different information for prediction:

- Search result ranking – when users submitted a search query, the search engine can rely on the search query and past search history in the search session (if any) to predict the ERel of results. The predicted ERel can be used for ranking search results. Previous studies show that ad hoc search models and previous search activities (e.g., past search queries and clicks) are effective for predicting topical relevance of results (X. Shen et al., 2005; Agichtein, Brill, & Dumais, 2006; Bennett et al., 2012; D. Guan et al., 2013; Luo et al., 2014). Here the presented work extends these approaches to predict ERel, which is related to but different from topical relevance (as the previous chapter examined).
- Search engine evaluation and optimization – search engine can rely on users' interaction with a result (e.g., click dwell time) and those afterwards in the search session (e.g., follow-up search queries and clicks) as implicit feedback to predict the ERel of results. Topical relevance judgments, when available, may also help the prediction. The predicted ERel can be useful in different ways. For example, the search engine can diagnose its performance based on the predicted ERel of the results. The search engine can also use the predicted ERel as relevance labels to optimize search result ranking. Similar approaches for predicting topical relevance labels received great success in the past decades in search engine companies (Agichtein, Brill, Dumais, & Ragno, 2006; Joachims et al., 2005, 2007). This study hopes to generalize these approaches for topical relevance to the case of ERel.

Therefore, we study two different ERel prediction tasks:

- **RQ2.1** – How well can we predict ERel before users are presented the result?

- **RQ2.2** – How well can we infer ERel from users' interaction with the result and those afterwards in the search session?
- **RQ2.3** – How well can we infer ERel based on both implicit feedback information in RQ2.2 and explicit topical relevance judgments?

Note that although both information before and after searchers clicked on the search results is available in the scenario of RQ2.2 and RQ2.3, normally we do not use the former information in RQ2.2 and RQ2.3. This is because an important application of RQ2.2 and RQ2.3 is to automatically generate training labels for search result ranking (similar to the scenario of RQ2.1), which also relies on information such as past search activities. Using both information in RQ2.2 and RQ2.3 has the risks of overfitting if such generated labels were applied to train ranking models that included similar features.

The rest of this chapter introduces the techniques for predicting ERel, and evaluates their effectivenesses.

6.2 APPROACHES

This section describes the techniques for predicting ephemeral relevance (ERel), focusing on the prediction features. We examine these features by their correlations with ERel and other search result judgments.

6.2.1 Regression

We use Gradient Boosted Regression Trees (GBRT) (Friedman, 2001, 2002), along with different sets of features, to predict ephemeral relevance (ERel). GBRT is a generic supervised learning technique for regression and classification problems based on gradient boosting. We also examined other regression models such as linear regression and Poisson regression. They all performed worse than GBRT using the same sets of features in our task. Thus, we only present results using the GBRT regression.

We focus on comparing the effectiveness and contribution of different sets of features in predicting ERel. We divide the features into different groups based on the types of user interaction information involved in computing the features:

- No user interaction – this group of features does not rely on any user interaction information. They can be applied in any situation to predict ERel. For example, features in this group include the characteristics of the result (e.g., the length of the result document, whether it is from the .gov domain).
- Only current query – this group of features uses the current search query, e.g., the query-document text similarity measures. These features can be applied to RQ2.1.
- Past user interaction – this group of features relies on past search activities in a search session, such as previous search queries and clicked results. They can only be applied when such information is available. For example, at the beginning of a search session, such information is not available. These features can be applied to RQ2.1.
- User interaction afterwards – this group of features relies on users' interaction after they were presented the result (for example, the time spent on the result, and follow-up searches and clicks in the session). Such information can be applied to RQ2.2.

These features, based on the factors they modeled, can also be divided into different types, such as: query-document text similarity (§ 6.2.2), the relation between the result and other search queries (§ 6.2.3), the relation between the result and query reformulation (§ 6.2.4), the relation between the result and clicks (§ 6.2.5), effort and understandability (§ 6.2.6), credibility (§ 6.2.7), and session-level global information (6.2.8)

6.2.2 Text Similarity

Text similarity features are based on the state-of-the-art ad hoc search models. These models only rely on the similarity between the current search query and the result document for relevance ranking. These features help examine how well ad hoc search models can help predict ERel of results.

QL is a normalized version of the query likelihood language model (Zhai & Lafferty, 2001), as in Equation 6.1. It normalizes the standard query likelihood score ($\log P(q|D)$) by

Table 28: Correlation of text similarity features with searchers' ERel and other judgments.

	ERel	TRel	Usef	Effort	Novel	Relia	Under
QL	0.22	0.17	0.19	0.02	0.20	0.20	0.12
SDM	0.22	0.18	0.19	0.04	0.20	0.20	0.14
BM25	0.24	0.19	0.21	−0.11	0.19	0.26	0.23

- Light , dark , and darker shadings indicate the correlations are significant at 0.05, 0.01 and 0.001 levels, respectively. Unshaded cells are not significant at 0.05 level.

query length ($\frac{1}{|q|}$), because the standard QL score is not comparable among queries of different lengths. Here $|q|$ refers to the number of words in the search query (query length). When computing QL, the document model is smoothed using Dirichet smoothing ($\mu = 3500$), with the Clueweb12 dataset as the background corpus (all following features based on language modeling use the same setting).

$$\text{QL} = \frac{1}{|q|} \cdot \log P(q|D) = \frac{1}{|q|} \cdot \sum_{w} \log P(w|D) \tag{6.1}$$

SDM is a normalized version of the sequential dependence model (Metzler & Croft, 2005). SDM improves QL by taking into account query term dependency. Similar to QL, SDM is also normalized by query length. When computing SDM, the weights for unigram, bigram (#2), and skip gram (#uw8) were set to 0.80, 0.15, and 0.05, respectively (Bendersky, Metzler, & Croft, 2010).

BM25 is a normalized version of the BM25 model (S. Robertson, Zaragoza, & Taylor, 2004). Similar to QL and SDM, BM25 is also normalized by the query length $\frac{1}{|q|}$. When computing BM25, k_1 and b were set to 1.2 and 0.75, respectively. Corpus statistics were computed based on the Clueweb12 dataset.

Table 28 reports the correlations (Pearson's r) of QL, SDM, and BM25 with participants' ratings on different variables for the clicked results. The table shows that all the three typical ad hoc search models have significant positive correlations with both ERel and topical relevance, yet the strengths of the correlations are weak (with Pearson's r about 0.2). This suggests that it is necessary to further take into account contextual information (such as past search activities) for better prediction of ERel. However, it should also be noted that the correlations reported in Table 28 are probably underestimated, because the majority of the clicked results are top-ranked search engines results, which are a subset of documents with overall high text similarity with the search query. Therefore, results in Table 28 do not mean text similarity features only correlate with ERel at about $r = 0.2$ in the whole web collection.

6.2.3 Other Queries

The similarity of the result with other queries in a session may also help predict ERel. This is based on previous findings (X. Shen et al., 2005; Bennett et al., 2012; D. Guan et al., 2013) that using past search queries in a session as relevance feedback for the current query can improve search performance (as measured by topical relevance). In addition, being similar to follow-up search queries in a session also confirms that the content of the clicked results are probably of interest to the users—otherwise they would not further search similar contents.

The feature **PWD PrevQ** and **PWD NextQ** compute the statistics (maximum, minimum, or mean values) of the normalized log query likelihood (QL) scores between the result and each other queries in the session. **PrevQ** and **NextQ** stand for considering queries before or after the current search query. The label **max**, **min**, or **avg** stand for using the maximum, minimum, or mean value of features, respectively. In addition, considering search queries are not always well formulated, two variants of the features considering only queries with clicks (assumed to be of better quality than those without clicks) were computed (labeled as **w/ clicks**). For example, **PWD PrevQ max w/ clicks** refers the maximum log QL score between the result and previous search queries with at least one click.

Table 29 reports the correlations of these features with users' judgments on ERel and topical relevance. Note that within a session, a clicked result does not always have the previous query or follow-up query to compute these features. The reported correlations are

Table 29: Correlation of features using other queries with searchers' ERel and topical relevance judgments.

	N	ERel	TRel	Usef	Effort	Novel	Relia	Under
PWD PrevQ max	483	0.21	0.17	0.16	0.11	0.16	0.13	0.12
PWD PrevQ min	483	0.22	0.15	0.17	0.14	0.21	0.10	0.08
PWD PrevQ avg	483	0.24	0.18	0.19	0.13	0.21	0.13	0.12
PWD NextQ max	528	0.21	0.22	0.23	−0.03	0.21	0.18	0.17
PWD NextQ min	528	0.20	0.15	0.19	−0.01	0.19	0.19	0.17
PWD NextQ avg	528	0.22	0.22	0.23	−0.03	0.22	0.23	0.19
PWD PrevQ max w/ clicks	460	0.23	0.19	0.17	0.14	0.18	0.16	0.10
PWD PrevQ min w/ clicks	460	0.22	0.15	0.16	0.14	0.22	0.13	0.10
PWD PrevQ avg w/ clicks	460	0.24	0.19	0.18	0.14	0.21	0.16	0.12
PWD NextQ max w/ clicks	508	0.20	0.22	0.20	−0.03	0.20	0.16	0.14
PWD NextQ min w/ clicks	508	0.19	0.17	0.20	−0.01	0.19	0.20	0.20
PWD NextQ mean w/ clicks	508	0.20	0.21	0.21	−0.03	0.20	0.20	0.18

- N stands for the number of results that can compute this feature.
- Light, dark, and darker shadings indicate the correlations are significant at 0.05, 0.01 and 0.001 levels, respectively. Unshaded cells are not significant at 0.05 level.

among results that have the information to compute these features.

As the figure shows, both the similarities of the result with previous queries (**PrevQ**) and follow-up queries (**NextQ**) have significant correlations with ERel and topical relevance. The strengths of the correlations for previous queries and follow-up queries seem comparable. In addition, restriction on queries does not seem to make obvious difference.Using different statistics (max, min, or avg) does not seem to show difference in terms of the correlation strengths as well. In addition, the table shows that the **PWD PrevQ** and **PWD NextQ** features differ in their correlations with Effort and Under, suggesting they are also sufficiently different.

The final regression model included **PWD PrevQ mean** and **PWD NextQ mean** as features. Other features were dropped due to their similarities with the two adopted features.

6.2.4 Query Reformulation

In addition to the similarity with other queries in a session, this section investigates more complex document-query similarity features by examining specific query reformulation patterns.

A query reformulation refers to the transition from one search query to another. The vocabulary difference of the two queries can be divided into three sets of words: ADD – the set of words added to the new query; RMV – the set of words removed from the old query; KEEP — the set of words included in both queries. These different sets of words may indicate different intentions of users. Previous work (D. Guan et al., 2013; Luo et al., 2014) also showed that considering these three sets of words separately can lead to improvements compared with those considering them altogether.

Similarly, this section considers the probability of these three sets of words in the result document separately. Table 30 and 31 list all the examined features and their correlations with users' judgments. The label **PrevQ** or **NextQ** stand for whether the sets of words are derived from query reformulations happened prior to or after clicking on the result. **ADD**, **RMV**, and **KEEP** stand for the three corresponding sets of words discussed above. **max**, **min**, or **avg** stand for using the maximum, minimum, or average values of the log

Table 30: Correlation of query reformulation features (using previous query reformulations) with searchers' ERel and other judgments.

	N	ERel	TRel	Usef	Effort	Novel	Relia	Under
PWD PrevQ ADD max	456	0.03	0.01	0.01	−0.07	−0.09	0.07	0.10
PWD PrevQ ADD min	456	0.13	0.04	0.08	−0.01	0.08	0.08	0.10
PWD PrevQ ADD avg	456	0.10	0.02	0.06	−0.02	0.03	0.07	0.11
PWD PrevQ RMV max	388	−0.01	0.01	−0.02	−0.06	−0.12	−0.02	0.06
PWD PrevQ RMV min	388	0.12	0.08	0.09	−0.04	0.02	0.07	0.08
PWD PrevQ RMV avg	388	0.09	0.08	0.07	−0.06	−0.03	0.04	0.12
PWD PrevQ KEEP max	479	0.13	0.11	0.08	−0.03	−0.03	0.04	0.00
PWD PrevQ KEEP min	479	0.15	0.10	0.13	−0.06	−0.00	0.06	0.03
PWD PrevQ KEEP avg	479	0.15	0.10	0.12	−0.05	−0.01	0.06	0.03

- N stands for the number of results that can compute this feature.
- Light , dark , and darker shadings indicate the correlations are significant at 0.05, 0.01 and 0.001 levels, respectively. Unshaded cells are not significant at 0.05 level.

Table 31: Correlation of query reformulation features (using follow-up query reformulations) with searchers' ERel and topical relevance judgments.

	N	ERel	TRel	Usef	Effort	Novel	Relia	Under
PWD NextQ ADD max	348	0.07	0.10	0.06	−0.07	0.00	0.06	0.02
PWD NextQ ADD min	348	0.08	0.00	0.07	−0.00	0.02	0.07	0.03
PWD NextQ ADD avg	348	0.07	0.04	0.07	−0.05	−0.01	0.05	0.04
PWD NextQ RMV max	326	0.24	0.19	0.20	−0.04	0.10	0.17	0.16
PWD NextQ RMV min	326	0.20	0.07	0.15	0.02	0.10	0.21	0.12
PWD NextQ RMV avg	326	0.24	0.11	0.18	−0.02	0.11	0.21	0.16
PWD NextQ KEEP max	375	0.10	0.14	0.07	−0.04	0.02	0.03	0.02
PWD NextQ KEEP min	375	0.10	0.15	0.18	−0.09	0.11	0.11	0.01
PWD NextQ KEEP avg	375	0.12	0.19	0.19	−0.10	0.10	0.12	0.03

- N stands for the number of results that can compute this feature.
- Light, dark, and darker shadings indicate the correlations are significant at 0.05, 0.01 and 0.001 levels, respectively. Unshaded cells are not significant at 0.05 level.

Table 32: Correlation of click dwell time with searchers' ERel and topical relevance judgments.

	N	ERel	TRel	Usef	Effort	Novel	Relia	Under
dwell	736	0.35	0.31	0.34	0.32	0.30	0.24	0.06
logdwell	736	0.44	0.38	0.43	0.36	0.41	0.34	0.12

- N stands for the number of results that can compute this feature.
- Light , dark , and darker shadings indicate the correlations are significant at 0.05, 0.01 and 0.001 levels, respectively. Unshaded cells are not significant at 0.05 level.

probabilities for the set of words as features. For example, **PWD PrevQ ADD max** refers to the maximum log probability of the ADD words in the clicked result where the ADD words are derived from query reformulations in past queries of the session.

As Table 30 shows, the probability of the KEEP words in previous queries ("PWD PrevQ KEEP") generally shows significant weak positive correlations with ERel. In contrast, Table 31 shows that the probability of the removed words (RMV) in searchers' follow-up query reformulations seems a sign that the clicked result has high ERel. This is probably because a reason to remove a word in follow-up queries is that searchers' information needs related to the removed words were satisfied in the clicked result, such that searchers do not need to further search similar contents.

Overall Table 30 and 31 shows that separately looking into the probability of these different types of words in query reformulation may be useful for predicting ERel. The regression model included **PWD PrevQ ADD min**, **PWD PrevQ RMV min**, **PWD PrevQ KEEP min**, **PWD NextQ RMV max**, **PWD NextQ RMV min**, **PWD NextQ KEEP min**, and **PWD NextQ KEEP mean** as features.

6.2.5 Click

Click dwell time is widely adopted as an important feature for inferring result relevance (Kelly & Belkin, 2004; White & Kelly, 2006; Y. Kim et al., 2014a; Y. Kim, Hassan, White,

Table 33: Correlation of click similarity features with searchers' ERel and topical relevance judgments.

	N	ERel	TRel	Usef	Effort	Novel	Relia	Under
PWD click PrevQ avg	460	0.29	0.27	0.25	0.17	0.27	0.26	0.11
PWD click PrevQ max	460	0.24	0.24	0.20	0.14	0.18	0.22	0.12
PWD click PrevQ min	460	0.24	0.23	0.22	0.12	0.25	0.23	0.10
PWD click NextQ avg	508	0.25	0.32	0.31	0.07	0.24	0.17	0.08
PWD click NextQ max	508	0.14	0.25	0.21	0.04	0.15	0.10	0.06
PWD click NextQ min	508	0.28	0.31	0.31	0.05	0.24	0.17	0.10

- N stands for the number of results that can compute this feature.
- Light, dark, and darker shadings indicate the correlations are significant at 0.05, 0.01 and 0.001 levels, respectively. Unshaded cells are not significant at 0.05 level.

& Zitouni, 2014b). Table 32 also examines whether result dwell time indicates result ERel. Both the raw dwell time and the log value were examined (because the distribution of dwell time is not linear). Results show log dwell time has a significant moderate correlation with ERel (and stronger than that for the raw dwell time), confirming that click dwell time is probably also an important feature for predicting ERel in addition to its success in inferring topical relevance. The regression model included the log dwell time as a feature.

In addition, this study also examines the similarity of the result with other clicked results in the session as features. The similarity is measured using the normalized QL score, where the content of the clicked result is considered as the query representation. Table 33 examines these features. Results show that being similar to clicked results in both previous queries and follow-up searches are positively correlated with the result's ERel. The regression model included all features in Table 33.

6.2.6 Effort and Understandability

As Chapter 5 examined, the difficulty in understanding contents in a clicked result is related to the amount of effort spent on the result, and they have complex relations with the amount of useful information searchers acquired from the result (ERel). Therefore, this section also includes features that may indicate the understandability and the effort spent on the results. Table 34 shows these features and their correlations with users' judgments.

The first group of features is related to the length of the clicked result document—the longer the document, the longer time (and probably the more effort) searchers may spend to examine the result (Smucker & Clarke, 2012; Y. Kim et al., 2014b). This group of features include: the number of characters (#chars), words (#words), and sentences(#sents) in the document. In addition, searchers may intentionally only read a part of the result relevant to their information needs. Therefore, "#sents w/ query term" computes the number of sentences with the occurrences of at least one query term, as an indicator for the length of the clicked result document fragments that are relevant to the query's information needs.

Table 34 shows that these features show significant and positive correlations with the effort spent, indicating they may capture the potential effort spent on the results to some

Table 34: Correlation of effort and understandability features with searchers' ERel and topical relevance judgments.

	N	ERel	TRel	Usef	Effort	Novel	Relia	Under
#chars	736	0.07	0.14	0.09	0.16	0.11	0.08	−0.03
#words	736	0.06	0.13	0.06	0.14	0.09	0.07	−0.03
#sents	736	0.08	0.11	0.06	0.15	0.10	0.07	−0.03
#sents w/ query term	736	0.09	0.16	0.12	0.15	0.12	0.07	−0.01
ARI	736	−0.08	−0.00	−0.06	0.03	−0.02	0.06	−0.11
LIX	736	−0.07	0.00	−0.05	0.04	−0.01	0.10	−0.11
CLI	736	0.01	0.05	0.01	0.06	0.01	0.17	−0.05
%heading	736	0.03	0.06	0.06	0.01	0.01	0.06	−0.02
%link	736	0.06	0.13	0.09	0.04	0.07	0.13	0.00
%table	736	0.03	−0.02	−0.05	−0.04	−0.00	−0.07	0.05
%em	736	0.10	0.10	0.09	−0.03	0.07	0.06	0.05
%div	736	0.09	0.13	0.10	0.03	0.07	0.10	0.01
%li	736	0.08	0.13	0.09	0.03	0.06	0.10	0.01
%p	736	0.09	0.10	0.08	0.01	0.06	0.07	0.02

- N stands for the number of results that can compute this feature.
- Light , dark , and darker shadings indicate the correlations are significant at 0.05, 0.01 and 0.001 levels, respectively. Unshaded cells are not significant at 0.05 level.

extent. In addition, all these features also shows significant correlations with TRel, but mostly no correlation for ERel and Usef. This may suggest that the TREC style relevance judgments are more likely to be influenced by the length of the result—for example, the criteria count results with minimal relevant information as marginally relevant, but the chances of containing minimal relevant information in a result may increase as the length of the result increases.

In addition, three widely adopted readability indicators were included into regression features, as readability may be closely related to the understandability of results. ARI stands for the Automated Readability Index (Senter & Smith, 1967), which measures readability based on the average length of words and sentences (long words and long sentences affect readability). LIX is another measure for readability based on sentence length and the proportion of long words (more than 6 letters) (Björnsson, 1983). CLI stands for the ColemanLiau index (Coleman & Liau, 1975), which is also based on sentence and word lengths.

Table 34 shows that ARI and LIX do weakly positively correlate with searchers' assessments on the difficulty in understanding the content of the result (Under). CLI does not show any significant correlation with UNDER, but has a significant correlation with the reliability of the result (Relia). Table 34 suggests that these readability indicators may be helpful in predicting ERel (although they do not directly correlate with ERel).

In addition, the effort and readability of web pages may also relate to its structure. Therefore, this study also examines related features. "%heading" measures the proportion of the texts in heading-like HTML tags such as <h1>and <h2>. "%link" counts the proportion of anchor texts in the result document. "%table", "%em", "%div", "%li", and "%p", respectively, measures the amount of texts in table (e.g., <table>, <tr>, <td>), emphasis (e.g., <b>, <i>, <em>, <strong>), div (<div>), list (<li>), and paragraph (<p>) HTML tags. Some of the features also show correlations with ERel and/or others. The regression model included all features in Table 34 except %heading and %table.

6.2.7 Credibility

This study uses the domain of the result URL as features for their credibility. For example, URLs from the .gov domain may be more reliable. Table 35 examines these features. As

Table 35: Correlation of credibility features with searchers' ERel and topical relevance judgments.

	N	ERel	TRel	Usef	Effort	Novel	Relia	Under
.gov	736	0.05	0.08	0.04	0.11	0.07	0.24	−0.03
.edu	736	−0.03	0.02	0.01	0.03	−0.03	0.02	−0.08
.net	736	−0.06	−0.07	−0.09	−0.01	−0.05	−0.06	0.00
.com	736	−0.01	−0.05	0.01	−0.13	−0.07	−0.07	0.05

- N stands for the number of results that can compute this feature.
- Light , dark , and darker shadings indicate the correlations are significant at 0.05, 0.01 and 0.001 levels, respectively. Unshaded cells are not significant at 0.05 level.

the results show, there is a significant weak positive correlation between the .gov domain and reliability. In addition, other domains such as .edu, .net, and .com also have some weak correlations with the assessed result properties.

6.2.8 Session

ERel of a result may also relate to the situation of visiting the result in a session. For example, with the time goes by (and when searchers have a better understanding of the topic), it may become easier or harder for them to acquire useful information from results. On the one hand, with a better understanding on the search task, searchers may be easier to locate and comprehend the contents, and are more likely to acquire useful information. On the other hand, with more knowledge, the novelty of the results may decline, and it may be more difficult for searchers to acquire useful information.

This section examines the time spent, the number of queries issued, and the number of clicks before and after searchers were presented with the results. Table 36 reports the correlations of these features with users' judgments.

As the results show, the time spent in a search session before searchers were presented with the result (**time prev**) has a significant negative correlation with the result's novelty and consequently its ERel and topical relevance. This shows that with the time goes by, it

Table 36: Correlation of session features with searchers' ERel and topical relevance judgments.

	N	ERel	TRel	Usef	Effort	Novel	Relia	Under
time prev	736	−0.12	−0.14	−0.14	−0.03	−0.23	−0.06	0.00
time post	736	0.03	0.11	0.07	0.03	0.11	0.00	0.02
#queries prev	736	−0.10	−0.11	−0.14	−0.02	−0.16	−0.11	−0.01
#queries post	736	−0.03	−0.02	−0.09	−0.04	−0.03	−0.10	0.01
#click prev	736	−0.13	−0.13	−0.15	−0.03	−0.21	−0.11	−0.04
#click post	736	−0.07	−0.03	−0.07	−0.09	−0.01	−0.11	−0.01
#queries	736	−0.09	−0.09	−0.16	−0.04	−0.12	−0.14	−0.00
#click	736	−0.18	−0.15	−0.20	−0.11	−0.20	−0.20	−0.05

- N stands for the number of results that can compute this feature.
- Light , dark , and darker shadings indicate the correlations are significant at 0.05, 0.01 and 0.001 levels, respectively. Unshaded cells are not significant at 0.05 level.

typically becomes more difficult to satisfy searchers novelty requirement. On the other hand, the time left in a session shows a positive correlation with novelty and topical relevance. This suggests that prediction of ERel may benefit from taking into account its time in a session.

However, it has been observed that, regardless of before or after clicking on a result, the number of queries issued and the number of results clicked are overall negatively correlated with the clicked result's ERel. This is probably due to that the number of queries and clicks in a session are also indicators of the session's search performance. For example, as Smith and Kantor (2008) showed, searchers may compensate limited search performance by spending more effort such as submitting more queries and clicking on more results. Therefore, the number of queries and clicks in a session serve as indicators for the overall search performance of the session. As the table shows, both the number of total queries and clicks in the session are negatively correlated with ERel and topical relevance judgments, showing that in sessions with more searches and clicks, overall searchers are less likely to acquire useful information from the results.

The regression model included **time prev**, **time post**, **#queries prev**, **#queries post**, **#click prev**, and **#click post** as features.

6.2.9 Summary of Features

As this section described, rich sets of features involving different types of information show certain correlations with searchers' judgments on ERel, topical relevance, and other properties of results. This suggests that these features are probably useful for predicting ERel as well. Table 37 summarizes the features included for prediction in the regression models. Due to the limited size of the collected data and the high similarity between some features, not all features were included into the regression models to avoid overfitting in training.

6.3 EVALUATION

6.3.1 Experiment Settings

This section evaluates the effectiveness of the described approaches in order to answer RQ2.1–2.3 and examine the hypotheses H2.1–H2.5. The experiment compares the effectivenesses of different sets of features in predicting ERel. The experiment generates 10 random partitions of the dataset. On each partition, a 10-fold cross validation is performed to evaluate the effectiveness of an approach (a set of features), where 7 folds are used for training, 2 folds are used for validation, and 1 fold is used for testing. All evaluated approaches use the same 10 partitions and cross-validation settings to make sure that results on each test fold are comparable.

The effectiveness of ERel prediction is evaluated using the following measures:

- Normalized root mean square error (NRMSE) – a measure for the differences between the predicted ERel and the real ERel (smaller values indicate better effectivenesses). The differences (RMSE) are normalized to the interval $[0, 1]$.
- Correlation (Pearson's r and Spearman's ρ) between the predicted ERel and the real ERel (higher values indicate better effectivenesses).

For each approach (feature set), the experiment yields results on 100 test folds in total (10 test folds for each partition, and 10 random partitions). The following sections report

Table 37: A summary of prediction features adopted in the ERel regression model.

	None	**CurrentQ**	**Before**	**After**
textsim		QL, SDM, BM25		
otherQ			PWD PrevQ avg	PWD NextQ avg
qref			PWD PrevQ ADD min, PWD PrevQ RMV min, PWD PrevQ KEEP min	PWD NextQ RMV max, PWD NextQ RMV min, PWD NextQ KEEP min, PWD NextQ KEEP mean
click			PWD click PrevQ avg, PWD click PrevQ max, PWD click PrevQ min	logdwell, PWD click NextQ avg, PWD click NextQ max, PWD click NextQ min
effort	#chars, #words, #sents, #sents w/ query term, ARI, LIX, CLI, %link, %em, %div, %li, %p			
relia	.gov, .edu, .net, .com			
session			time prev, #queries prev, #clicks prev	time post, #queries post, #clicks post

the mean values of the measures on the 100 test folds. Differences between approaches are compared by whether the mean values of the approaches on the 100 test folds are significantly different using paired *t*-test.

Note that results in this section should be interpreted with respect to how these ERel ratings were collected. User Study 1 only collected ERel judgments for clicked results—most of them are top-ranked results returned by current search engines. In fact, all the clicked results have a mean topical relevance rating (TRel) of 1.66 (from 0–3), showing that these results generally have moderate to high topical relevance. Assuming that existing search engines are capable of retrieving results with high topical relevance, the results in this section mostly stand for the effectivenesses of the features in predicting ERel among those with high topical relevance.

6.3.2 Can we predict ERel before searchers are presented the result?

This section evaluates the sets of features that are available before searchers are presented with the result. This includes: text similarity features (textsim); similarity of the result with previous search queries (PrevQ otherQ); query reformulation features based on previous search queries (PrevQ qref); click features in previous searches (PrevQ click); session-level statistics in previous searches (PrevQ session); effort and understandability features (effort); reliability features (relia). Results in this section examine the following research question and hypotheses:

- **RQ2.1** – How well can we predict ERel before users are presented the result?
- **H2.1** – The ERel of a result is related to the characteristics of the result document.
- **H2.2** – We can predict the ERel of a result based on the text similarity between a query and the result.
- **H2.3** – We can predict the ERel of a result based on past search history in the session.

Table 38 reports the results for these feature sets individually and their combination. The combination of these feature sets is representative of the prediction accuracy one can achieve at the time of search result ranking (when users issued the query, but has not been presented with the search results). A better prediction of ERel at this moment can help search engines better rank search results (for example, by predicted ERel).

Results verify the correctness of H2.2—text similarity between search query and results can predict the ERel of a result. As Table 38 shows, text similarity features (textsim) altogether yield a prediction of ERel with Pearson's $r = 0.237$ with the real ERel values. This shows that existing ad hoc search models (text similarity measures) are generally predictive of ERel. However, the correlation values are weak, suggesting that among the clicked results (with moderate to high topical relevance), text similarity has only limited additive values in predicting the ERel of results.

Overall results also verify the correctness of H2.3—users' past search activities in a session are predictive of the ERel of a result. The prediction performance is at least comparable to that using only text similarity features, suggesting the importance of contextual information in a search session in predicting ERel. Among features using previous search activities, those based on previous clicks (PrevQ click) have the best performance (NRMSE 0.327, and $r = 0.241$), which is comparable to the text similarity features (the differences are not significant at 0.05 levels). Similarity of the result with previous search queries (PrevQ otherQ) are less effective, with NRMSE 0.330 and $r = 0.206$ (significantly worse than PrevQ click at 0.05 level). The query reformulation features are also predictive of the results' ERel, but significantly worse than using previous clicks and queries (significantly worse than PrevQ click and PrevQ otherQ at 0.01 level).

In addition, results also verify that H2.1 is correct. As the table shows, the effort and understandability features (effort) based on only the result's own characteristics is the most effective individual set of features among those examined in Table 38. These features yield NRMSE of 0.324, and Pearson's $r = 0.270$, which are significantly better than those for text similarity features (textsim) and click features (PrevQ click). This suggests that the effort and understandability features are very important for predicting ERel among results with high topical relevance. Note that both text similarity features and those using previous queries or clicks are in nature measuring the text similarity between representations of searchers' information needs and the result documents. This further confirms the limited additive value of text similarity in predicting the useful information acquired by the user.

Reliability features alone seem not very predictive of the result's ERel.

Combining different sets of features ("PrevQ ALL + textsim + effort + relia") further improve the effectiveness of prediction to a significantly better level than using any individual feature sets alone. The Pearson's r between the predicted ERel and the real value is 0.320, and the Spearman's ρ is 0.309.

It should be noted that although the prediction accuracy looks overall rather limited, this is a very hard problem—to further predict the usefulness of results among a subset of pre-selected highly relevant results. First, the collected data are from top-ranked results by search engines. Second, the collected results are those clicked by users (indicating their feedback of the results' values assessed based on the outlook of the result snippet). The inclusion of features already takes into account all major ideas in previous work, e.g., using previous queries and clicks as relevance feedback (X. Shen et al., 2005; Bennett et al., 2012), considering detailed query reformulation patterns (D. Guan et al., 2013; Luo et al., 2014), and the very recent work on effort-based judgments and prediction (Yilmaz, Verma, Craswell, Radlinski, & Bailey, 2014; Verma, Yilmaz, & Craswell, 2016). Results suggest that these approaches are generally predict of the ERel of results as well. In addition, the better effectiveness of the effort features indicate that these features are probably those missing by existing search engines.

6.3.3 Can we predict ERel after searchers are presented the result?

This section evaluates the sets of features based on information that are only available after searchers interact with the result. This includes: similarity of the result with follow-up search queries (NextQ otherQ); query reformulation features based on follow-up search queries (NextQ qref); click features in follow-up searches (NextQ click); session-level statistics in follow-up searches (NextQ session). Results in this section examine the following research question and hypotheses:

- **RQ2.2** – How well can we infer ERel from users' interaction with the result and those afterwards in the search session?
- **H2.4** – We can infer the ERel of a result afterwards based on how searchers interact

with the result and the SERP showing the result.

- **H2.5** – We can infer the ERel of a result afterwards based on how searchers behave in follow-up searches in the session.

Table 39 reports the effectivenesses of these features. As the table shows, the click features are the most effective individual set of features, with NRMSE 0.302 and Pearson's $r = 0.443$. This largely outperforms using other feature sets alone, or the combination of features available before searchers are presented the result. This is due to the high correlation between click dwell time and ERel—as previous sections reported, the log dwell time alone has a Pearson's $r = 0.44$ with the ERel of the result (in the whole dataset without using cross validation). This is also not surprising considering click dwell time were generally considered as a strong predictor of result relevance in almost all previous work. Here the experiment confirms that click dwell time is also indicative of the result's ERel. This also verifies the correctness of H2.4.

In addition, results in Table 39 verifies H2.5. Generally searchers' interaction with the search engine in follow-up searches (not directly connected to the clicked search result) are capable of predicting the result's ERel. In contrast to those reported in the previous section, it has been found that the text similarity with queries in previous searches have better prediction performance than those in follow-up searches. But the query reformulation features in follow-up searches better predict ERel than those in previous searches. This suggests the different role of features before and after searchers are presented with the result.

Combining all the features (including the effort and reliability features) can predict ERel by a higher level of accuracy. The predicted values and the real values are moderately correlated ($r = 0.517$). Note that this is achieved without using any explicit feedback among a subset of results that are generally highly relevant. This shows that the proposed approaches are highly effective.

6.3.4 Do explicit and implicit feedbacks help each other in predicting ERel?

As the previous section shows, implicit feedback information in a session after the searchers were presented with the result can predict the ERel of the result reasonably well, with a Pearson's $r = 0.517$ between the predicted ERel and the actual value. This section further examines how well the prediction compares with those using explicit feedbacks.

- **RQ2.3** – How well can we infer ERel based on both implicit feedback information in RQ2.2 and explicit topical relevance judgments?

Table 40 reports the effectiveness of the approaches in previous sections and those using explicit feedback on topical relevance. TRel and Usef refers to the prediction performance using TRel or Usef alone as the predictor. In addition, the implicit feedback-based approach in the last section was combined with users' Usef ratings.

Results show that using implicit feedback alone (RQ2.2 – NextQ All + effort + relia + Usef) is still not comparable to those based on explicit topical relevance ratings in predicting ERel. This is not surprising. However, implicit feedback is almost free for search engine companies, but explicit topical relevance requires human labor.

In addition, results show that combining implicit feedback signals in a search session and the explicit feedback on topical relevance (Usef) can better predict the ERel of results compared with using only explicit feedback alone (Usef). As the table shows, "RQ2.3 – NextQ All + effort + relia + Usef" produces prediction values with Pearson's $r = 0.784$ with the actual value and NRMSE 0.209. Both are significant better than those using Usef alone as the predictor. This further confirms the value of implicit feedback in a session in ERel prediction.

6.4 ANSWERS TO RQ2

To sum up, results in this section answered RQ2 and verified the correctness of the five hypotheses. Overall results show that ERel of a result can be predicted based on: 1) the result's own characteristics (especially its readability); 2) the text similarity between search queries and the result; 3) past search activities in a search session (such as previous search queries and clicks); 4) searchers' interaction with the result, especially click dwell time; and 5) searchers' interaction in follow-up searches in the same session.

Among the five different types of information, 2, 3, and 4 were widely used in previous

Table 38: Evaluation of ERel prediction performance using different sets of features that are available before searchers are presented with the result (RQ2.1).

Features	**NRMSE (lower is better)**	**Pred. r (higher is better)**	**Pred. ρ (higher is better)**
textsim	0.327	0.237	0.194
PrevQ otherQ	0.330	0.206	0.143
PrevQ qref	0.332	0.164	0.104
PrevQ click	0.327	0.241	0.200
PrevQ session	0.336	0.100	0.069
effort	0.324	0.270	0.240
relia	0.337	0.011	0.008
PrevQ ALL + textsim + effort + relia	**0.319**	**0.320**	**0.309**

- Reported values are the mean values on the 100 test folds.
- NRMSE: normalized root mean square error (lower values are better); Pred. r and ρ: the Pearson's r and Spearman's ρ between the groundtruth values and the predicted ones on the test folds (higher values are better).
- All differences with "PrevQ ALL + textsim + effort + relia" are significant at 0.001 level by paired t-test.
- textsim: text similarity features; PrevQ otherQ: similarity with previous queries; PrevQ qref: query reformulation features using previous queries; PrevQ click: click features in previous searches; PrevQ session: session statistics in previous searches; effort: effort and understandability features; relia: reliability features.

Table 39: Evaluation of ERel prediction performance using different sets of features that are available after searchers are presented with the result (RQ2.2).

Features	**NRMSE (lower is better)**	**Pred. r (higher is better)**	**Pred. ρ (higher is better)**
NextQ otherQ	0.332	0.181	0.144
NextQ qref	0.329	0.219	0.130
NextQ click	0.302	0.443	0.369
NextQ session	0.333	0.142	0.117
effort	0.324	0.270	0.240
relia	0.337	0.011	0.008
NextQ ALL + effort + relia	**0.289**	**0.517**	**0.457**

- Reported values are the mean values on the 100 test folds.
- NRMSE: normalized root mean square error (lower values are better); Pred. r and ρ: the Pearson's r and Spearman's ρ between the groundtruth values and the predicted ones on the test folds (higher values are better).
- All differences with "NextQ ALL + effort + relia" are significant at 0.001 level by paired t-test.
- NextQ otherQ: similarity with follow-up queries; NextQ qref: query reformulation features using follow-up queries; NextQ click: click features in follow-up searches; NextQ session: session statistics in follow-up searches; effort: effort and understandability features; relia: reliability features.

Table 40: Comparison of ERel prediction performance with predictions using explicit topical relevance judgments (RQ2.3).

Features	**NRMSE (lower is better)**	**Pred. r (higher is better)**	**Pred. ρ (higher is better)**
RQ2.1 – PrevQ ALL + effort + relia	0.319 ***	0.320 ***	0.309 ***
RQ2.2 – NextQ All + effort + relia	0.289 ***	0.517 ***	0.457 ***
RQ2.3 – NextQ All + effort + relia + USEF	**0.209 *** **	**0.784 ** **	**0.749 *** **
TREL	0.252 ***	0.662 ***	0.563 ***
USEF	0.222	0.752	0.685

- *, **, and *** indicate the value is significantly different from USEF at 0.05, 0.01, and 0.001 levels. All differences with “NextQ All + USEF” are significant at least at 0.01 level.
- NRMSE: normalized root mean square error (lower values are better); Pred. r and ρ: the Pearson’s r and Spearman’s ρ between the groundtruth values and the predicted ones on the test folds (higher values are better).

work for predicting or inferring topical relevance of results. Here this study confirms that they are effective predictors for ERel as well. Using 1 is not common in predicting topical relevance (probably due to the fact that topical relevance judgments are affected by understandability of result by a smaller extent), despite a few recent work on predicting effort (Yilmaz et al., 2014; Verma et al., 2016). But here the results show that 1 is a very useful indicator for result's ERel. As the last chapter examined, this is probably due to the complex relation between ERel, effort, and the understandability and reliability of results. In addition, using 5 is rare, to the best of my knowledge, in the field of information retrieval. This study shows that searchers' follow-up interaction in a session (such as later query reformulations and clicks) are also useful predictors of ERel for results in previous searches.

Overall these five different types of information can be applied to predict ERel of results in two different scenarios, solving both issues in search result ranking and search engine evaluation and optimization. Results also show that the prediction of ERel based on only implicit feedback information is very effective, with Pearson's $r > 0.5$ between the predicted and observed values. This can provide large-scale and free training data (though relatively low quality compared with explicit feedback) for optimizing search engines.

7.0 QUERY REFORMULATION IN A SEARCH SESSION

This chapter examines and predicts searchers' query reformulation behavior in a search session. Here query reformulation behavior is decomposed into users' choices of word changes in query reformulation, such as whether to remove or to retain a word in the next query, and whether or not to add a word into the next query. This chapter proposes seven hypotheses related to RQ3:

- **RQ3** – What affects users' choices of word changes in a query reformulation during a search session? Can we predict such choices of word changes?
- [**H3.1**] – Searchers' choices about a word in a query reformulation are related to the word itself.
- [**H3.2**] – Searchers' choices about a word in a query reformulation are related to the connection between the word and search queries in a session.
- [**H3.3**] – Searchers' choices about a word in a query reformulation are related to the connection between the word and retrieved results in a session.
- [**H3.4**] – Searchers' choices of word changes in a query reformulation are related to the situation in a search session where the query reformulation happens.
- [**H3.5**] – Searchers' choices of word changes in a session are related to the task's goal and product.
- [**H3.6**] – Searchers' choices of word changes in a session are related to the searcher.

Firstly, § 7.1 describes an analysis framework for query reformulation behavior by decomposing query reformulations into searchers' choices of word changes. Secondly, § 7.2 and 7.3 separately analyze the relation of different factors on two different types of word changes (removing or retaining a word, and adding a word or not). Further, § 7.4 and 7.5 describe

Table 41: An example of query reformulation patterns by Rieh & Xie (2006).

Pattern	**Meaning**	**Example**
Specification	adding words or using more specific words	us primaries → us primaries 2016
Generalization	removing words or using more general words	linear regression matlab → linear regression
Replacement with Synonyms	replacing a word with another with a similar meaning	statistics software → statistics package
Parallel Movement	moving from a subtopic to another	depression symptoms → depression definition

and evaluate approaches for predicting word changes in query reformulation. Finally, § 7.6 concludes an answer to RQ3 and discusses the hypotheses.

7.1 WORD CHANGES IN QUERY REFORMULATION

7.1.1 Decomposing Query Reformulation

Query reformulation refers to the transition from a search query to the next one in a search session. The differences of the two queries may indicate certain intentions of the users. Previous studies (Bruza & Dennis, 1997; Anick, 2003; Rieh & Xie, 2006; Huang & Efthimiadis, 2009; Jansen, Booth, & Spink, 2009) categorized query reformulations into different "types" or "patterns". Table 41 shows an example of the patterns summarized by Rieh and Xie (2006). These patterns help understand how query reformulations look like and what the possible intentions are, but they have several limitations:

- First, no evidence shows that searchers think and make decisions with regard to these "types" or "patterns" when they reformulate queries. For example, searchers do not necessarily choose one pattern out of others when they reformulate queries.
- Second, the query reformulation types are not specific—they do not look into individual words.

In contrast, this study focuses on individual and specific word changes in query reformulation. Let Q_i and Q_{i+1} be the sets of words included in two consecutive queries q_i and q_{i+1} in a search session. All words included into this query reformulation ($q_i \rightarrow q_{i+1}$) can be divided into three types:

- Retained words (words included in both queries) — $Q_i \cap Q_{i+1}$.
- Removed words (words included in q_i but not in q_{i+1}) — $Q_i - Q_{i+1}$.
- Added words (words included in q_{i+1} but not in q_i) — $Q_{i+1} - Q_i$.

When discussing $q_i \rightarrow q_{i+1}$, following discussions also call q_i *the current query*, and call q_{i+1} *the next query* or *the new query*. These three sets of words stand for finer-grained decisions in query reformulation compared with the query-level reformulation patterns. Words in the current query (q_i) are either retained or removed in the next one (q_{i+1}). Words that are not included in the current query (q_i) are either added to the next query (q_{i+1}) or not. This chapter decomposes a user's decisions in a query reformulation into the following two types of events:

- For each word in the current query, to decide whether to retain or to remove the word in the next query.
- For each word from *a candidate set*, to decide whether or not to add the word to the next query.

Here the candidate set refers to the set of words considered by the users for whether or not to add to the next query. The candidate set does not have overlap with the current query. Of course, it is difficult (if possible) to determine the actual set of words considered by the users. Thus this study uses some heuristics to determine the candidate set. Section 7.3.2 further described details.

Table 42 shows an example to better illustrate the differences between the query reformulation patterns in previous studies and the word changes examined in this chapter. The example shows four search queries in a search session. For each query reformulation, the table also shows the corresponding patterns (by Rieh and Xie's (2006) definition) and word changes.

Table 42: An example of query reformulation patterns and word changes.

	Query	Reformulation Pattern	Word Changes
1	depression symptoms	-	-
2	depression definition	Parallel Movement	retain: depression remove: symptoms add: definition
3	depression treatment	Parallel Movement	retain: depression remove: definition add: treatment
4	depression treatment cost	Specification	retain: depression retain: treatment add: cost

As the example shows, this study examines specific word changes. For example, from the first query *depression symptoms* to the second one *depression definition*, the examined word changes include: retaining *depression*, removing *symptoms*, and adding *definition*. These word changes indicate the possible decisions of the user in this query reformulation—whether to remove or to retain *depression* (and the user decided to retain), whether to remove or to retain *symptoms* (and the user decided to remove), and whether or not to add *definition* (and the user decided to add it to the next query).

7.1.2 Dependent Variables

Let Q_i and Q_{i+1} be the sets of words included in two consecutive queries q_i and q_{i+1}. Formally, this study separately examines the following two dependent variables:

- **[RMV]** – For each word $w \in Q_i$, $\text{RMV}(w, q_i, q_{i+1}) = 1$ if the user removed w in $q_i \rightarrow q_{i+1}$. Otherwise $\text{RMV}(w, q_i, q_{i+1}) = 0$ (the user retained the word).
- **[ADD]** – For each word $w \in C$ (a candidate set), $\text{ADD}(w, q_i, q_{i+1}) = 1$ if the user added the word to q_{i+1}. Otherwise $\text{ADD}(w, q_i, q_{i+1}) = 0$ (the user did not add the word to q_{i+1}).

The two dependent variables (RMV and ADD) cover all *vocabulary* differences between the two queries. Other differences between the two queries are not examined in this study, either because they are less important (for example, changes in the sequence of words, e.g.,

Table 43: Session-level independent variables for analyzing word changes in query reformulation.

Hypothesis	**Variable**	**Explanation**
H3.5	`goal`	0 if the goal of the search task is clear, or 1 if it is amorphous (Y. Li & Belkin, 2008).
H3.5	`product`	0 if the task looks for factual information, or 1 if for intellectual understanding (Y. Li & Belkin, 2008).
H3.6	`familiarity`	User's self-rated familiarity regarding the topic of the task using a five-point likert scale.
H3.6	`avg_num_rmv`	Average number of removed words in other sessions by the same user.
H3.6	`avg_num_add`	Average number of added words in other sessions by the same user.

from *statistics R* to *R statistics*) or too rare in the collected search log (for example, multiple occurrences of the same word in a query, which was only observed in 2 out of 388 queries in the collected data).

7.1.3 Independent Variables

A major purpose of this chapter is to analyze the influence of different factors on word changes, as characterized by the dependent variables. Table 43, 44, and 45 list all independent variables examined in this study. These independent variables are divided into three groups depending on their relations to the dependent variables:

- Session-level variables are measured at the session-level as a whole. Each examined word within a search session shares the same influence from these variables.
- Query-level variables are measured at the query-level. Each examined word with regard to the same query reformulation shares the same influence from these variables.
- Word-level variables are specific to each examined word.

7.1.3.1 Session-level Independent Variables *Session-level variables* measure factors related to the characteristics of the task and the searcher. Within a search session, every word

Table 44: Query-level variables for analyzing word changes in query reformulation.

Hypothesis	**Variable**	**Explanation**
H3.4	`q_length`	Length of the current query (excluding stop words).
H3.4	`q_duration`	Time duration from the submission of the current query to that of the new query.
H3.4	`q_clickpos`	Position of the lowest ranked clicked results on the SERP, or 0 if no click.
H3.4	`num_query`	Number of submitted queries in the session (including the current query).
H3.4	`num_click`	Number of past clicks in the session (including clicks on the current query's SERP).
H3.4	`duration`	Time duration from the beginning of the session to the submission of the new query.

change in each query reformulation shares the same influence from session-level variables (because they are from the same session performed by the same searcher). Table 43 shows the session-level variables, their meanings, and related hypotheses.

User Study 2 followed Li and Belkin's faceted task classification framework (Y. Li & Belkin, 2008) and considered three characteristics of search tasks. `goal` is a variable for the goal of a search task, which is either specific (0) or amorphous (1). `product` is variable for the targeted product of a search task, which is either factual (0) or intellectual (1). `familiarity` is a variable for users' self-rated familiarity on the topic of the task using a five-point Likert scale.

In addition, this study suspects that users' choices of word changes in a query reformulation may also depend on individual preference. For example, some searchers may prefer to add or remove words in general. In order to capture this factor, the average number of added and removed words during a query reformulation in other sessions performed by the same searcher are adopted as surrogates for the users' personal preferences for adding (`avg_num_add`) or removing (`avg_num_rmv`) words in query reformulation.

7.1.3.2 Query-level Variables *Query-level variables* measure factors related to past user activities in a session. These factors apply to a query reformulation as a whole. Each

Table 45: Word-level independent variables for analyzing word changes in query reformulation.

Hypothesis	**Variable**	**Explanation**
H3.1	`idf`	IDF of the word in the ClueWeb09 collection.
H3.2	`p(w|pastq)`	Probability of the word appearing in past queries of the session.
H3.2	`avg_jaccard`	Average Jaccard similarity of the word with other words in the current query.
H3.3	`#click_hasw`	Number of clicked results whose title/snippet/URL contains the word.
H3.3	`#skip_hasw`	Number of skipped results whose title/snippet/URL contains the word.
H3.3	`freq_w_suggest`	Frequency of the word in query suggestions if users viewed the area, or 0 otherwise.

word change in a query reformulation shares the same influence from the query-level variables (because they are related to the same query reformulation event). Table 44 shows the query-level variables, their meanings, and related hypotheses.

This study suspects users' choices of word changes in a query reformulation is directly influenced by the most recent search. Thus Table 44 included variables for the characteristics of the current query, such as query length (`q_length`), and user activities on its SERP, including the time spent on the SERP (`q_duration`) and the lowest rank of the clicked results (`q_clickpos`). In addition, Table 44 included variables indicating the time of a session when the reformulation happened, including the number of previous queries submitted (`num_query`), the number of previous clicks (`num_click`), and the total time spent in the session starting from the submission of the current query (`duration`).

7.1.3.3 Word-level Variables

Word-level variables measure factors directly related to the word being examined. Different words in the same query reformulation can be affected differently by these variables. Table 45 shows the word-level variables, their meanings, and related hypotheses.

`idf` indicates word specificity (Spärck Jones, 1972), i.e., to what extent the word is general or specific. `p(w|pastq)` is the probability that the word w was included in past queries of the session (from the current query to the current query). `p(w|pastq)` indicates the centrality of the word w to the task. The following table shows an example, where the task is to find information on the symptoms and treatments of depression. The word *depression* is included in every query and expresses the main theme of the task, which is unlikely to be removed in query reformulation.

1	**depression** symptoms
2	**depression** definition
3	**depression** treatment
4	**depression** treatment cost

`avg_jaccard` measures the connection between the word being examined and other words in the current query using Jaccard similarity based on their co-occurrences in documents. Here the Jaccard similarity of two words is defined as the Jaccard similarity between the sets of documents containing the two words in the Clueweb09 dataset. When examining a word w with regard to a query q, `avg_jaccard` is the mean value of the Jaccard similarity between w and each word in q that is not w. If there is no other words than w in the query, `avg_jaccard` is set to the mean value of `avg_jaccard` in other query reformulations.

In addition, this study also hypothesizes that users' choices in word changes are related to the occurrences of the examined word in results displayed on the current query's SERP. This study separately considers clicked results (`#click_hasw`) and skipped results (`#skip_hasw`). Here *skipped results* refer to those that users viewed their summaries but did not click on. This study relies on eye movement to determine whether or not searchers viewed a result, based on whether or not an eye fixation was observed. In addition, the occurrences of the word in different elements of result summaries (including titles, snippets, and URLs) are examined separately as different variables.

Moreover, `freq_w_suggest` measures the occurrences of words in query suggestions displayed on the current query's SERP. The screen area displaying query suggestions are considered as a whole (due to the limited accuracy of the eye-tracking device in tracking small items). If the SERP provides query suggestions and the study observed the user's eye fixations on that area, `freq_w_suggest`'s value is set to the frequency of the word in all query suggestions. Otherwise, `freq_w_suggest` is set to 0.

7.1.4 Analysis Approach

The following sections examine the influence of the independent variables on the dependent variables using hierarchical (multilevel) logistic regression, a technique that models variables with more than one variance component (Gelman & Hill, 2006). More specifically, it deals with a regression model with binary outcome (dependent variable), and varying coefficients of independent variables.

This study adopts hierarchical logistic regression in analysis instead of vanilla logistic regression because the observations in this study are nested—the analysis examines multiple word changes nested within a query reformulation, and there can be multiple query reformulations nested within a session as well. In such case, the observations are not independent of each other, because some of them share the same contexts at the query and/or session levels. This violates the independence assumption to apply vanilla logistic regression. In contrast, hierarchical models can handle such issues. The analysis conducted in the following sections uses a hierarchical model with three levels to study word changes (level 1) during a query reformulation (level 2) in a search session (level 3).

However, it should be noted that the analysis approach does not consider another dependency issue in word changes—the decision on one word may relate to those on other words. This is a limitation of the analysis approach.

7.2 REMOVING OR RETAINING A WORD

This section reports results for whether to remove or retain a word in query reformulation (RMV). When studying a reformulation from an old query to a new query, the analysis examines each word in the old query as an instance. This is to assume that users consider each word in the current query and make decisions on whether to remove or retain the word in the next query. This yields 687 observations from 203 query reformulations, where 304 (44%) are positive (RMV = 1).

Table 46 reports results for the hierarchical logistic regression analysis, where RMV is

Table 46: Hierarchical logistic regressions: RMV (removing a word instead of retaining it in the next query) as dependent variable.

Step	Variable Name	Model 1	Model 2	Model 3		
		exp(B)	exp(B)	exp(B)	95% CI	
1	`constant`	0.236	0.039	0.068	-	-
	`product (I)`	0.974	0.682	0.825	0.545	1.250
	`goal (A)`	1.001	0.926	0.929	0.638	1.353
	`familiarity`	1.018	1.064	1.085	0.917	1.283
	`avg_num_rmv`	2.123	1.626	1.559	1.063	2.285
2	`q_length (log)`		2.067	1.675	1.117	2.511
	`q_duration (log)`		1.334	1.570	1.236	1.993
	`q_clickpos`		1.042	1.130	1.038	1.230
	`num_query`		1.213	1.165	1.077	1.261
	`num_click`		0.953	0.959	0.897	1.026
	`duration`		0.905	0.885	0.804	0.974
3	`idf`			0.928	0.845	1.019
	`p(w|pastq)`			0.216	0.123	0.378
	`avg_jaccard (log)`			0.851	0.749	0.966
	`#click_title_hasw`			0.734	0.590	0.912
	`#click_snippet_hasw`			0.946	0.767	1.165
	`#skip_title_hasw`			0.921	0.797	1.065
	`#skip_snippet_hasw`			0.916	0.813	1.032
	`freq_w_suggest`			0.941	0.844	1.049
	-2 LL (null model: 943.3)	917.3	877.1	802.5		
	Omnibus Tests of Model Coefficients	***	***	***		

- Light, dark, and darker shadings indicate the coefficients are significant at 0.05, 0.01, and 0.001 levels, respectively. Unshaded cells are not significant at 0.05 level.
- *, **, and *** indicate the model is statistical significant from its baseline model at 0.05, 0.01, and 0.001 levels, respectively. Model 1 uses the null model as the baseline (with only a constant term). Model 2 uses Model 1 as the baseline, and Model 3 uses Model 2 as the baseline.

the dependent variable. exp(B) $>$ 1 suggests a positive influence of the independent variable on RMV, and exp(B) $<$ 1 indicates a negative one. Model 1 includes only the session-level variables and constant. Model 2 further includes the query-level variables. Model 3 includes all variables.

The values of some variables are transformed by taking natural log values to make them linear, including: `q_length`, `q_duration`, and `avg_jaccard`. This study examines multicollinearity issues using the variance inflation factor (VIF). A value greater than 5 or 10 usually suggests that the model may have multicollinearity issues (Menard, 2002). After examining variables' VIF, `avg_num_add` was removed due to its high correlation with `avg_num_rmv` (Pearson's $r = 0.85$). Similarly, `#click_url_hasw` is removed because of its high correlation with `#click_title_hasw` (Pearson's $r = 0.86$), and `#skip_url_hasw` is removed for its correlation with `#skip_title_hasw` (Pearson's $r = 0.83$). All the variables included in Table 46 satisfy VIF < 5.

7.2.1 Influence of the Session-level Variables

The examined session-level variables show some influence on whether to remove or retain a word in query reformulation. Model 1 explains the collected data significantly better than a baseline model including only constant ($p < 0.001$ in Omnibus test). However, the magnitude of change in -2 log likelihood is small, indicating only a mild influence.

Among these variables, only users' average number of removed words in other sessions (`avg_num_rmv`) consistently shows a significant positive effect in all three models, indicating that users are more likely to remove a word in a session if they removed words more frequently in other sessions. This suggests that a user's overall preference to remove a word may affect their decisions on removing or retaining a word in a specific session. Some users may prefer to remove words in query reformulations in general, and they are likely to do so in a specific session as well. However, here it remains unclear whether such preference is related to other factors (e.g., search expertise) due to the limited information collected in the User Study.

Task product (`product`), goal (`goal`), and topic familiarity (`familiarity`) have no significant influence on the dependent variable in any of the three models. This indicates that the examined task characteristics may not directly affect removing or retaining a word.

7.2.2 Influence of the Query-level Variables

The examined query-level variables also show some influence on RMV. After including the query-level variables, Model 2 significantly improves over Model 1 ($p < 0.001$). Over half of the query-level variables show significant influence in both Model 2 and Model 3. This suggests that removing or retaining a word were affected by these variables.

The length of the current query (`q_length`) shows a significant positive influence—users removed words more frequently in relatively longer queries. This is not surprising considering that longer queries are more likely to include words that are unnecessary for the task.

Results also suggest that users are more likely to remove a word if they spent a relatively longer time on the current query's SERP (`q_duration` shows a significant positive influence) and clicked on results on the current query's SERP (`q_clickpos` shows a significant positive influence). This indicates a situation that the current query retrieved relevant information (such that users were willing to spend time on examining the results, instead of quickly reformulating to the next query without clicking). A further examination of the collected data found that this is probably because the majority of the removal happened when users had (successfully) finished exploring a facet of the task (a subtask) and switched to another subtask (or the main task). Section 7.2.4 discusses more details.

In addition, results suggest that users are more likely to remove a word if they submitted many queries in a session (a significant positive influence of `num_query`), but they are less likely to do so with the time goes by in a session (a significant negative influence of `duration`). The two trends are seemingly conflicting with each other, since the time spent in a session (`duration`), unsurprisingly, has a moderate correlation with the number of issued queries (`num_query`) ($r = 0.52$). But the two variables may also stand for two different factors.

Submitting more queries does not necessarily mean a better search progress, because querying and SERP examination themselves provide limited relevant information. Submitting many queries may even indicate limited search performance, e.g., previous studies reported that users will compensate limited search performance by searching more frequently

(Smith & Kantor, 2008; Azzopardi, 2014). The number of clicks (num_click), in contrast, may better indicate the amount of relevant information acquired in a search session, even though not all clicked results are relevant. In the collected data, duration better correlates with num_click ($r = 0.61$), while num_query only slightly correlates with num_click ($r = 0.33$).

A possible interpretation is that the results for num_query, num_click, and duration in Table 46 indicate two different factors for whether to remove or retain a word in query reformulation. On the one hand, users are more likely to remove a word when the session has limited search effectiveness, which is supported by the significant positive influence of num_query and its connection with limited search performance reported in previous studies (Smith & Kantor, 2008; Azzopardi, 2014). On the other hand, users are less likely to remove a word after acquiring more relevant information, as suggested by the negative influence of duration and num_click. Note that in Table 46, num_click shows no significant influence on removing or retaining, but this is in fact influenced by the relatively high correlation between num_click and duration ($r = 0.61$). After excluding duration, num_click shows a significant negative effect as well in the reduced model (exp(B) = 0.917, $p = 0.005$).

7.2.3 Influence of the Word-level Variables

The examined word-level variables show a strong influence on whether to remove or retain a word in query reformulation. After including the word-level variables, Model 3 significantly improves over Model 2 ($p < 0.001$). The -2 log likelihood reduces by 74.6 compared with Model 2. The magnitude of change is greater than that for the combination of the session and query-level variables compared with the null model (66.2). This suggests the examined word-level variables are more salient factors than session and query-level variables for removing or retaining a word in query reformulation—this is also not surprising because these word-level variables are specific to the examined words, while other variables are not.

The frequency of using a word in past queries of the same session (p(w|pastq)) shows a significant negative influence on removing the word—users are less likely to remove a word that appeared frequently in previous queries. As discussed in Section 7.1.3.3, this is probably because words repeatedly used in a session represent the main theme of the task, which is less likely to be excluded in queries.

Results also suggest that users are more likely to remove a word that does not co-occur frequently with other words in the current query (`avg_jaccard` shows a significant negative influence on RMV). This usually happens when the examined word is off-topic, overspecific, or misspelled. In these cases, the examined word co-occurs with other query words only in a limited number of (if any) documents, causing a low value of `avg_jaccard`.

In addition, results for `#click_hasw` and `#skip_hasw` suggest that searchers are more likely to remove a word in query reformulation if the word appeared less often in the current query's results. Although only `#click_title_hasw` shows a significant negative effect in Table 46, this is affected by the relatively strong correlations among these variables ($r = 0.77$ between `#click_title_hasw` and `#click_snippet_hasw`, and r=0.68 between `#skip_title_hasw` and `#skip_snippet_hasw`). After removing `#click_title_hasw` and `#skip_title_hasw` from Model 3, both `#click_snippet_hasw` (exp(B) = 0.798, p = 0.010) and `#skip_snippet_hasw` (exp(B) = 0.876, p = 0.003) have significant negative influence. Similarly, `#skip_title_hasw` will show a significant negative influence (exp(B) = 0.861, p = 0.006) if `#skip_snippet_hasw` was removed from the model.

Results in Table 47 further confirms this finding. Table 47 compares occurrences of the removed and retained words in different result elements. In addition to the clicked and skipped results, the table also examined *unclicked* results—any results displayed on the SERP that users did not click on, regardless of whether or not they viewed the results. Table 47 clearly suggests that removed words appear significantly less often in the current query's results compared with the retained words. This applies to all clicked, skipped, and unclicked results, and is consistent among different elements of the results. Therefore, this section does not hope to over-emphasize the significant influence of word occurrences in clicked result titles (`#click_title_hasw`), as showed in Table 47. It seems in general users are more likely to remove words that appeared less often in the retrieved results. This may happen when the word is ineffective (cannot retrieve any results containing the word when combining with other query words), or when the word is not central to the main theme of the task.

Table 47: Mean values (with standard errors) of variables for removed and retained words.

Variable	**Removed**	**Retained**	
	mean (S.E.)	mean (S.E.)	
`#click_title_hasw`	0.63 (0.06)	1.03 (0.08)	***
`#click_snippet_hasw`	0.97 (0.08)	1.27 (0.09)	**
`#click_url_hasw`	0.46 (0.05)	0.78 (0.06)	***
`#skip_title_hasw`	1.17 (0.09)	1.75 (0.09)	***
`#skip_snippet_hasw`	2.02 (0.11)	2.49 (0.11)	**
`#skip_url_hasw`	0.82 (0.07)	1.30 (0.08)	***
`#unclick_title_hasw`	2.69 (0.16)	3.85 (0.15)	***
`#unclick_snippet_hasw`	4.25 (0.15)	5.15 (0.13)	***
`#unclick_url_hasw`	2.03 (0.13)	2.91 (0.13)	***
`freq_w_suggest`	0.35 (0.08)	0.64 (0.10)	*

- *, **, and *** indicate the variable's values between removed and retained words are significantly different at 0.05, 0.01, and 0.001 levels, respectively, by a two-tail Welch's *t*-test.

7.2.4 Qualitative Analysis

To better interpret results in Table 46, two annotators (one of them is the author of this dissertation) manually examined all 304 cases of word removal and divided them into two groups: removing (probably) with satisfaction (SAT remove), and removing (probably) due to dissatisfaction (DSAT remove). The Cohen's κ is 0.72. They further discussed the disagreements in annotation and came into a final decision. 199 (65%) removed words are classified as SAT remove, and 105 (35%) are DSAT remove. SAT remove is more frequent than DSAT remove in the collected data.

SAT remove stands for the cases that a word has successfully served its purpose in a query and is removed afterwards. It usually happens when users finished exploring one facet of the task (a subtopic or a subtask) and was about to switch to another subtopic or the main task. Table 48 shows an example. The table reports queries, the number of clicks, and the time spent on each query in a search session collected in User Study 2. Removed words were highlighted. Removed words in the third, fourth, fifth, and sixth queries were labeled SAT removes. These words all indicate different subtopics related to sunspot, the main theme of the task. The user clicked on some results and spent relatively longer time on each of these queries' SERPs, indicating they might have acquired some useful information from the results, and the removed words might have successfully served their purposes.

In contrast, DSAT remove stands for the cases that a word did not fulfill the searcher's purpose of using it in the query. The query usually retrieves low quality results, such that the user did not click on any results and spent only a short duration on the SERP. As the table shows, in the second query, *new* and *phenomena* were labeled as DSAT removes. The searcher showed a similar intent in the second and the third queries – to find out when sunspots were first observed – but did not click on any results for the second query. A further examination on the second query's SERP found that the retrieved results seem not useful—none included both *new* and *phenomena* in the summaries. All the web pages had also been re-examined and found little useful information. Therefore, the user rephrased the query and removed the two words in the next query.

The greater popularity of SAT remove in the collected data explains why results in Table

Table 48: An example of SAT remove and DSAT remove.

	Query	**Label**	**#click**	**time**
1	what are sunspots	-	1	93s
2	are sunspots a **new phenomena**	DSAT	0	14s
3	when were sunspots **first observed**	SAT	2	103s
4	are sunspots **random** or **patterned**	SAT	2	150s
5	sunspots and **earths climate**	SAT	1	142s
6	sunspots **11 year cycle**	SAT	1	94s
7	sunspots magnetic fields	-	1	-

46 suggest that users tend to remove a word when they spent a relatively longer time on the current query's SERP (`q_duration`) and clicked on its results (`q_clickpos`). However, the annotators did also label a substantial number of DSAT removes (35%). This was concealed by the positive influence of `q_duration` and `q_clickpos`.

7.2.5 Summary of Findings

To summarize, results in this section disclose that users remove a word mainly for two reasons and depend on different factors.

Firstly, the word has already fulfilled its purpose and becomes less necessary (SAT remove)—users need to remove the word in query reformulation in order to move forward in the session. This is also supported by the observed significant influence of a few variables reported in Table 46: users spent a longer time on the current query (`q_duration`) and clicked on results on the current query's SERP (`q_clickpos`) before they remove a word; removed words appeared less frequently, but still quite often, in the retrieved results (`#skip_snippet_hasw`; searchers are less likely to remove a word if it is related to the main theme of the task (`p(w|pastq)`).

Secondly, the word caused limited search performance in the current query (DSAT remove). This conflicts with the positive effects of `q_duration` and `q_clickpos`, but is supported by the qualitative analysis. DSAT remove is also consistent with the observed influence of `avg_jaccard_log` and `#skip_snippet_hasw`.

Moreover, removing a word seems also relate to individual preference in query reformulation (`avg_num_rmv`), query length (`q_length`), overall search performance of the session (`num_query`), and the amount of acquired relevant information in the session (`duration` and `num_click`).

7.3 ADDING A WORD OR NOT

7.3.1 Adding a Brand-new Word and Reusing

Section 7.1.2 defined the dependent variable for adding a word or not (ADD). However, during analysis it has been observed that adding a word or not can be further divided into two cases, with probably very different intentions of the searchers.

The first case is to add a brand-new word—a word that was not included in any previous queries of the search session. Following sections refer to this case as *adding a new word* (NEW). The second case is to reuse a word—a word that was included in some previous queries, but had been removed in the current query. Following sections refer to this case as *reusing a word* (REUSE). The rest of this section separately considers these two types of added words because they may stand for different intentions of users. Any added words belong to either of the two types.

7.3.2 Candidate Sets

As Section 7.1.2 discussed, it is difficult (if possible) to determine the actual sets of words searchers considered in their minds for whether or not to add to the new query. Thus this study relies on some heuristics to determine the candidate sets (the sets of words searchers considered for whether or not to add to the new query).

The candidate set for REUSE includes all words in previous queries excluding those in the current query. This is to assume that during a query reformulation, users reconsider each previously used word that was excluded in the current query, and decide whether or not to reuse it in the next query. For each session, this rule was applied to extract the candidate sets of REUSE starting from the reformulation from the second to the third query (by definition, it requires at least two previous queries to examine reusing). On average the candidate set for REUSE includes 4.4 words. This yields 668 instances of REUSE among 151 query reformulations, where 53 (7.9%) are positive (REUSE = 1).

The candidate set for NEW includes words from three sources: task description (showing on the top area of the search interface), result summaries (both titles and snippets) viewed by the user, and query suggestions viewed by the user. Here the considered result summaries and query suggestions are only restricted to those displayed on the current query's SERP (otherwise the size of the candidate set goes very large). For the same result, this study also does not consider words in the clicked web pages. By definition, the candidate set for NEW excludes words included in previous queries. This is to assume that in a query reformulation, users consider whether or not to add new words related to the task (in task description) and those they recently viewed in result summaries and query suggestions (if any) displayed on the SERP. This rule was applied to extract candidate sets for each query reformulation. On average the candidate set for an examined instance has 71.0 words. In total, 14,413 instances from 203 query reformulations were extracted, where 152 (1.05%) are positive (NEW = 1).

As Table 49 shows, the candidate set (all three sources) covers about half (54.9%) of the observed added new words. Task description and viewed result summaries are the major two sources of added new words. In contrast, query suggestions viewed by the users include only 1.3% of the added new words. This study restricts its scope to this candidate set when examining NEW.

7.3.3 Models

Separate hierarchical logistic regression models were estimated for NEW and REUSE. Table 50 and 51 report the results. Similarly, Model 1 includes only the session-level variables and

Table 49: Percentage of new words in query reformulations found in different sources.

Source	**Percentage**
Task description	43.3%
Result summaries viewed by the user	26.7%
Query suggestions viewed by the user	1.3%
All three sources	54.9%
Result titles viewed by the user	15.2%
Result snippets viewed by the user	23.5%

constant, Model 2 further includes the query-level variables, and Model 3 uses all variables.

For NEW, `p(w|pastq)` was excluded because by definition, all words in the candidate sets for NEW should not appear in previous queries. For both NEW and REUSE, `avg_num_rmv` was excluded to avoid serious multicollinearity issues (VIF ≥ 5), because it has a high correlation with `avg_num_add` ($r = 0.85$ for NEW and $r = 0.91$ for REUSE). Similarly, `click_url_hasw` ($r = 0.71$ with `click_title_hasw`) and `skip_url_hasw` ($r = 0.87$ with `skip_title_hasw`) were excluded in REUSE's model. The variables included in the reported models all satisfy VIF < 5.

7.3.4 Influence of the Session-level Variables

Results in Table 50 and 51 suggest that session-level variables have no significant influence on adding a word or not in query reformulation. This is consistent for both adding a new word (NEW) and reusing a word (REUSE). None of the session-level variables show any significant influence on on the two dependent variables in Model 1, Model 2, or Model 3. In addition, for both dependent variables, Model 1 cannot explain the collected data significantly better than baseline models involving only constant ($p = 0.073$ for NEW and $p = 0.122$ for REUSE). This indicates that the included task and user characteristics may

Table 50: Hierarchical logistic regressions: NEW (whether or not to add a brand-new word to the next query) as dependent variable.

Step	**Variable Name**	**Model 1**	**Model 2**	**Model 3**		
		exp(B)	**exp(B)**	**exp(B)**	**95% CI**	
1	`constant`	0.007	0.009	0.048	-	-
	`product (I)`	0.791	0.817	0.916	0.618	1.357
	`goal (A)`	0.836	0.846	0.801	0.557	1.153
	`familiarity`	1.052	1.051	1.059	0.909	1.233
	`avg_num_add`	1.340	1.352	1.317	0.899	1.930
2	`q_length (log)`		0.751	0.615	0.426	0.890
	`q_duration (log)`		1.075	1.051	0.824	1.339
	`q_clickpos`		1.031	1.031	0.961	1.106
	`num_query`		0.994	1.009	0.925	1.101
	`num_click`		0.974	0.958	0.898	1.023
	`duration`		0.957	0.952	0.867	1.046
3	`idf`			1.092	0.990	1.203
	`avg_jaccard (log)`			1.395	1.202	1.619
	`#click_title_hasw`			2.399	1.688	3.410
	`#click_snippet_hasw`			0.688	0.494	0.957
	`#click_URL_hasw`			1.281	0.957	1.717
	`#skip_title_hasw`			1.442	1.045	1.990
	`#skip_snippet_hasw`			0.752	0.575	0.985
	`#skip_url_hasw`			0.775	0.548	1.097
	`freq_w_suggest`			0.317	0.049	2.058
−2 LL (null model: 1686.2)		1677.6	1666.8	1598.125		
Omnibus Tests of Model Coefficients				***		

- Light, dark, and darker shadings indicate the coefficients are significant at 0.05, 0.01, and 0.001 levels, respectively. Unshaded cells are not significant at 0.05 level.
- *, **, and *** indicate the model is statistical significant from its baseline model at 0.05, 0.01, and 0.001 levels, respectively. Model 1 uses the null model as the baseline (with only a constant term). Model 2 uses Model 1 as the baseline, and Model 3 uses Model 2 as the baseline.

Table 51: Hierarchical logistic regressions: REUSE (whether or not to reuse a word to the next query) as dependent variable.

Step	**Variable Name**	**Model 1**	**Model 2**	**Model 3**		
		exp(B)	**exp(B)**	**exp(B)**	**95% CI**	
1	`constant`	0.011	0.009	0.006	-	-
	`product (I)`	2.121	1.380	1.031	0.337	3.153
	`goal (A)`	0.907	1.176	1.155	0.403	3.313
	`familiarity`	0.883	0.935	0.847	0.566	1.267
	`avg_num_add`	2.743	2.354	2.703	0.799	9.143
2	`q_length (log)`		1.006	0.913	0.487	1.712
	`q_duration (log)`		1.467	1.386	0.878	2.188
	`q_clickpos`		0.921	0.902	0.731	1.114
	`num_query`		1.026	1.067	0.946	1.205
	`num_click`		0.974	0.980	0.821	1.168
	`duration`		0.879	0.851	0.681	1.064
3	`idf`			1.020	0.860	1.209
	`p(w|pastq)`			0.965	0.740	1.259
	`avg_jaccard (log)`			4.663	1.003	21.683
	`#click_title_hasw`			1.693	0.458	6.253
	`#click_snippet_hasw`			2.253	0.891	5.694
	`#skip_title_hasw`			0.824	0.443	1.530
	`#skip_snippet_hasw`			1.364	0.821	2.268
	`freq_w_suggest`			3.152	0.810	12.265
	−2 LL (null model: 370.3)	363.0	356.3	339.5		
	Omnibus Tests of Model Coefficients			*		

- Light , dark , and darker shadings indicate the coefficients are significant at 0.05, 0.01, and 0.001 levels, respectively. Unshaded cells are not significant at 0.05 level.
- *, **, and *** indicate the model is statistical significant from its baseline model at 0.05, 0.01, and 0.001 levels, respectively. Model 1 uses the null model as the baseline (with only a constant term). Model 2 uses Model 1 as the baseline, and Model 3 uses Model 2 as the baseline.

Table 52: An example of word reusing in query reformulation.

No.	Query
1	differences in dehumidifier
2	differences in dehumidifier 500 sq ft room
3	hygrometer
4	hygrometer amazon
5	**dehumidifier** ACH

not directly influence users' decisions on whether or not to add a word in query reformulation, regardless of adding a brand-new word or reusing a word.

7.3.5 Influence of the Query-level Variables

Results show that the query-level variables have limited influence on adding a word or not in the next query. This applies to both adding a new word (NEW) and reusing a word (REUSE). As Table 50 shows, only one query-level variables shows a significant influence on adding new words in Model 3, and none have any significant influence on reusing. For both NEW and REUSE, their Model 2 do not significantly improve over Model 1. Even taking into account the limited sample size for REUSE, results suggest very limited influence of the query-level variables on adding a new word and reusing a word in query reformulation.

The length of the current query (`q_length`) shows a significant negative influence on adding a new word in Model 3. This is not surprising because longer queries are usually more specific. Adding a new word can make it overspecific, which is less likely to be adopted by the searchers.

7.3.6 Influence of the Word-level Variables

Word-level variables show relatively strong influence on whether or not to add a new word to the next query (NEW), and they also show a clear influence on reusing words in query reformulation (REUSE). After including the word-level variables, Model 3 for both dependent variables significantly improve over Model 2 ($p < 0.001$ for NEW, and $p < 0.030$ for REUSE). This indicates that the word-level variables are more salient factors for adding a word in query reformulation compared with the session and query-level variables.

Results suggest that users are more likely to add new words that co-occur frequently with existing words in the current query (`avg_jaccard` shows a significant positive influence on NEW). This is not surprising because words with low `avg_jaccard` values are more likely off-topic and may retrieve low quality results.

Results also show that users are more likely to add a new word to the next query if it appeared frequently in results' titles they viewed on the current query's SERP, regardless of whether or not they clicked on the results (both `#click_title_hasw` and `#skip_title_hasw` show significant positive influence on NEW). On the contrary, users are less likely to add a new word if it appeared frequently in the result snippets they viewed on the current query's SERP (both `#click_title_hasw` and `#skip_title_hasw` show significant negative influence on NEW). Results in Table 53 also agrees with these trends. The new words added to the next query (NEW = 1) appeared in significantly more clicked ($p < 0.01$) and skipped result titles ($p < 0.05$) compared with other words in the candidate sets (NEW = 0). The added new words also appeared in significantly fewer skipped result snippets ($p < 0.05$), although the difference is not significant for clicked result snippet.

This indicates that the occurrences of a new word in result titles may play a more important role on users decisions of whether or not to add the word to the next query. This is probably because results' titles are more eye-catching than other SERP elements, as most current search engines show results' titles using a larger font size than other SERP elements. However, as Table 49 shows, only 15.2% of the added new words can be located in result titles, in contrast to 23.5% in snippets and 26.7% in summaries (both titles and snippets). This indicates result snippets still provide valuable ideas for new words in query reformulation, but the new words do not necessarily appear more often than other words in result snippets. In fact, both results in Table 50 and Table 53 suggest that they appear less often in snippets compared with other words that were not added to the next query. This is probably because result snippets are noisy, including both relevant and off-topic words.

Table 53: Mean values (S.E.) of variables for new words added and not added to the next query.

Variables	**NEW = 1**	**NEW = 0**	
`#click_title_hasw`	0.243 (0.054)	0.079 (0.003)	**
`#click_snippet_hasw`	0.270 (0.058)	0.286 (0.005)	
`#click_url_hasw`	0.224 (0.050)	0.108 (0.004)	*
`#skip_title_hasw`	0.257 (0.056)	0.186 (0.004)	*
`#skip_snippet_hasw`	0.487 (0.074)	0.605 (0.006)	**
`#skip_url_hasw`	0.191 (0.047)	0.221 (0.006)	

- *, **, and *** indicate significance at 0.05, 0.01, and 0.001 levels, respectively, by a two-tail Welch's *t*-test.

In contrast, only `p(w|pastq)` (how frequently the word was used in past queries of the session) shows a significant effect on reusing a word (REUSE). None of the other word-level variables show any significant effects. This indicates that while reusing a word, users usually simply reuse the word associated to the main theme of the task.

I manually examined the cases of reusing a word in the collected data. It has been found that reusing a word in query reformulation mostly happens when users revert from a subtask to the main task, or to another subtask. Table 52 shows an example. The user first explored differences in dehumidifiers in the first two queries, and then switched to look for information related to hygrometer in the third and the fourth queries. After finishing exploring hygrometer, the user reverted back to continue to explore dehumidifier and thus reused the word *dehumidifier* in the fifth query. This example explains why users are more likely to reuse words related to the main theme of the task (with high `p(w|pastq)` values).

7.3.7 Summary of Findings

To summarize, results in this section suggest that adding a brand-new word ($\mathrm{Y}_{\mathrm{new}}$) and reusing a word ($\mathrm{Y}_{\mathrm{reuse}}$) stand for different intentions of users. Such distinctions were not identified in previous studies. Users exploit highly related, unused words from the results they viewed and include them into the next query, as suggested by the positive influence of `avg_jaccard`, `#click_title_hasw`, and `#skip_title_hasw`. In contrast, they reuse a word when reverting from a subtask to the main task or another subtask, as suggested by the positive influence of `p(w|pastq)` and manual analysis.

Compared with removing or retaining a word, results show that adding a word is less likely influenced by the session and query-level variables, as suggested by the limited effects of session and query-level variables on both NEW and REUSE. This indicates that users may make decisions on whether or not to add a word mostly based on local factors that are directly related to the word. Such decisions are less likely influenced by task characteristics or past user activities in the session.

7.4 PREDICTING WORD CHANGES IN QUERY REFORMULATION

7.4.1 Tasks

The previous two sections examined the influence of different factors (as measured by the independent variables) on searchers' choices of word changes. This section further investigates RQ3 by designing approaches for predicting word changes in query reformulation based on various features in a search session. Formally, for a query reformulation from q_i to q_{i+1} in a search session, the task is:

- [**RMV prediction**] – for each word $w \in q_i$, predict whether or not the word will be removed (instead of retained) in q_{i+1}.
- [**ADD prediction**] – for each word w from the candidate sets for NEW and REUSE words (described in Section 7.3.2), predict whether or not the word will be added to the next query q_{i+1}. Note that this section does not further separately study adding a new word (ADD) and reusing a word (REUSE).

The two tasks assume that all information of the current query (including searchers' activities on the current query's SERP) and previous queries are available. This stands for a prediction task for word changes at the moment that the searchers are about to reformulate (e.g., when the searchers are about to input content in the search box). This is similar to the task scenario for query auto-completion (Bar-Yossef & Kraus, 2011; Shokouhi & Radinsky, 2012; Shokouhi, 2013) without receiving any inputs of the new query. This section examines the two tasks separately. Both task can be considered as binary classification problems.

7.4.2 Classifier

This study adopts a feature-based approach for predicting word changes. It uses Gradient Boosted Regression Trees (GBRT) (Friedman, 2001, 2002), along with different sets of features, to predict word changes in query reformulation. GBRT is a general supervised learning technique for regression and classification problems based on gradient boosting. It has the advantage of handling features with non-linear relationships with the targets. In addition, GBRT is robust for features of heterogeneous nature. These advantages make GBRT a suitable option for the purpose of predicting word changes, because the adopted features vary a lot. I also examined other classification models and found that linear approaches such as logistic regression regression and support vector machines generally performed worse than GBRT using the same sets of features. Therefore, this dissertation only presents results using the GBRT classification.

This study focuses on comparing the effectivenesses and contribution of different sets of features in predicting word changes. The adopted features can be divided into different groups based on the types of information involved in computing the features. In addition, each group is designed with regard to an hypothesis:

- Word features (H3.1) – features in this group only consider information of the word itself.
- Query-word relation features (H3.2) – features in this group consider the relation of the word and search queries, including both the current search query and previous queries in a search session.
- SERP-word relation features (H3.3) – features in this group consider the relation of the word and SERP elements, including the retrieved result summaries and the displayed query suggestions in a search session.

- Query features (H3.4) – it is assumed that searchers' choices of words are also related to the situation of formulating the query. Therefore, features in this group include the characteristics of the current query (but independent of the word) such as how many results the searchers clicked on.
- Session features (H3.4) – in addition to the query features, this section also included features related to the time point of a session where the query reformulation happens (session features).

The rest of this section describes these features (some of them are identical to the variables examined in the last section). These features are also examined by their correlations (Pearson's r and Spearman's ρ) with the two dependent variables in the dataset. Table 54, 55, 56, 57, and 58 report these features and their correlations. Note that the reported correlations with RMV are based on the raw RMV observations (687 instances), but those with ADD are based on a large bootstrap sample ($N = 20,000$) where positive and negative instances are of equal size (because the raw observations for both ADD and REUSE are highly biased to the negative instances). Note that due to the large N, it is easy to observe statistically significant correlations in the bootstrap sample.

7.4.3 Word Features

Word features only use information of the target word. These features include:

- `#char` – the number of characters in the target word. It is assumed that the length of the word is related to many factors such as the cost of inputting the word into the query box, as well as the complexity of the word, etc. Therefore, the length of the word may be a predictor of word changes.
- `idf` – IDF (inverse document frequency) (Spärck Jones, 1972) is a measure for word specificity. Here the IDF of the word is computed using the Clueweb09 dataset (in order to be consistent with the last section).

Table 54 reports the correlation of the word features with RMV and NEW. As the table shows, word character length (`#char`) has a significant and weak negative correlation with

Table 54: Correlation of word features with RMV and ADD.

Feature	**RMV**		**ADD**	
	Pearson	Spearman	Pearson	Spearman
`#char`	−0.11	−0.14	0.04	0.06
`idf`	−0.05	−0.13	−0.08	0.01

- Light , dark , and darker shadings indicate the correlations are significant at 0.05, 0.01, and 0.001 levels, respectively. Unshaded cells are not significant at 0.05 level.
- The correlations for NEW are based on a large bootstrap sample ($N = 10,000$) of the raw observations with an equal number of positive and negative instances.

RMV, indicating that in generally searchers seem less likely to remove longer words.

7.4.4 Query-word Relation Features

Query-word relation features are computed based on the target word and search queries (the current query and/or past search queries in the session). Table 55 reports these features and their correlations with RMV and NEW. This group of features includes the following:

- `p(w|pastq)` – `p(w|pastq)` is a feature for the chances of observing the word in past queries, which was also included in the analysis models in the previous sections. As the previous sections examined, words that were frequently included in past queries are less likely to be removed from search queries, and are very likely to be reused (if they had been removed from the current query).

- `jaccard` – `jaccard` is a set of features for the Jaccard similarity between the examined word and a query term other than the examined word in the current query. Here the Jaccard similarity of two words are computed based on the sets of documents where the term appeared in (using the Clueweb09 dataset). Because the current query may include many query terms, this study computes the Jaccard similarity of the examined word with

Table 55: Correlation of query-word features with RMV and ADD.

Feature	**RMV**		**ADD**	
	Pearson	Spearman	Pearson	Spearman
`p(w|pastq)`	−0.27	−0.26	0.12	0.10
`jaccard_max`	−0.03	0.02	0.11	0.10
`jaccard_min`	−0.07	−0.04	0.20	0.21
`jaccard_mean`	−0.05	0.01	0.18	0.18

- Light, dark, and darker shadings indicate the correlations are significant at 0.05, 0.01, and 0.001 levels, respectively. Unshaded cells are not significant at 0.05 level.
- The correlations for NEW are based on a large bootstrap sample ($N = 10,000$) of the raw observations with an equal number of positive and negative instances.

each word in the current query, and uses their maximum, minimum, and mean values as features.

As reported in Table 55, `p(w|pastq)` also shows a significant weak and negative correlation with RMV, and a positive one with NEW, confirming the findings in previous sections. In addition, the `jaccard` features (especially `jaccard_min` and `jaccard_mean`) show significant weak and positive correlations with ADD, suggesting that the added words are likely those with high Jaccard similarity with every word in the current query.

7.4.5 SERP-word Relation Features

The SERP-word relation features are computed based on the occurrences of the examined word in SERP elements such as the result snippets and the displayed query suggestions. Table 56 reports these features and their correlations with RMV and ADD. This group of features include the following:

- `freq_currq` – `freq_currq` is a set of features for the frequency of the word in SERP elements displayed on the current query's SERP. For example, `freq_currq_clicked_title`

Table 56: Correlation of SERP-word features with RMV and ADD.

Feature	**RMV**		**ADD**	
	Pearson	Spearman	Pearson	Spearman
`freq_currq_clicked_title`	−0.15	−0.15	0.16	0.13
`freq_currq_clicked_snippet`	−0.10	−0.09	−0.01	−0.07
`freq_currq_clicked_url`	−0.15	−0.13	0.11	0.12
`freq_currq_unclicked_title`	−0.20	−0.21	0.14	0.10
`freq_currq_unclicked_snippet`	−0.17	−0.17	−0.11	−0.02
`freq_currq_unclicked_url`	−0.18	−0.19	0.07	0.09
`freq_currq_suggestion`	−0.12	−0.11	0.08	0.11
`freq_pastq_clicked_title`	−0.14	−0.18	0.13	0.10
`freq_pastq_clicked_snippet`	−0.06	−0.13	−0.08	−0.08
`freq_pastq_unclicked_title`	−0.21	−0.24	0.09	0.09
`freq_pastq_unclicked_title`	−0.10	−0.22	−0.10	−0.01

- Light, dark, and darker shadings indicate the correlations are significant at 0.05, 0.01, and 0.001 levels, respectively. Unshaded cells are not significant at 0.05 level.
- The correlations for NEW are based on a large bootstrap sample ($N = 10,000$) of the raw observations with an equal number of positive and negative instances.

Table 57: Correlation of query features with RMV and ADD.

Feature	**RMV**		**ADD**	
	Pearson	Spearman	Pearson	Spearman
currq_length	0.12	0.12	−0.09	−0.06
currq_clickpos	0.01	0.02	0.04	0.04
currq_noclick	−0.04	−0.04	−0.05	−0.05
currq_dwell	0.05	0.03	0.05	0.06

- Light, dark, and darker shadings indicate the correlations are significant at 0.05, 0.01, and 0.001 levels, respectively. Unshaded cells are not significant at 0.05 level.
- The correlations for NEW are based on a large bootstrap sample ($N = 10,000$) of the raw observations with an equal number of positive and negative instances.

means the frequency of the word in the titles of the clicked results displayed on the current query's SERP (the mean value of the frequency in different clicked results' titles).

- freq_pastq – similarly, freq_pastq is a feature for the frequency of the word in SERPs retrieved by previous queries in the session.

Results in Table 56 confirm findings in previous sections. The frequencies of the examined words in the current SERP's elements show negative correlations with RMV (this is consistent with the negative influence of #click_title_hasw and others on RMV reported in Table 46). In addition, the frequencies of the examined words in titles and snippets show positive and negative correlations, respectively, with ADD (this is also consistent with the positive influence of #click_title_hasw and #skip_title_hasw, and the negative influence of #click_snippet_hasw and #skip_snippet_hasw on NEW in Table 50). Moreover, the freq_pastq features also show certain correlations with both RMV and ADD, suggesting that the occurrences of the examined word in previous SERPs retrieved in a session may also be helpful predictors of word changes.

7.4.6 Query and Session Features

Query features compute the characteristics of the current query. Session features characterize the time of the session where the query reformulation happens. Both the two groups of

Table 58: Correlation of session features with RMV and ADD.

Feature	RMV		ADD	
	Pearson	Spearman	Pearson	Spearman
`timespent`	0.04	0.05	-0.06	-0.06
`timespent_perq`	-0.05	-0.08	0.03	0.02
`numq`	0.14	0.11	-0.08	-0.06
`numclick`	-0.06	-0.03	-0.08	-0.09
`numclick_perq`	-0.11	-0.13	-0.02	-0.02

- Light , dark , and darker shadings indicate the correlations are significant at 0.05, 0.01, and 0.001 levels, respectively. Unshaded cells are not significant at 0.05 level.
- The correlations for NEW are based on a large bootstrap sample ($N = 10,000$) of the raw observations with an equal number of positive and negative instances.

features are independent of the examined words. Table 57 and 58 report the correlation of these features with RMV and ADD.

Query features includes:

- `currq_length` – the length of the current search query.
- `currq_clickpos` – the rank of the lowest clicked result on the current query's SERP.
- `currq_noclick` – the value is 1 if there is no click on the current query's SERP, or 0 if there is any click.
- `currq_dwell` – the time spent on the current query's SERP (log value).

Session features include:

- `timespent` and `timespent_perq` – the time spent in the session starting from the submission of the first query, and the average time spent on a query in previous searches.
- `numq` – the number of submitted queries in previous searches.
- `numclick` and `numclick_perq` – the total number of clicked results in previous searches, and the average number for each previous search.

7.5 EVALUATION

7.5.1 Experiment Settings

This section evaluates the effectivenesses of the described approaches in order to answer RQ3. The experiment compares the effectivenesses of different sets of features in predicting word changes. The experiment separately evaluates the effectiveness of predicting removing or retaining (RMV) and adding a word or not (ADD).

As the previous section reported, the dataset for RMV is relatively balanced on both positive and negative cases. As a result, this section directly performs experiments on this dataset. The experiment generates 10 random partitions of the dataset. On each partition, a 10-fold cross validation is performed to evaluate the effectiveness of an approach (a set of features), where 7 folds are used for training, 2 folds are used for validation, and 1 fold is used for testing. All evaluated approaches use the same 10 partitions and cross-validation settings to make sure that results on each test fold are comparable.

However, the dataset for ADD is strongly biased towards the negative cases due to the issue of candidate sets. As a result, this study randomly samples a balanced dataset to perform the experiments. The experiment separates the positive and negative instances, and generates 10 random partitions for both positive and negative cases. For each partition, the experiment further generate 10 folds of the positive cases. For each fold, an equal number of negative cases are randomly sampled from the original sample of the negative samples to make the distribution of the dataset equal on both positive and negative instances.

7.5.2 Evaluation

The effectiveness of the prediction is evaluated using the following measures:

- Precision, recall, and F1 on the positive cases (for RMV, the positive cases are the removed word; for ADD, the positive cases are the added words).
- Precision, recall, and F1 on the negative cases (for RMV, the positive cases are the retained word; for ADD, the positive cases are the words that were not added to the queries).

- Average F1 of the positive and negative cases.
- Overall accuracy of the prediction.

For each approach (feature set), the experiment yields results on 100 test folds in total (10 test folds for each partition, and 10 random partitions). The following sections report the mean values of the measures on the 100 test folds. Differences between approaches are compared by whether the mean values of the approaches on the 100 test folds are significantly different using paired *t*-test.

7.5.3 Predicting Removing or Retaining

This section evaluates the proposed approaches for predicting whether to remove or to retain a word. The focus of the experiment is to compare the effectivenesses of different sets of features in prediction. Table 59 reports the results.

7.5.3.1 Word Features As Table 59 shows, word features are the most effective set of features among the five for predicting whether to remove or to retain a word, although only two features are included. This suggests that searchers' choices of whether to remove or to retain a word in a query reformulation can be predicted by, and are likely related to, the word itself, independent of other issues. As Table 54 shows, the correlations suggest that searchers retained long words and specific words (with high IDF) more often in the collected dataset. But it requires further studies to verify whether these factors have influence on searchers' choices. Here the results only indicate word features are very likely useful ones for predicting searchers' decisions on removing or retaining.

7.5.3.2 Query-word Relation Features Query-word relation features are also strong predictors compared with word features. Both groups of features have similar prediction performance. This suggests that the relations of the target word with the search queries are also important evidence for predicting searchers' word changes.

7.5.3.3 SERP-word Relation, Query, and Session The rest three groups of features are only marginally useful for prediction. The performance looks weak compared with the other two groups of features, especially considering the number of predictors included. This

Table 59: Effectiveness of prediction models for whether to remove or to retain a word in query reformulation.

Feature	**Avg F1**	**Accuracy**	**Remove**			**Retain**		
			F1	**P**	**R**	**F1**	**P**	**R**
Word	0.69	0.70	0.63	0.68	0.59	0.74	0.70	0.78
Query-word	0.68	0.69	0.63	0.66	0.61	0.73	0.71	0.75
SERP-word	0.62	0.63	0.55	0.60	0.51	0.69	0.65	0.73
Query	0.63	0.64	0.58	0.61	0.55	0.69	0.67	0.72
Session	0.63	0.64	0.56	0.60	0.53	0.69	0.66	0.73
All	**0.75**	**0.76**	**0.72**	**0.73**	**0.71**	**0.79**	**0.78**	**0.79**

- Reported values are the mean values on the 100 test folds.
- All differences with “All” are significant at least at 0.01 level by paired *t*-test (except for the recall on retained words using “Word” features).

suggests that they are less important predictors for searchers decision on removing or retaining. This also suggests that searchers decisions on removing or retaining is less likely related to SERP-word relations compared with query-word relations. However, the performance of these features all outperform a null model baseline (based on prior probability of classes), showing that they are still useful features for prediction.

7.5.3.4 Overall Prediction Effectiveness After combining all sets of features, the prediction performance reaches an overall accuracy of 0.76. The performance of using all features is also significantly better than using any individual sets of features, showing that it is beneficial and generally successful to combine different groups of features. These also verifies that the designed different sets of features are effective and complementary to each other. In addition, no individual set of features can reach a similar level of effectiveness, indicating that the users' decisions on removing or retaining are probably rather diverse and relate to different factors.

The prediction performance is also balanced on both removed and retained words, which a sightly worse performance on predicting removed words. The final model can predict removed words with a precision of 0.73 and a recall of 0.71. In contrast, the overall prediction precision on retained word is 0.78, with a recall of 0.79.

7.5.4 Predicting Adding or not

This section further evaluates the prediction performance for whether or not to add a word in the next query. Table 60 reports the prediction performance on the randomly generated balanced datasets.

7.5.4.1 SERP-word Relation Features As Table 60 shows, SERP-word relation features are the most effective set of features among the five sets of features. This set of features, along, reaches a prediction performance that is comparable to those using the combination of all features. This also suggests that SERP-word relation is a very important factor related to searchers' decisions on whether or not to add a specific word in the next query, which is also consistent with the findings in Section 7.3. In addition, this is also different from the

Table 60: Effectiveness of prediction models for whether or not to add a word in query reformulation.

Feature	**Avg F1**	**Accuracy**	**Add**			**Do not Add**		
			F1	**P**	**R**	**F1**	**P**	**R**
Word	0.65	0.66	0.69	0.64	0.75	0.62	0.68	0.56
Query-word	0.61	0.61	0.61	0.63	0.59	0.60	0.58	0.63
SERP-word	0.74	0.74	0.75	0.76	0.74	0.72	0.71	0.73
Query	0.52	0.52	0.50	0.51	0.49	0.53	0.52	0.54
Session	0.56	0.56	0.55	0.57	0.52	0.58	0.56	0.60
All	**0.76**	**0.76**	**0.78**	**0.77**	**0.79**	**0.74**	**0.76**	**0.73**

- Reported values are the mean values on the 100 test folds.
- All differences with “All” are significant at least at 0.01 level by paired *t*-test (except for the recall on words that were not added to the queries using “SERP-word” features).

results reported for predicting removing or retaining, where the SERP-word relation features seem less effective compared with others.

7.5.4.2 Word Features Word features is the second most effective set of features among the five sets for predicting whether or not to add a query term in the next query. This suggests that searchers' decisions on whether or not to add a query term is also likely related to the characteristics of the word itself. In addition, it has been observed that word features has the best recall in predicting added words, suggesting a unique contribution of the word features to this prediction problem.

7.5.4.3 Query-word Relation Features Query-word relation features are also somewhat effective in predicting added words. However, the effectiveness of prediction is not comparable to SERP-word features and word features. This suggests that query-word relation is not as important as SERP-word features in predicting whether or not to add a new word in the next query. This is also consistent with the findings in Section 7.3, where it was found that the most typical cases of adding a word into the next query is to perform relevance feedback—searchers find useful and interesting words from the results they examined and add them to the search queries.

7.5.4.4 Query and Session Features Query and session features are those independent of the examined word itself. They are indicative of the situation in a search session where the query reformulation takes place. As the table shows, these two sets of features generally have very limited effectiveness in prediction. This is also consistent with the findings in Section 7.3, where results suggest that query-level independent variables (similar to the query and session features in this section) do not show any significant influence on ADD. This is also different from those for predicting removing or retaining, suggesting that the decisions on adding a word or not is probably related to the situation of a search session by a smaller extent compared with those for removing or retaining.

7.5.4.5 Overall Prediction Effectiveness Overall the prediction performance on the collected dataset achieves a moderately accurate level. Using all five sets of features is relatively more effective in predicting whether or not a word will be added into the query, with 0.77 precision and 0.79 recall. In comparison, the performance on the negative instances are relatively lower, with 0.76 in precision and 0.73 in recall. This verifies the effectiveness of our approaches for predicting query term addition. However, it suggests that other sets of features did not provide much additive values to the SERP-word relation features.

7.5.5 Comparing Two Prediction Tasks

As the last sections examined, using the same five sets of features, this section successfully trained effective models to predict both decisions on removing or retaining a word, and those on adding a word or not. However, the effectivenesses of the feature sets are very different for these two problems, showing that the nature of the decisions related to the two types of word changes are very different. However, the proposed approaches are effective in both tasks, showing that the design of features are generic and effective for predicting different types of word changes in query reformulation.

7.6 ANSWERS TO RQ3

This section examined the influence of different factors on two types of word changes in query reformulation during a search session—removing or retaining a word, and adding a word or not. In addition, techniques for predicting such changes have been proposed and verified for their effectivenesses. Results and analysis in this section answered the following research questions and examined the correctness of the hypotheses.

- **RQ3** – What affects users' choices of word changes in a query reformulation during a search session? Can we predict such choices of word changes?

Results show that users' choices of word changes in a query reformulation are affected by different factors depending on the types of the word changes in query reformulation. Whether to remove or retain a word is affected by a diverse sets of factors, including the word itself, its relation to the current and previous search queries, its occurrences in search results, as well as the situation of the session where the query reformulation happened. In contrast, whether or not to add a word to the query is influenced by a smaller set of factors, as suggested in the collected data, mainly including: the occurrences of the word in the current query's SERP (the case of adding a new word), and the relation of the word to previous queries (in the case of reusing a word).

Based on a rich set of features designed according to the hypotheses, effective classification models can be trained to predict searchers' choices of word changes with moderate accuracy (about 75% accuracy for both types of word changes).

- **[H3.1]** – Searchers' choices about a word in a query reformulation are related to the word itself.

The experimental results in Section 7.5 showed that the word features alone are effective for predicting searchers' choices on both types of word changes. However, related independent variables did not show significant influence in the hierarchical logistic regression models, suggesting that the influence of the factor is limited, after taking into account other factors. Overall this hypothesis is likely true, but it seems not a major factor for searchers' decisions on their word changes.

- **[H3.2]** – Searchers' choices about a word in a query reformulation are related to the connection between the word and search queries in a session.

Both the analysis in Section 7.2 and 7.3, and the experimental results in Section 7.5, suggest that this hypothesis is likely true. The connection of the word with other words in the current search query shows a significant influence on removing or retaining the word. In addition, the chances of observing the word in past queries have significant influence on both removing or retaining the word, and reusing the word. Moreover, experimental results show that the query-word relation features are also effective in predicting both types of word changes.

- **[H3.3]** – Searchers' choices about a word in a query reformulation are related to the connection between the word and retrieved results in a session.

Both the analysis in Section 7.2 and 7.3, and the experimental results in Section 7.5, suggest that this hypothesis is likely true. The occurrences of the word in different SERP elements (such as result title, snippet, and URL) show different influence on removing or retaining a word and adding a new word or not, but the influence in both cases are significant in the hierarchical logistic regression models. In addition, experimental results show that the SERP-word relation features are also very effective in predicting both types of word changes (especially for adding a word or not).

- **[H3.4]** – Searchers' choices of word changes in a query reformulation are related to the situation in a search session where the query reformulation happens.

Both the analysis in Section 7.2 and 7.3, and the experimental results in Section 7.5, suggest that this hypothesis is likely true for removing or retaining a word. Some related variables show significant influence in the hierarchical logistic regression models for removing or retaining a word. Experimental results also verified that the query and session features are effective in predicting removing or retaining a word (although not as effective as other features). But both the hierarchical logistic regression models and the prediction results suggest these factors are not closely related to whether or not to add a word.

- **[H3.5]** – Searchers' choices of word changes in a session are related to the task's goal and product.

For both types of word change decisions, results show that task goal and product do not have any direct influence on these decisions, suggesting that users make their decisions of word use and changes in a search session mainly based on factors that are specific to the words. However, many previous studies (J. Liu et al., 2010; C. Liu et al., 2012; Jiang et al., 2014) also showed that searchers have different activity patterns in a search session. As searchers' decisions on word changes are related to their search activities in a search session (as H3.2–3.4 examined), task goal and product may still have an indirect influence on these word change decisions—task goal and product influence searchers' overall search activity patterns, and subsequently these patterns may influence their choices on word changes.

- **[H3.6]** – Searchers' choices of word changes in a session are related to the searcher.

Due to the collected data in User Study 2, the examination of user-related factors are restricted to their topic familiarity and their preferences of adding and removing words (as measured by the average number of added and removed words in their query reformulations in other sessions) in this chapter. Only the users' preferences of removing words show a significant influence on removing or retaining. Overall it requires further investigations to understand the influence of user-rated factors on their query reformulation behavior.

To conclude, this section shows that searchers' query reformulation behavior (and specifically their decisions on word changes) generally relate to many factors in a search session (H3.2–H3.5 are all related to factors in a search session) as well as those independent of the search session (H3.1). This indicates that search session is crucial for understanding and modeling users' query reformulation behavior.

8.0 CLICK DECISION IN A SEARCH SESSION

This chapter examines and predicts whether or not searchers would click on a result after they viewed its summary displayed on a SERP (search result page). This is referred as users' click decisions. A result summary refers to the result's title, snippet, and URL displayed on the SERP. This issue was examined in many previous studies (Clarke et al., 2007; Cutrell & Guan, 2007; Z. Guan & Cutrell, 2007; Y. Yue et al., 2010). In contrast, this chapter extends these analysis to a search session in order to examine the influence of a search session on click decisions. This chapter examines five hypotheses related to RQ4:

- **RQ4** – What affects users' decisions on whether or not to click on a search result's link after viewing its summary displayed on a SERP? Can we predict such decisions?
- **[H4.1]** – Click decision is related to the task's product and goal.
- **[H4.2]** – Click decision is related to the searcher's characteristics.
- **[H4.3]** – Click decision is related to the time of the session when the decision was made.
- **[H4.4]** – Click decision is related to previous search queries in the session.
- **[H4.5]** – Click decision is related to previously clicked results in the session.

8.1 CLICK DECISION

8.1.1 Purpose

Clicking behavior is a widely studied topic in the past decade. Many studies found that higher-ranked results generally received more clicks (Joachims et al., 2005; Cutrell & Guan, 2007). Previous work on this topic proposed and explored at least five factors that may help explain observed clicks in a search log.

- **Examination** – Eye-tracking studies (Joachims et al., 2005; Cutrell & Guan, 2007; Z. Guan & Cutrell, 2007) showed that searchers' visual attention focuses on top-ranked results. These studies suggested that a major reason for the fewer observed clicks at lower ranks is that users are less likely to examine the results at lower ranks.
- **Rank** – Joachims et al. (2005) also observed that, although the observed chances of viewing the top two results were similar, searchers still clicked on the first one more often, even when the first one is less relevant than the second one. They hypothesized that there exists a trust factor—searchers trust the quality of results ranked higher by the search engines and thus are more likely to click on results at higher ranks. But trust is only one possible interpretation. More generally, this chapter refers to this factor as rank factor – searchers prefer to click on a higher-ranked result than a lower-ranked one even they viewed (examined) both.
- **Other clicked results** – Craswell et al. (2008) hypothesized that (the cascade model hypothesis), the chance of clicking on a result declines if results at higher ranks are relevant. This is because if searchers are satisfied by visiting results at higher ranks, they do not need to further visit the lower-ranked ones.
- **Result relevance** – Many reported that searchers generally have higher chances to click on more relevant results (Yilmaz, Shokouhi, Craswell, & Robertson, 2010). This is not surprising considering that a major purpose of providing result summaries is to inform users the relevance of the results before they actually visit the results.
- **Attractiveness** – Many reported that more attractive summaries receive more clicks, for example, those with the occurrences of query terms in their titles (Clarke et al., 2007; Y. Yue et al., 2010). This is called the attractiveness bias (Y. Yue et al., 2010).

In light of these findings, this section extends the analysis of the factors to those in a search session. The analysis focuses on users' click decision – whether or not to click on a result after viewing its summary. This excludes the examination factor mentioned above. In addition, the analysis excludes the influence of other clicked results by only considering click

Table 61: Distribution of total fixation time (in milliseconds) on clicked and unclicked snippets during the first browse of a SERP (counting from the SERP was first displayed to switching to a result web page or another SERP).

Percentile	**Clicked Snippets**	**Unclicked Snippets**
10	259	120
20	439	180
30	580	279
40	913	380
50	1067	489
60	1280	677
70	1453	857
80	2025	1116
90	2831	1695

decisions during the first browse of each SERP (from the SERP was first displayed to the first click or query reformulation). Other factors mentioned above are included as controls in the analysis. The purpose of the analysis is to:

- examine whether factors related to a search session influence searchers' click decisions *in addition to the currently identified factors.*

The rest of this section introduces the design of the analysis, including the dependent variable, independent variables, control variables, and the statistical analysis approach.

8.1.2 Dependent Variable

The dependent variable of the analysis is whether or not searchers clicked on a result provided they viewed the result's summary. This study relies on the eye fixations recorded by an eye-tracking device in User Study 2 to determine whether or not searchers viewed a result's summary. Practically this is determined by whether or not the total duration of the eye fixations on the result exceeds the threshold 300ms. The data recorded by the eye-tracking device is a sequence of observed eye-fixations, along with the duration for each eye fixation.

An eye fixation can be linked to a result based on the position of the eye fixation on the screen and the result displayed at that position on the screen. Here the total duration of eye fixations for a result refers to the sum of duration for all eye fixations that are linked to the result.

The 300ms threshold is set according to the thresholds adopted in previous eye-tracking studies (Joachims et al., 2005; Cutrell & Guan, 2007; Z. Guan & Cutrell, 2007). Table 61 shows the distribution of the total fixation duration on clicked and unclicked results from the log. Single fixation duration less than 100ms was removed due to their noisy nature. According to the table, the threshold included about 85% of the clicked results and about 65% of the unclicked results. Other results are considered as not viewed—at least the duration seems too short to claim that the searchers viewed the results. In total, the click decisions on 528 viewed results were collected, where 152 are positive (were clicked on).

In addition, this analysis only considers the click decisions during the first browse of each SERP—starting from the time the SERP was first displayed on the screen, to that the searcher switched to a result web page or a new SERP. This is to exclude the factor of other clicked results on the same SERP.

8.1.3 Independent Variables

Similar to the analysis of word changes, the independent variables are divided into three groups depending on their relations to the dependent variable. In addition, some control variables are included.

- Session-level variables are measured at the session-level as a whole. Each examined click decision within a search session shares the same influence from these variables. Session-level variables are related to H4.1 and H4.2.
- Query-level variables are measured at the query-level. Each examined click decision from the same SERP shares the same influence from these variables. Query-level variables are related to H4.3.
- Result-level variables are specific to each examined click decision. Result-level variables are related to H4.4 and H4.5. In addition, all the control variables are at the result-level.

Table 62: Session-level independent variables for analyzing click behavior.

Hypothesis	**Variable**	**Meaning**
H4.1	`product`	The product of the search task (*factual* or *intellectual*).
H4.1	`goal`	The goal of the search task (*specific* or *amorphous*).
H4.2	`familiarity`	Users' ratings on their familiarity to the task topic.
H4.2	`avg_num_click`	The average number of clicks in other sessions performed by the same searcher.

8.1.3.1 Session-level Independent Variables The session-level independent variables are similar to those included in the analysis of word changes in a search session. Table 62 shows these variables and their related hypotheses. These variables include: `goal` (either specific or amorphous); `product` (either factual or intellectual); `familiarity` (searchers' responses to their familiarity with the topic of the task using a 5-point scale).

In addition, `avg_num_click` computes the average number of clicks in other sessions performed by the same searcher. This variable intends to capture searchers' characteristics such as their frequencies of clicking in a search session—some searchers may prefer to constantly click on and check different results.

8.1.3.2 Query-level Independent Variables Table 63 shows the query-level independent variables and their related hypotheses. The query-level variables include the time spent in the session (`duration`), the number of past search queries (`numq`), and the number of past

Table 63: Query-level independent variables for analyzing click behavior.

Hypothesis	**Variable**	**Meaning**
H4.3	`duration`	Time duration from the beginning of the session to the submission of the new query.
H4.3	`numq`	Number of submitted queries in the session (including the current query).
H4.3	`numclick`	Number of past clicks in the session (including clicks on the current query's SERP).

Table 64: Result-level independent variables for analyzing click behavior.

Hypothesis	**Variable**	**Meaning**
H4.5	`visited`	Whether or not the result was visited (clicked on) in previous searches of the same session.
H4.4	`log P(pastq|title)`	The text similarity between the result's title and past search queries (using a normalized query likelihood score).
H4.4	`log P(pastq|snippet)`	The text similarity between the result's snippet and past search queries (using a normalized query likelihood score).
H4.5	`log P(clicked|title)`	The text similarity between the result's title and past clicked results' titles (using a normalized query likelihood score).
H4.5	`log P(clicked|snippet)`	The text similarity between the result's snippet and past clicked results' snippets (using a normalized query likelihood score).

clicks (`numclick`) in the session. All the three query-level variables were also included as query-level variables in the analysis of word changes in a search session. These variables are indicative of the time point of a session where the click decision was made. They are used to verify H4.3.

8.1.3.3 Result-level Independent Variables Table 64 shows the result-level independent variables and their related hypotheses. In addition, all the control variables (introduced in the next section) are also at the result level.

- `visited` is a binary variable for whether or not the result was visited (clicked on) in previous searches of the same session.

- `log P(pastq|title)` and `log P(pastq|snippet)` measure the text similarity between the result's title/snippet and previous search queries in the session. The similarity is measured using a normalized query likelihood (QL) score, where the combination of all past queries are considered as a long query. The QL score is normalized using the length of the long query (`pastq`) to ensure that the scores are comparable between different

Table 65: Independent variables for analyzing click behavior: control variables.

Variable	Meaning
`URLslash`	The number of slashes in the result's URL (excluding those in `http://`).
`length_title`	Length of the result's title (number of words).
`length_snippet`	Length of the result's snippet (number of words).
`freq(q,title)`	Sum of the query terms' frequencies in the result's title.
`freq(q,snippet)`	Sum of the query terms' frequencies in the result's snippet.
`freq(q,URL)`	Sum of the query terms' frequencies in the result's URL (by subsequence matching).
`rank`	The rank of the result on the SERP.
`P(click|r)`	The chances of clicking on results with the same relevance level.

instances.

$$\texttt{log P(pastq|title)} = \sum_w P(w|\texttt{pastq}) \cdot \log P(w|\texttt{title})$$

$$P(w|\texttt{pastq}) = \frac{1}{|\texttt{pastq}|} \cdot \sum_{q \in \texttt{pastq}} P(w|q)$$

- `log P(clicked|title)` and `log P(clicked|snippet)` measure the text similarity between the result's title/snippet and previous clicked results' titles/snippets in the session. The similarity is also measured using a normalized query likelihood score, where the concatenation of all clicked results' titles or snippets are considered as a long query.

8.1.3.4 Control Variables

The analysis also includes some control variables to take into account factors identified in previous work. Table 65 shows the control variables.

- `URLslash` is a variable for the number of slashes in the result's URL. It is included into the analysis because Clarke et al. (2007) found that this variable has significant influence on click behavior in their studies.

- `length_title` and `length_snippet` are variables for the length of the result summary. It is included into the analysis because Clarke et al. (2007) found that some similar variable shows significant influence on click behavior.

- `freq(q,title)`, `freq(q,snippet)`, and `freq(q,URL)` are variables for frequency of the query terms in result title, snippet, and URL, respectively. Some similar variables were found to have significant influence on clicking behavior (Clarke et al., 2007).
- `rank` – `rank` is a variable for the rank of the result displayed on the SERP. Here `rank` is included to consider the rank factor independent of examination, e.g., searchers are more likely to click on higher-ranked results because they trust the result's quality (Joachims et al., 2005).
- `P(click|r)` – `P(click|r)` is the overall probability of clicking on results with the same relevance grade after viewing their summaries. This value is estimated from the whole search log – the estimated probability for *highly relevant*, *relevant*, and *non-relevant* results are 0.54, 0.50, and 0.26, respectively. The purpose of including this variable is to control the factor related to the perceived relevance of the result from its summary (Yilmaz et al., 2010). As the estimated probabilities also show, searchers are more likely to click on results with higher relevance grades.

8.1.4 Analysis Approach

Similar to the analysis of word changes, the following section examines the influence of the independent variables on the dependent variable using hierarchical (multilevel) logistic regression, to take into account the dependency of the three groups of variables (Gelman & Hill, 2006). The analysis conducted in the following section uses a hierarchical model with three levels to study click decisions (level 1) on a query's SERP (level 2) in a search session (level 3).

8.2 ANALYSIS

This section analyzes the influence of different factors on searchers' click decisions after they viewed the result's summary. Table 66 reports the results of the hierarchical logistic regression. Model 1 included only the session-level variables. Model 2 further included the query-level variables. Model 3 included all three groups of variables (including the

Table 66: Hierarchical logistic regressions: click decision as dependent variable.

	Variable Name	Model 1 exp(B)	Model 2 exp(B)	Model 3 exp(B)	95% CI	
1	constant	0.236	0.291	0.110	-	-
	product (I)	1.214	1.273	1.014	0.617	1.666
	goal (A)	0.854	0.869	0.710	0.443	1.137
	familiarity	0.922	0.933	0.892	0.735	1.083
	avg_num_click	1.066	1.062	1.081	1.008	1.159
2	duration		0.999	0.999	0.998	1.001
	numclick		1.015	1.017	0.952	1.087
	numq		0.973	0.988	0.883	1.105
3	visited			0.294	0.141	0.611
	log P(pastq|title)			0.920	0.803	1.055
	log P(pastq|snippet)			1.116	0.982	1.269
	log P(clicked|title)			4.440	1.511	13.05
	log P(clicked|snippet)			0.610	0.170	2.188
	(control) URLslash			1.048	0.889	1.237
	(control) length_title			1.106	0.973	1.258
	(control) length_snippet			0.981	0.939	1.025
	(control) freq(q,title)			1.334	1.024	1.738
	(control) freq(q,snippet)			0.907	0.780	1.056
	(control) freq(q,URL)			1.128	0.906	1.404
	(control) rank			0.858	0.777	0.947
	(control) P(click|r)			159.8	25.46	1003.4
	−2 LL (null model: 633.9)	624.3	622.2	545.0		
	Omnibus Tests of Model Coefficients	*		***		

- Light, dark, and darker shadings indicate the coefficients are significant at 0.05, 0.01, and 0.001 levels, respectively. Unshaded cells are not significant at 0.05 level.
- *, **, and *** indicate the model is significantly better than its baseline model at 0.05, 0.01, and 0.001 level by Omnibus tests of model coefficients. Model 1's baseline model is the null model (with only a constant term); Model 2's baseline model is Model 1; Model 3's baseline model is Model 2.

control variables). Similar to the previous chapter, multicollinearity issues in the models are examined by whether or not the variance inflation factor (VIF) of the variable in these models exceeds the threshold of 5. All included variables satisfy VIF < 5.

8.2.1 Influence of the Control Variables

As Table 66 shows, Model 3 confirms the influence of many control variables reported in previous studies.

- `P(click|r)` shows a significant positive influence on result clicking ($p < 0.001$), suggesting that searchers are significantly more likely to click on a result if it is relevant. This is not surprising and consistent with those reported in many previous studies (Yilmaz et al., 2010).
- `rank` also shows a significant "positive" influence on result clicking ($p < 0.01$)—note that smaller values of `rank` stand for higher-ranked results, thus the negative influence of the variable means that searchers are more likely to click on a result if it is at a higher rank (note that this is under the condition that searchers viewed the result's summary, otherwise it can be interpreted as the examination factor). This also confirms the existence of a rank-related factor independent of examination on click decision. Joachims et al. (2005) explained this as a trust bias, but it is not the only interpretation.
- Among the many result summary characteristics examined by Clarke et al. (2007), only the frequency of the query terms in result title (`freq(q,title)`) shows a significant positive influence in result clicking ($p < 0.05$). The positive influence of `freq(q,title)` is also consistent with Clarke et al.'s (2007) and Y. Yue et al.'s (2010) findings. In addition, `freq(q,snippet)` and `freq(q,URL)` do not show any significant influence on clicking, which is also consistent with Clarke et al.'s (2007) findings.
- However, the model suggests that `URLslash` and `length_snippet` do not have any significant influence on click decision, which conflicts with those reported by Clarke et al. (2007). They found in an experiment that user preferred to click on results with shorter snippets and more URL slashes (more URL hierarchies). A possible reason for this inconsistency is that Clarke et al. (2007) analyzed the influence of different factors

separately (without any controls), while these variables might not have significant influence in their analysis when other factors were considered. Another possible reason is the differences in experiment setting—this study determines the dependent variables based on eye-tracking, while Clarke et al. (2007) used online swapping experiments.

Despite a few inconsistencies, the influence of the control variables are generally consistent with those reported in previous studies, suggesting that the experiment and analysis settings are reasonable and valid. In addition, although these variables were included as controls for other variables to test the hypotheses, other variables related to a search session also served as controls for these variables as well. Therefore, results in this section also confirm the influence of these control variables on click decision in addition to a wider range of other independent variables.

8.2.2 Influence of the Session-level Variables

The session-level variables show statistically significant influence on click decision. As Table 66 shows, Model 1 explains the collected data significantly better than the null model (with only a constant term). The -2 log likelihood is 624.3 for Model 1, compared with 633.9 of the null model ($p < 0.05$). However, the magnitude of improvement is practically small, showing the limited influence of the session-level variables. This is also not surprising considering session-level variables are not specific to each click.

`avg_num_click` (the average number of clicks in other sessions performed by the same searcher) shows a significant positive influence in both Model 1 (exp(b) $= 1.066$, $p < 0.01$) and Model 3 (exp(b) $= 1.081$, $p < 0.05$). This suggests that searchers are more likely to click on a result in a specific session if they tend to click on results more frequently in other sessions. This indicates that personal search activity preference is a possible factor influencing result clicking in a session.

Other session-level variables show no significant influence in any of the three models, suggesting that result clicking is unlikely influenced by factors such as task goal, product, and user's topic familiarity.

8.2.3 Influence of the Query-level Variables

The query-level variables show no significant influence on click decision in neither Model 2 nor Model 3. In addition, after including the query-level variables, model 2 also did not significantly improve over Model 1 at 0.05 level. This suggests that these variables do not have important influence on result clicking. These variables are indicative of the time point of a session where the click decision was made. Their insignificant influence in the models suggest that searchers' click decisions are less likely influenced by the time point of a session.

8.2.4 Influence of the Result-level Variables

The result-level variables show some significant influence on searchers' click decisions. After including the result-level variables, Model 3 significantly improved over Model 2. The -2 log likelihood is 525 for Model 3, compared with 622.2 for Model 2 ($p < 0.001$). As the previous sections reported, many control variables have significant influence on the dependent variable. In addition, among the variables related to a search session, `visit` and `log P(clicked|title)` also show significant influence on the dependent variable.

The significant negative influence of `visited` ($p < 0.01$) is unsurprising—users are less likely to click on a result again if they clicked on the link in previous searches of the same session. This is also generally consistent with Shokouhi et al.'s (2013) findings, although they showed that sometimes searchers also revisit clicked results.

In addition, `log P(clicked|title)` – the text similarity between the result's title and previously clicked results' titles – shows a significant positive influence on users' click decisions ($p < 0.01$). This suggests that in a search session, users are significantly more likely to click on results with similar titles to what they clicked on before. This seems unsurprising, but note that the significant influence was observed at $p < 0.01$ with many other variables as controls (including almost all known factors). This indicates that the similarity of a result's title with previously clicked results' titles stand for a unique factor that is sufficiently different from the existing ones.

To further interpret its influence, `log P(clicked|title)` was replaced with two variables `log P(clicked, R|title)` and `log P(clicked, NR|title)`. They separately measure the text similarity of the result's title with the titles of previously clicked relevant results (`log P(clicked, R|title)`) and non-relevant results (`log P(clicked, NR|title)`). Table 67 presents the results of the new model, where the influence of other factors are consistent with the models presented in Table 66.

Table 67 shows that `log P(clicked, R|title)` has a significant positive influence on click decision ($p < 0.05$), while `log P(clicked, NR|title)` has none. This clarifies that searchers are more likely to click on results with similar titles to the relevant results they visited before.

`log P(clicked, R|title)` stands for a new source of attractiveness bias that has not been observed before in previous studies. It shows that searchers' click decisions are influenced by the correct click decisions they made in previous searches of the session—once they clicked on a result and found it was relevant, they are more likely to click on results with similar titles in follow-up searches. In contrast, the do not "learn" much from failed click decisions, as suggested by the insignificant influence of `log P(clicked, NR|title)`.

The significant influence of `log P(clicked, R|title)` suggests that applications such as click models (Chuklin et al., 2015) should take into account session-level factors better model searchers' clicks. For example, the importance of clicking on a result that is similar to previously clicked results in a session should be penalized, because this observed click is likely affected (boosted) by previous clicks. Other applications based on user clicks, such as online interleaved experiments may also benefit from taking into account such a session-level attractiveness bias, which is missing in current studies (Hofmann et al., 2012).

In addition, results show that the text similarity of the result with previous search queries do not show any significant influence on click decision, suggesting that users' click decisions are less likely affected by previous search queries.

8.3 PREDICTION

The task of predicting click decisions can be considered as a binary classification problem. It predicts, after viewing a result summary, whether or not searchers will click on the link.

Table 67: Hierarchical logistic regressions, separately considering previously clicked relevant results and non-relevant results: click decision as dependent variable.

	Variable Name	Model 1	Model 2	Model 3		
		exp(B)	exp(B)	exp(B)	95% CI	
1	constant	0.236	0.291	0.099	-	-
	product (I)	1.214	1.273	1.014	0.617	1.667
	goal (A)	0.854	0.869	0.710	0.443	1.137
	familiarity	0.922	0.933	0.893	0.735	1.084
	avg_num_click	1.066	1.062	1.081	1.008	1.159
2	duration		0.999	0.999	0.998	1.001
	numclick		1.015	1.017	0.952	1.087
	numq		0.973	0.988	0.883	1.105
3	visited			0.293	0.141	0.610
	log P(pastq|title)			0.920	0.803	1.054
	log P(pastq|snippet)			1.117	0.983	1.270
	log P(clicked,R|title)			4.830	1.279	18.239
	log P(clicked,NR|title)			0.870	0.218	3.467
	log P(clicked|snippet)			0.608	0.170	2.183
	(control) URLslash			1.048	0.888	1.236
	(control) length_title			1.107	0.974	1.259
	(control) length_snippet			0.981	0.939	1.025
	(control) freq(q,title)			1.332	1.022	1.736
	(control) freq(q,snippet)			0.907	0.779	1.055
	(control) freq(q,URL)			1.129	0.907	1.405
	(control) rank			0.858	0.778	0.947
	(control) P(click|r)			160.2	25.486	1007
	−2 LL (null model: 633.9)	624.3	622.2	544.9		
	Omnibus Tests of Model Coefficients	*		***		

- Light, dark, and darker shadings indicate the coefficients are significant at 0.05, 0.01, and 0.001 levels, respectively. Unshaded cells are not significant at 0.05 level.
- *, **, and *** indicate the model is significantly better than its baseline model at 0.05, 0.01, and 0.001 level by Omnibus tests of model coefficients. Model 1's baseline model is the null model (with only a constant term); Model 2's baseline model is Model 1; Model 3's baseline model is Model 2.

Here the purpose of the task is to serve as a benchmark to examine whether session-level information can help existing ones (without using session information) in modeling users' click decisions after viewing result summary.

Note that although this task itself does not have a direct application value, it has many connections with other applications. For example, click models (Chuklin et al., 2015) and online interleaved experiments (Hofmann et al., 2012; Radlinski & Craswell, 2013; Hofmann, Whiteson, & Rijke, 2013; Chuklin, Schuth, Hofmann, Serdyukov, & de Rijke, 2013; Schuth, Sietsma, Whiteson, Lefortier, & de Rijke, 2014) are all influenced by the issue of attractiveness bias, which is essentially what the click decision prediction task models. Therefore, if a set of features works for the click decision prediction task examined here, they have the chance to help solve similar issues in other occasions. This study, due to the lack of large-scale search log and user traffic for online experiments, did not further examine real applications.

Similar to Chapter 7, this chapter uses Gradient Boosted Regression Trees (GBRT) classifier (Friedman, 2001, 2002) for prediction. The purpose of the experiment is to examine whether or not session information can help prediction of users' click decisions. The rest of this section introduces and evaluates the features for prediction. Some of them are the same as the variables examined in the previous section.

8.3.1 Baseline Features

Table 68 introduces the baseline feature set. The table also reports the correlation (Pearson's r) of each feature with the dependent variable. The baseline feature set included many result-level variables examined in the last section, such as `length_title`, `length_snippet`, `URLslash`, `freq(q,title)`, `freq(q,snippet)`, and `freq(q,URL)`.

The table confirms that `length_snippet` does have a weak and negative correlation with click decision ($p < 0.05$)—this is consistent with Clarke et al.'s (2007) findings that searchers clicked on results with shorter snippets more frequently. However, as the last section also showed, `length_snippet` has no significant influence in the hierarchical logistic regression models. This indicates the limited additive value of `length_snippet` compared

Table 68: Baseline features (do not use session information) for predicting click decisions.

Feature	**Meaning**	**Pearson's *r***
`ARI_title`	The ARI index of the result's title.	0.031
`ARI_snippet`	The ARI index of the result's snippet.	−0.041
`CLI_title`	The CLI index of the result's title.	0.043
`CLI_snippet`	The CLI index of the result's snippet.	0.003
`LIX_title`	The LIX index of the result's title.	0.045
`LIX_snippet`	The LIX index of the result's snippet.	−0.033
`length_title`		−0.034
`length_snippet`		−0.099
`URLslash`		0.044
`.gov/.edu`	Whether the result is from a .gov or .edu domain.	0.019
`Homepage`	Whether the title contains "home page".	−0.044
`Official`	Whether the title contains official.	0.069
`freq(q,title)`		0.158
`freq(q,snippet)`		−0.012
`freq(q,URL)`		0.147

- Light, dark, and darker shadings indicate the correlations are significant at 0.05, 0.01, and 0.001 levels, respectively. Unshaded cells are not significant at 0.05 level.
- Unexplained features have the same meanings as the corresponding variables in Table 65.

Table 69: Session-related features for predicting click decisions.

Feature	**Meaning**	**Pearson's r**
`duration`		-0.014
`numclick`		0.073
`numq`		-0.029
`visited`		-0.030
`log P(pastq|title)`		0.059
`log P(pastq|snippet)`		0.062
`log P(clicked|title)`		0.101
`log P(clicked|snippet)`		0.062
`log P(clicked,R|title)`		0.102
`log P(clicked,R|snippet)`		0.059
`log P(clicked,NR|title)`		0.088
`log P(clicked,NR|snippet)`		0.088
`freq(ADD,title)`	Frequency of the added words in title.	0.094
`freq(ADD,snippet)`	Frequency of the added words in snippet.	-0.087
`freq(KEEP,title)`	Frequency of the retained words in title.	0.148
`freq(KEEP,snippet)`	Frequency of the retained words in snippet.	0.059
`freq(RMV,title)`	Frequency of the removed words in title.	0.011
`freq(RMV,snippet)`	Frequency of the removed words in snippet.	-0.019

- Light , dark , and darker shadings indicate the correlations are significant at 0.05, 0.01, and 0.001 levels, respectively. Unshaded cells are not significant at 0.05 level.
- Unexplained features have the same meanings as the corresponding variables in Table 65.

with other variables in explaining the variance of click decisions. Similarly, the table shows that `freq(q,URL)` also has a weak and positive correlation with click ($p < 0.001$), although it did not show any significant influence in the hierarchical logistic regression models. This is probably due to the high correlation between `freq(q,URL)` and `freq(q,title)` ($r = 0.72$).

In addition, this study examines many other possible features for this task, yet none of them showed any significant correlation with click. This study have also examined their value to the prediction task using incremental feature selection, where `length_snippet`, `freq(q,title)`, and `freq(q,URL)` are used as the base set of features. Still, none of them were selected (had any improvements over the base set). The examined features include:

- readability of the result's title and snippet, as measured by ARI, CLI, and LIX indexes
- whether or not the result's URL comes from the .gov or .edu domain
- whether or not the title contains the word "official" (Clarke et al., 2007)
- whether or not the title contains the phrase "home page" (Clarke et al., 2007)

The final baseline feature set only included three features: `length_snippet`, `freq(q,title)`, and `freq(q,URL)`.

8.3.2 Session Features

Table 69 introduces the set of features using session-level information. In addition to the variables examined in the hierarchical logistic regression models, the session features also included query-reformulation features:

- The added (ADD), retained (KEEP), and removed (RMV) words are extracted from the query reformulation from the previous query to the current query. The frequencies of these words in the result's title and snippet are used as features, e.g., `freq(ADD,title)` refers to the sum of the frequencies of ADD words in the title. As the table shows, `freq(ADD,title)`, `freq(ADD,snippet)`, and `freq(KEEP, title)` show significant correlations with click decision, suggesting that query reformulation in a session may be helpful for predicting click decisions. Note that the three features were not included into the last section's hierarchical logistic regression models for analysis because they are highly correlated with `freq(q, title)` and `freq(q, snippet)`. Here this section also does not hope to overstate any of their influence on users' click decisions. But the significant correlations suggest they may be useful prediction features.

The final session-related feature set included all the features with a significant correlation at 0.05 level in Table 69.

8.3.3 Experiment Settings

Similar to previous sections, the experiment in this section generates 10 random partitions of the dataset. On each partition, a 10-fold cross validation was performed to evaluate the effectiveness of an approach (a set of features), where 7 folds were used for training, 2 folds were used for validation, and 1 fold was used for testing. All evaluated approaches used the same 10 partitions and cross-validation settings.

The effectiveness of the prediction is evaluated using the precision, recall, and F1 of both positive and negative cases, as well as the average F1 and accuracy. For each approach (feature set), the experiment yields results on 100 test folds in total (10 test folds for each partition, and 10 random partitions). The following sections report the mean values of the measures on the 100 test folds. Differences between approaches are tested by whether their mean values of metrics on the 100 test folds are significantly different using paired *t*-test.

8.3.4 Results

Table 70 reports the results of the experiment. As the table shows, the baseline features have the highest precision in predicting clicked results, but the recall is very limited (0.15). The session features, alone, have relatively lower precision, but higher recall compared with the baseline features. The session features alone also significantly outperforms the baseline features in F1 on the clicked results ($p < 0.05$). After combining both sets of features, the prediction model ("All") significantly outperforms both the baseline and the session features in recall, and consequently in F1 and overall performance (average F1). This suggests that the session features are complementary to the baseline features in predicting users' click decisions. Overall results indicate that session features and existing ones are complementary

Table 70: Effectiveness of prediction models for click decisions.

Feature	**Avg F1**	**Accuracy**	**Click**			**Skip**		
			F1	**P**	**R**	**F1**	**P**	**R**
Baseline	0.54 **	**0.73**	0.24 **	**0.61** **	0.15 **	**0.83**	0.74	**0.96** *
Session	0.55 *	0.72	0.28 *	0.55	0.18 *	**0.83**	0.74	0.94
All	**0.59**	**0.73**	**0.35**	0.56	**0.26**	**0.83**	**0.75**	0.92

- Reported values are the mean values on the 100 test folds.
- *, **, and *** indicate the value is significantly different from "All" at 0.05, 0.01, and 0.001 level, respectively, by paired t-test on the 100 test folds.

and helpful to each other in modeling users' click decisions.

8.4 ANSWERS TO RQ4

This chapter examined the influence of session search history on users' click decisions in a session – whether or not searchers would click on a result after viewing its summary. In addition, it also examined whether or not session-level information can help the prediction of users' click decisions. Results answered the research question and verified the hypotheses.

- **RQ4** – What affects users' decisions on whether or not to click on a search result's link after viewing its summary displayed on a SERP? Can we predict such decisions?

The analysis in this section confirmed a few previous findings, including: 1) the relevance of the result has a positive influences on users' decisions to click on the result; 2) the rank of the result also has a positive influence (users prefer to click on higher-ranked results), which is independent of result examination; 3) the occurrences of the query terms in result title also have a positive influence; and 4) visiting a result reduces the chances of revisiting the result in a search session.

In addition, analysis also disclosed a few session-related factors that may influence users' click decisions. These factors include: 1) the overall frequency of a searcher to click on results; and 2) the similarity of the result's title with previously clicked relevant results' titles. Moreover, experimental results show that session-level features can help existing ones (without using session information) in predicting searchers' click decisions.

- [**H4.1**] – Click decision is related to the task's product and goal.

As Section 8.2 discussed, task product and goal did not show any significant influence on users' click decisions.

- [**H4.2**] – Click decision is related to the searcher's characteristics.

The analysis in Section 8.2 showed that searchers' overall frequencies of result clicking may influence their click decisions in a specific session.

- **[H4.3]** – Click decision is related to the time of the session when the decision was made.

As Section 8.2 discussed, the examined variables indicating the time of the session did not show any significant influence on users' click decisions.

- **[H4.4]** – Click decision is related to previous search queries in the session.

As Section 8.2 discussed, variables measuring the similarity between the result and previous queries did not show any significant influence on users' click decisions.

- **[H4.5]** – Click decision is related to previously clicked results in the session.

As Section 8.2 discussed, results showed that searchers' click decisions are influenced by the correct click decisions they made in previous searches of the session—once they clicked on a result and found it was relevant, they are more likely to click on results with similar titles in follow-up searches. In contrast, the do not "learn" much from failed click decisions.

9.0 DISCUSSION AND CONCLUSION

The past decade witnessed the flourish of contextual and personalized retrieval techniques. Specifically, implicit feedback and behavioral modeling techniques have achieved huge success in commercial search engines. Nowadays search engines are generally capable of consistently and constantly improving search quality based on information from search logs and large user traffic. Search engines also have become more intelligent in providing much help for user interaction, such as query suggestion, query auto-completion, and so on.

We studied relevance and user activities in a search session. We consider relevance as a dynamic and contextual variable that depends on various factors, including the searcher, the result, the context (more specifically, a search session in the case of this dissertation), the actual interaction between the user and the result, and the task. This idea is not new, but previous studies had limited empirical analysis. In addition, we also examined the influence of a search session on two important search activities—query reformulation and click.

Our study offers a series of insights to current understandings of relevance and user activities. We also provide practical techniques for addressing a few critical IR challenges. This chapter summarizes these insights and contributions from different aspects.

9.1 RELEVANCE JUDGMENT AND IR EVALUATION

Evaluation method is a fundamental research area of information retrieval (IR). Unlike many other subfields of artificial intelligence, IR addresses highly dynamic problems because the ultimate goal of IR systems is to satisfy real users. The variability of user, information need, and search behavior makes IR evaluation intrinsically difficult and complex. In the past two decades, test collection-based IR evaluation methods (Sanderson, 2010) such as "the TREC approach" (Harman, 1992a; Hawking et al., 1999; Collins-Thompson et al., 2014) is the most popular one applied in IR research. Unfortunately, the method has limited agreements with user experience (Turpin & Scholer, 2006; Huffman & Hochster, 2007; Sanderson, Paramita, Clough, & Kanoulas, 2010; Jiang & Allan, 2016a, 2016b), making many search engine companies switch to use online evaluation techniques such as interleaved experiment (Chapelle, Joachims, Radlinski, & Yue, 2012; Radlinski & Craswell, 2013) and user experience prediction (Hassan, Jones, & Klinkner, 2010; Jiang, Hassan Awadallah, Shi, & White, 2015; Jiang, Hassan Awadallah, Jones, et al., 2015; Kiseleva et al., 2016). However, comparing to online evaluation methods, test collection-based evaluation still has the advantage of repeatability, which also makes it appropriate for automatic system optimization.

Our study offers insights for improving the status quo of test collection-based IR evaluation from the aspect of relevance judgment. Relevance judgment is a core component of test collection-based IR evaluation methods. The current de facto standard approach for relevance judgment asks external annotators to assess a list of preassigned search results one after another, without a real search context. In addition, most current work judged relevance using criteria that focus on topical relevance. We believe the discrepancy between the actual useful information acquired by the searcher and the current context-independent and topical relevance judgment is a critical limitation. We examined two alternatives to current relevance judgment method (TRel). The context-independent usefulness judgment (Usef) replaces the judgment criteria from topicality to usefulness. Ephemeral relevance (ERel) further takes into account context in relevance judgment, which is essentially a contextual usefulness judgment.

We evaluated both ERel and Usef and compared with TRel by correlating with user experience measures at six dimensions. The results offer suggestions to which type of relevance judgment should be collected in test collection-based IR evaluation (for the purpose of serving as a faithful indicator to users' search experience).

- Our study suggests that moving from topicality to usefulness as the main criterion for relevance judgment is fruitful. We showed the collected context-independent usefulness judgments (Usef) yield consistently stronger correlations with user experience measures than TREC-style relevance judgments (TRel) collected in the same setting (without a search context). This suggests future IR evaluation method can improve its faithfulness to user experience without any additional cost by simply switching to adopt a different relevance judgment criterion—usefulness.

- We also found that collecting usefulness judgment in a contextual manner (ephemeral relevance) can further yield slightly better correlations with many user experience measures than context-independent usefulness judgment (Usef). However, we also note that collecting contextual relevance/usefulness judgments requires a significant high cost and a more complex setting. It remains an open question whether or not it is worthwhile to collect relevance/usefulness judgments in a contextual manner. We suggest future studies perform detailed cost-benefit analysis to examine the trade-off between relevance judgment quality and cost.

9.2 EPHEMERAL RELEVANCE AND INFLUENCING FACTORS

Relevance is a key notion of information retrieval (IR). Most current search systems are designed and optimized to rank results based on relevance, and almost all offline methods for evaluating search systems are based on relevance judgments. Previous studies had numerous discussions on the notion of relevance (Borlund, 2003; Mizzaro, 1997; Saracevic, 1975, 2007, 1996) and its measurement (Xu & Chen, 2006; Xu & Wang, 2008; Y. Zhang et al., 2014). Most of these studies acknowledged that relevance depends on not only topic aboutness but also many other factors such as novelty, understandability, reliability, search task, search context, and users' interaction with the search results.

We designed a laboratory user study and collected ephemeral relevance (ERel) judgments—essentially ERel is a contextual usefulness judgment—as well as many other complementary search result assessments for explaining relevance/usefulness in a contextual setting. Our study offers empirical insights for understanding the influencing factors of relevance/usefulness as well as the dynamics of relevance/usefulness in a search session.

First, we offer new understandings regarding how users assess usefulness in contextual and context-independent settings and the influencing factors of usefulness. Similar to previous studies of relevance criteria, we found that both contextual and context-independent usefulness judgment criteria include at least five important factors: topicality, novelty, understandability, reliability, and scope. We also show that the collected contextual and context-independent usefulness judgment are significantly link to users' judgments of related factors, confirming the user-rated usefulness judgment criteria. We also found that users' criteria of usefulness judgment undergo slight but significant changes in contextual and context-independent settings, suggesting a subtle connection between relevance/usefulness criteria and the judgment context.

Second, our study discloses the dynamics of relevance/usefulness in a search session. We show that users' perceptions regarding the understandability and reliability of a search result undergo changes and the differences of usefulness in contextual and context-independent settings are significantly linked to these changes. We have also identified search effort as a significant influencing factor for contextual usefulness judgment (ephemeral relevance). These findings suggest that relevance is not a static and context-independent attribute of the document, but a changing state regarding both the result, the user, and the context.

Third, our analysis regarding how ephemeral relevance and effort change by different levels of understandability and reliability (Figure 8) disclosed the complex process of acquiring relevance information from a result. We believe the examples of understandability and reliability imply more principled mechanism of how users acquire useful information from an information object. We propose a hypothesis for the process as in Figure 9 to interpret the empirical observations in Figure 8:

- Figure 9: x-axis – The acquired amount of useful information and the effort spent on a result depend on not only how much useful information the result contains but also the efficiency of acquiring useful information from the information object (result).
- Figure 9: left side – When the efficiency of acquiring useful information declines, users will first try to spend more effort to compensate the limited efficiency, such that the amount of useful information acquired from the result still maintains at a certain level.
- Figure 9: right side – However, when the efficiency declines below certain threshold, users will abandon examining the result to avoid wasting effort. For example, users may

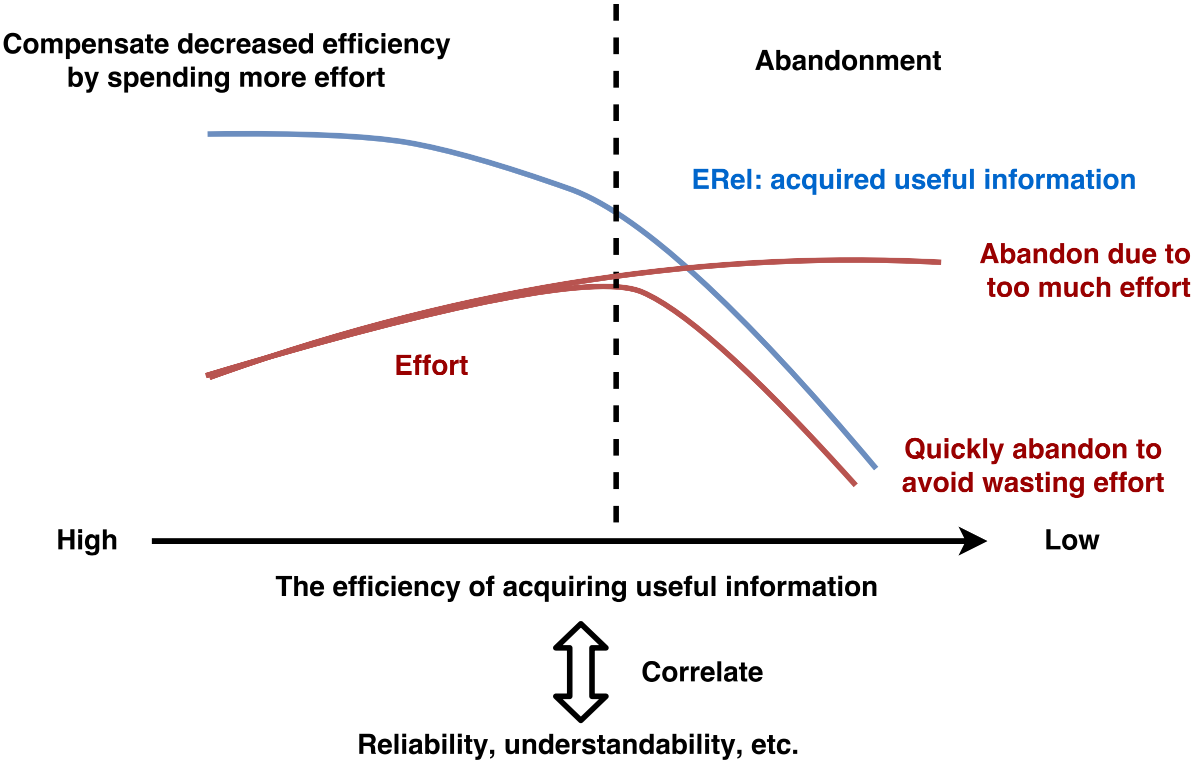

Compensate decreased efficiency by spending more effort
Abandonment
ERel: acquired useful information
Abandon due to too much effort
Effort
Quickly abandon to avoid wasting effort
High
Low
The efficiency of acquiring useful information
Correlate
Reliability, understandability, etc.

Figure 9: The adaptive process of acquiring useful information from information objects with different levels of efficiency.

quickly identify it is not worthwhile to continue spending more effort on the result.

- The efficiency of acquiring useful information from an object (result) relates to factors such as the understandability and reliability of the results. For example, users need to spend more effort if the result is difficult to understand. They also need to spend more effort on the less reliable results such as for the purpose of confirming the credibility of information.

The empirical observations in Figure 8 fit well with the hypothesis. Here we do not have the resource to fully verify this hypothesis, but we believe it may potentially offer a new understanding to users' interaction with the search results.

9.3 IMPLICIT FEEDBACK TECHNIQUES

Nowadays, implicit feedback techniques (Joachims, 2002; Joachims et al., 2005, 2007; Kelly & Teevan, 2003) have been widely applied to commercial search engines in all aspects of information retrieval (Joachims, 2002; Joachims et al., 2005, 2007; Kelly & Teevan, 2003; Agichtein, Brill, & Dumais, 2006; Agichtein, Brill, Dumais, & Ragno, 2006; Craswell et al., 2008; Richardson et al., 2007; Hassan et al., 2010). The success of implicit feedback techniques is one of the primary force for the development of IR systems in the past decade.

Our study also makes a practical contribution by providing effective implicit feedback techniques for predicting contextual usefulness judgments (ephemeral relevance). We found that we can predict ephemeral relevance using a wide range of features. We also separately examined two different tasks for predicting ephemeral relevance, which offer insights to potential applications in search result ranking and search engine evaluation, respectively:

- We show that ephemeral relevance can be predicted with a reasonable level of accuracy (about 0.3 correlation between the predicted and actual ERel values) using a wide range of features covering search result content, structure, and previous interactions. This prediction task offers practical insights on how to help search engines rank results by ephemeral relevance—considering ERel better correlates with user experience than conventional relevance judgment measures such as TRel, ranking search results by ERel may potentially improve users' search experience as well. The effectiveness of the prediction techniques suggest it is possible to rank search results by ERel on the fly based on the result itself and short-term search history. In addition, similar to many previous studies, we also find that a contextual prediction model (using both the information of the result and short-term search history) outperformed a baseline one that without using context information. This suggests that contextual search techniques are also helpful for IR applications targeting ephemeral relevance.

- We also show that ephemeral relevance can be predicted accurately using implicit feedback information, achieving as high as over 0.5 correlation between predicted and real values. This suggests that ephemeral relevance judgments can be predicted accurately from search logs without human judgments. Similar to previous studies using implicit feedback for inferring document relevance labels, the implication of the prediction experiment is twofold: first, we can automatically generate empheral relevance judgments from search logs; second, we can also aggregate predicted ERel judgments of results in past search logs as features to rank search results in the future. We have also make an interesting technical contribution by extending implicit feedback signals from those related to only a search result itself (such as click dwell time) to those in a search session (such as the relationship of a clicked result with follow-up searches and queries in the session). We observed that the set of generalized implicit feedback signals produced better prediction results. This method may also have potential to generalize to other implicit feedback applications.

9.4 QUERY REFORMULATION

Query reformulation is a widely studied topic in information retrieval. However, most previous studies only focused on query-level query reformulation patterns (Rieh & Xie, 2006; Huang & Efthimiadis, 2009) rather than particular word changes (Jiang & Ni, 2016). Using both hierarchical logistic regression and qualitative analysis, we provide insights on how different factors may affect specific choices of word changes in query reformulations during a task-based search session. This advances the state-of-the-art understanding of query reformulation from query-level patterns to word-level and finer-grained users decisions.

Results suggest that word-level variables may strongly influence all three types of word changes. This is not surprising because word-level variables are word-specific, while other two types of factors are not. In contrast, query-level variables only show certain influence on removing or retaining a word, but limited influence on adding and reusing a word. This indicates that removing or retaining a word may more likely be affected by past user interaction and situation in a session, while adding and reusing a word may not. Session-level variables exhibited limited direct influence on all types of word changes.

Comparing the three categories of factors, word-level factors show the strongest influence on all three types of word changes compared with the other two types of factors. Query-level factors show less influence, and session-level factors only exhibit limited direct influence. Results also suggest that these variables may influence different word changes in different ways. For example, word occurrences in result summaries show different patterns for all three types of word changes. This implies the different nature of these word changes. In addition, our analysis also help identify effective sets of features for predicting such word changes in an on-going search session.

Moreover, our study also discloses scenarios for different word changes. Users remove a word in query reformulation for two possible reasons. Firstly, it happens when users finished exploring a subtask and move forward to another. Secondly, it also happens when users try to correct bad-performing queries. In our collected data, the former is more prevalent. In contrast, adding a new word usually happens in relevance feedback—users exploit related, unused words from recently viewed result summaries and add them to new queries. Reusing a word mostly happens when searchers revert back from a subtask to the main task, and the reused words are highly related to the main theme of the task.

9.5 CLICK DECISION AND BIAS

Click behavior (Joachims et al., 2005; Craswell et al., 2008; Y. Yue et al., 2010) is one of the most widely studied search behavior, probably because click is the most important implicit feedback signal. Previous studies examined many factor that may influence click behavior, including position bias (Joachims et al., 2005), presentation bias (Y. Yue et al., 2010), trust bias (Joachims et al., 2005), and so on. However, most of these studies restrict their scope to a single query. We examined click decision in a search session using eye-movement data, which adds a few insights to current understandings of click behavior.

One of the most important finding related to our study of click decision lies in that we identified a new type of presentation bias in a search session. We found that searchers are likely to be influenced by the contents of previous clicks when they made new click decisions. For specifically, click on a relevant result in the past encourages users to click on results with similar information in the future. This finding seems minor compared with existing studies regarding the influencing factors of click (such as position bias, presentation bias of the current result), but we believe we make a substantial contribution considering previous studies have already had exhaustive research regarding click behavior and bias.

This newly identified presentation bias also offers insights to a few important IR applications. First, our finding suggests that future click models should take into account presentation bias factors in a search session to estimate more reliable “unbiased” click probabilities. Second, our finding also suggests that interleaved experiments should either account for such bias in experiemnt design or result analysis, or avoid long search sessions (to exclude such presentation bias).

9.6 LIMITATION

Despite many new insights and contributions, we acknowledge that our study also has many limitations. First, our study is entirely based on laboratory studies, which is limited in scale and may fail to replicate real scenarios. Second, we note that our experiment setting, especially the search task setting in both User Study 1 and 2, is not fully representative of regular web search information needs. The employed tasks are typically more complex than regular web search (such as navigational search). Third, we also note that the process of collecting contextual search result judgments in User Study 1 more or less has some impacts on user's search behavior, which consequently affects the results related to both RQ1 and

RQ2. In our study, it is difficult to assess the influence of these impacts. We suggest future researchers reexamine these issues when replicating or generalizing our work.

APPENDIX A

SEARCH TASKS

This appendix includes the description of tasks adopted in User Study 1 and 2. All tasks come from the TREC session track 2012 (Kanoulas et al., 2012) and 2013 (Carterette et al., 2013) except the training tasks. The Year and Topic ID correspond to the original topic ID in the specified year of the TREC session track.

A.1 TASK ASSIGNMENTS

Table 71 shows the tasks assigned to each group of users in User Study 1.

A.2 TRAINING TASKS

In User Study 1, the following training task was used for the participants from Simmons College.

> Who is the provost of Simmons College? Please find out the answer.

In User Study 1, the following training task was used for the participants from the University of Massachussets Amherst.

> Who is the provost of UMass? Please find out the answer.

Table 71: User Study 1 – the assigned task numbers for each group of participants.

Group	**F+S**	**F+A**	**I+S**	**I+A**
1	2012, No.1	2012, No.14	2012, No.18	2012, No.20
2	2012, No.11	2012, No.22	2012, No.2	2012, No.29
3	2012, No.15	2012, No.3	2012, No.38	2012, No.10
4	2012, No.30	2012, No.33	2012, No.41	2012, No.37
5	2012, No.23	2012, No.4	2012, No.48	2012, No.46
6	2012, No.32	2012, No.40	2012, No.7	2012, No.5
7	2013, No.33	2013, No.43	2013, No.51	2013, No.63

- "F" and "I" stand for tasks that generate *factual* and *intellectual* products.
- "S" and "A" stand for tasks with *specific* and *amorphous* goals.
- Task number refers to the topic ID in the TREC session track. For example, "2012, No.32" refers to the topic No.32 in the TREC 2012 session track.

Table 72: User Study 2 – the assigned task numbers for each group of participants.

Group	**F+S**	**F+A**	**I+S**	**I+A**
1	2012, No.1	2012, No.14	2012, No.18	2012, No.20
2	2012, No.11	2012, No.22	2012, No.2	2012, No.29
3	2012, No.15	2012, No.3	2012, No.38	2012, No.10
4	2012, No.30	2012, No.33	2012, No.41	2012, No.37
5	2012, No.23	2012, No.4	2012, No.48	2012, No.46
6	2012, No.32	2012, No.40	2012, No.7	2012, No.5

- "F" and "I" stand for tasks that generate *factual* and *intellectual* products.
- "S" and "A" stand for tasks with *specific* and *amorphous* goals.
- Task number refers to the topic ID in the TREC session track. For example, "2012, No.32" refers to the topic No.32 in the TREC 2012 session track.

In User Study 2, the following training task was used for all participants (the experiments were conducted at the University of Pittsbrugh).

> Does Pitt have fitness centers still open at 11pm? If not, you want to find some alternatives in the Shadyside area so that you can work out every night at 11pm.

A.3 FORMAL TASKS

Topic 01 (Factual+Specific)

You are writing a summary article about US tax code 403(b) retirement plans. Find as many relevant documents as you can that would help you in writing the summary. Aspects might include eligibility for a 403(b), tax benefits of 403(b) plans, the types of institutions that offer them to employees, withdrawal rules, contribution limits, instructions for rolling over into another retirement plan, and so on.

Topic 02 (Informational+Specific)

Suppose you are an employee at a US non-profit organization. You are trying to decide whether to contribute money to your employer-offered 403(b) retirement plan or to a personal IRA (Individual Retirement Account). Find resources that could help you decide which is best for your needs.

Topic 03 (Factual+Amorphous)

You have decided to teach yourself computer programming. Find some resources to help you begin. What languages might be good to start with (what are the pros and cons of various languages for a beginner)? For those langauges you might start with, what introductory resources are available? What software would you need to install on your computer to be able to practice? And so on.

Topic 04 (Factual+Amorphous)

You are writing a summary article about the Pocono Mountains region. Find as many relevant articles as you can describing the region, things to see and do there (such as national parks, resorts, shopping, etc), and communities of people living there.

Topic 05 (Informational+Amorphous)

You are planning a winter vacation to the Pocono Mountains region in Pennsylvania in the US. Where will you stay? What will you do while there? How will you get there?

Topic 07 (Informational+Specific)

A "designer dog" is a cross between two purebred dogs of two different breeds. "Designer" dogs are sometimes bred for practical purposes, but critics say they are more often bred as status symbols. Find resources describings pros and cons of the practice of crossing purebred dogs.

Topic 10 (Informational+Amorphous)

Lara Dutta of India was crowned Miss Universe in 2000, and between 1994 and 2000 women from India won two Miss Universe competitions, four Miss World competitions, and many less well-known competitions. To what extent can decisions and policies of the Indian government be credited with these wins?

Topic 11 (Factual+Specific)

Where is Port Arthur? When did the massacre occur? What was the final death toll of the massacre? Who was the killer? What was the killer's nationality? What were the names of the victims? What were the nationalities of the victims?

Topic 14 (Factual+Amorphous)

France won its first (and only) World Cup in 1998. Find information about the reaction of French people and institutions (such as stock markets), and studies about these reactions.

Topic 15 (Factual+Specific)

Where is Bollywood located? From what foreign city did Bollywood derive its name? What is the Bollywood equivalent of Beverly Hills? What is Bollywood's equivalent of the Oscars? Where does Bollywood rank in the world's film industries? Who are some of the Bollywood stars?

Topic 18 (Informational+Specific)

Merck and Co. is one of the largest pharmaceutical companies in the world, and a major lobbyist in Washington D.C. Find information about specific legislation or policies Merck has lobbied for or against. How has Merck's lobbying affected US policy?

Topic 20 (Informational+Amorphous)

Akira Kurosawa (1910-1998) was one of the most renowned Japanese film directors.

What are his lasting influences on Western cinema, Indian cinema, and Chinese and Japanese cinema?

Topic 22 (Factual+Amorphous)

Hydropower is considered one of the renewable sources of energy that could replace fossil fuels. Find information about the efficiency of hydropower, the technology behind it and any consequences building hydroelectric dams could have to environment.

Topic 23 (Factual+Specific)

What is pseudocyesis? What are the demographics of this condition? What is the earliest report on it? What are the causes according to psychology studies and biological studies?

Topic 29 (Informational+Amorphous)

You would like to write a report about interesting weddings traditions of different cultures, religions, and ethnic groups. Find information about wedding ceremonies that you think are the most fascinating and different than what you are used to.

Topic 30 (Factual+Specific)

What are the sunspots? Since when have they been observed? Does their numbers and locations follow any pattern or are they simply random? How do they affect the climate on earth? Has such an effect been discovered in the past? What is the average cycle of a sunspot activity?

Topic 32 (Factual+Specific)

What is depression? What are the major symptoms of depression? What medications, therapies and other treatments can be used to treat depression symptoms? Who performs therapy and what are the costs? Does health insurance pay for any of the treatments?

Topic 33 (Factual+Amorphous)

You think that one of your friends may have depression, and you want to search information about the depression symptoms and possible treatments.

Topic 37 (Informational+Amorphous)

You would like to buy a dehumidifier. On what basis should you compare different dehumidifiers?

Topic 38 (Informational+Specific)

You would like to buy a dehumidifier. You want to know what makes a dehumidifier good value for money.

Topic 40 (Factual+Amorphous)

One of your friends from Kenya invited you to attend a party in his house and have a taste of traditional swahili dishes. You would like to search and find some information about Swahili dishes.

Topic 41 (Informational+Specific)

A friend from Kenya is visiting you and you'd like to surprise him with by cooking a traditional swahili dish. You would like to search online to decide which dish you will cook at home.

Topic 46 (Informational+Amorphous)

A friend of yours was recently diagnosed with a type of collagen vascular disease. You want to find out more about the condition and what impact it may have on your friend's lifestyle, particularly whether they can continue to play sport.

Topic 48 (Informational+Specific)

You live in Connecticut and are considering a career move and being in the fire service interests you. You want to find out more about the Connecticut Fire Academy, what's involved in the training, where you need to go to attend training sessions and what kinds of skills you will develop during the programme

APPENDIX B

QUESTIONS FOR THE IMPORTANCE OF DIFFERENT FACTORS IN EREL AND TOPICAL RELEVANCE JUDGMENTS

[**Instruction for answering ERel judgments criteria**] You may still remember—each time you visited a web page and switched back to our system, we asked you the question "*How much useful information did you get from this webpage?*" Please weigh the importance of the following factors when you answered this question.

[**Instruction for answering topical relevance judgments criteria**] In the previous survey, we asked you to take a look at the web pages again and answer the question "How much useful information does this web page provide for the task?" Please weigh the importance of the following factors when you answered this question.

The following 15 statements are adapted from the questions in Xu and Chen's (2006) work. All the following statements were rated for their importance using a 7-point scale, from *not at all important* (1) to *very important* (7). The sequence of the statements were displayed in a random order.

[**TOP1**] The main content of the web page describes the task topic.

[**TOP2**] The web page is within the general domain of the task topic.

[**TOP3**] The subject area of the web page is related to the task topic.

[**NOV1**] The web page provides new information to me.

[**NOV2**] The web page describes information that I do not know.

[**NOV3**] The web page provides unique information that I am coming across for the first time.

[UND1] The content of the web page is easy for me to understand.

[UND2] I am able to follow the content of this web page with little effort.

[UND3] I find the web page easy to read.

[REL1] I think the content of the web page would be consistent with facts.

[REL2] I think the content of the web page would be reliable.

[REL3] I think the content of the web page would be likely true.

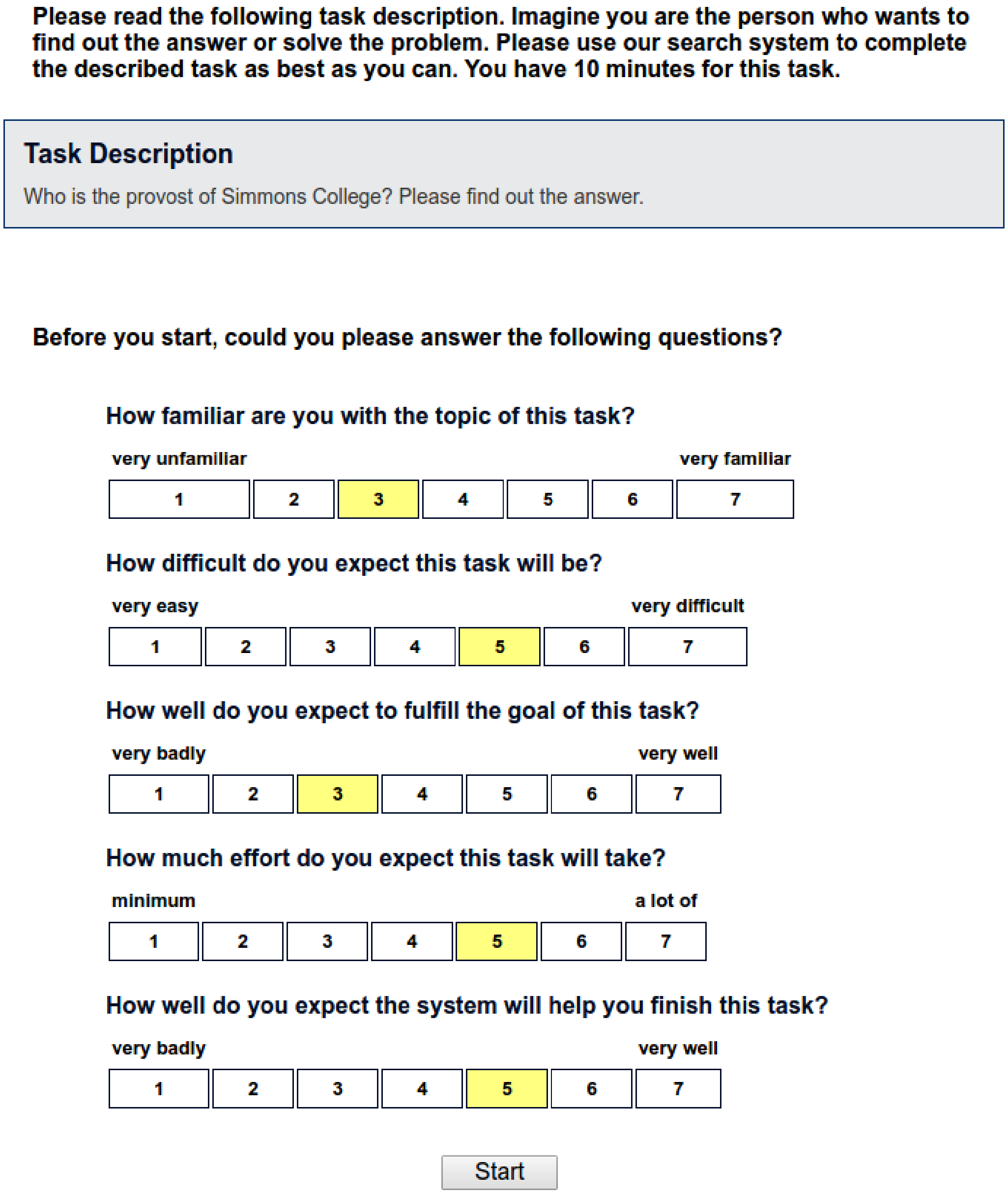
[SCP1] The scope of the web page is too general or too specific for me.

[SCP2] The scope of this web page is too broad or too narrow for me.

[SCP3] This web page gives too many or too few details for me.

APPENDIX C

USER STUDY 1: EPHEMERAL RELEVANCE – SCREENSHOTS OF THE EXPERIMENTAL SYSTEM



Please read the following task description. Imagine you are the person who wants to find out the answer or solve the problem. Please use our search system to complete the described task as best as you can. You have 10 minutes for this task.

Task Description

Who is the provost of Simmons College? Please find out the answer.

Before you start, could you please answer the following questions?

How familiar are you with the topic of this task?

very unfamiliar — very familiar

1 2 [✓ 3] 4 5 6 7

How difficult do you expect this task will be?

very easy — very difficult

1 2 3 4 [✓ 5] 6 7

How well do you expect to fulfill the goal of this task?

very badly — very well

1 2 [✓ 3] 4 5 6 7

How much effort do you expect this task will take?

minimum — a lot of

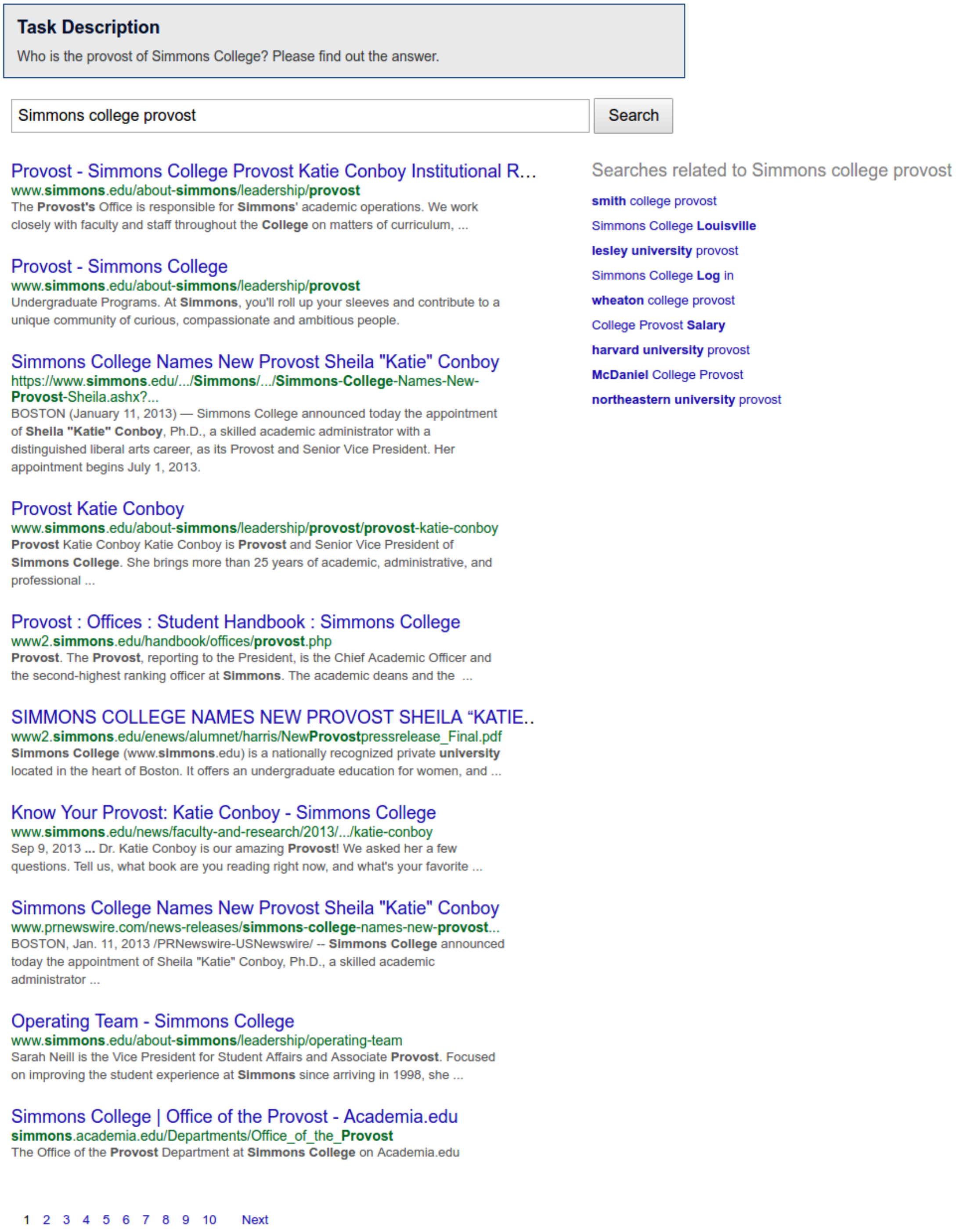
1 2 3 4 [✓ 5] 6 7

How well do you expect the system will help you finish this task?

very badly — very well

1 2 3 4 [✓ 5] 6 7

Figure 10: User Study 1: Ephemeral Relevance – a screenshot for the pre-task survey.



Task Description

Who is the provost of Simmons College? Please find out the answer.

Simmons college provost

Provost - Simmons College Provost Katie Conboy Institutional R...
www.**simmons**.edu/about-**simmons**/leadership/**provost**
The **Provost's** Office is responsible for **Simmons**' academic operations. We work closely with faculty and staff throughout the **College** on matters of curriculum, ...

Provost - Simmons College
www.**simmons**.edu/about-**simmons**/leadership/**provost**
Undergraduate Programs. At **Simmons**, you'll roll up your sleeves and contribute to a unique community of curious, compassionate and ambitious people.

Simmons College Names New Provost Sheila "Katie" Conboy
https://www.**simmons**.edu/.../**Simmons**/.../**Simmons-College**-Names-New-**Provost**-Sheila.ashx?...
BOSTON (January 11, 2013) — Simmons College announced today the appointment of **Sheila "Katie" Conboy**, Ph.D., a skilled academic administrator with a distinguished liberal arts career, as its Provost and Senior Vice President. Her appointment begins July 1, 2013.

Provost Katie Conboy
www.**simmons**.edu/about-**simmons**/leadership/**provost**/**provost**-katie-conboy
Provost Katie Conboy Katie Conboy is **Provost** and Senior Vice President of **Simmons College**. She brings more than 25 years of academic, administrative, and professional ...

Provost : Offices : Student Handbook : Simmons College
www2.**simmons**.edu/handbook/offices/**provost**.php
Provost. The **Provost**, reporting to the President, is the Chief Academic Officer and the second-highest ranking officer at **Simmons**. The academic deans and the ...

SIMMONS COLLEGE NAMES NEW PROVOST SHEILA "KATIE..
www2.**simmons**.edu/enews/alumnet/harris/New**Provost**pressrelease_Final.pdf
Simmons College (www.**simmons**.edu) is a nationally recognized private **university** located in the heart of Boston. It offers an undergraduate education for women, and ...

Know Your Provost: Katie Conboy - Simmons College
www.**simmons**.edu/news/faculty-and-research/2013/.../katie-conboy
Sep 9, 2013 **...** Dr. Katie Conboy is our amazing **Provost**! We asked her a few questions. Tell us, what book are you reading right now, and what's your favorite ...

Simmons College Names New Provost Sheila "Katie" Conboy
www.prnewswire.com/news-releases/**simmons-college**-names-new-**provost**...
BOSTON, Jan. 11, 2013 /PRNewswire-USNewswire/ -- **Simmons College** announced today the appointment of Sheila "Katie" Conboy, Ph.D., a skilled academic administrator ...

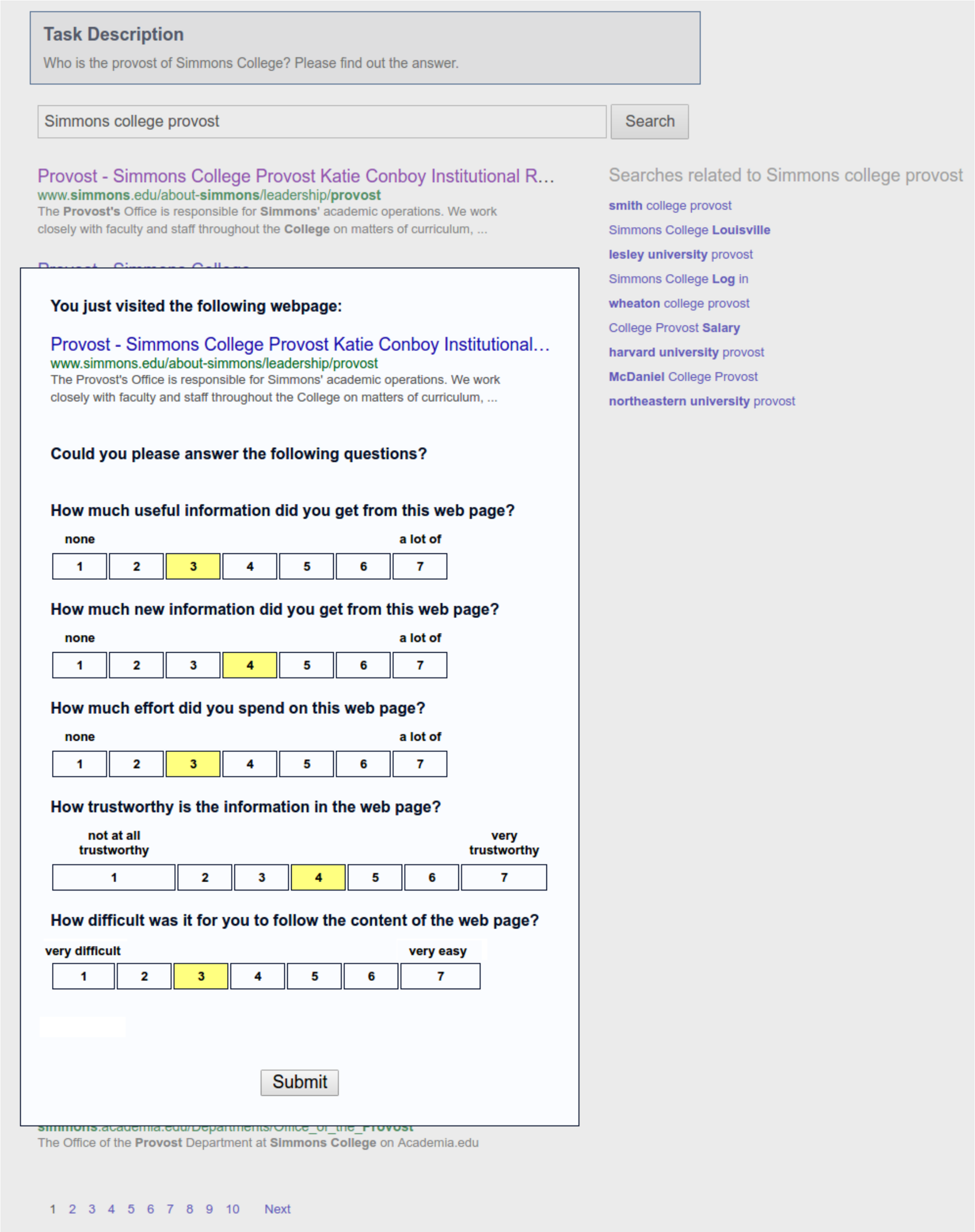
Operating Team - Simmons College
www.**simmons**.edu/about-**simmons**/leadership/operating-team
Sarah Neill is the Vice President for Student Affairs and Associate **Provost**. Focused on improving the student experience at **Simmons** since arriving in 1998, she ...

Simmons College | Office of the Provost - Academia.edu
simmons.academia.edu/Departments/Office_of_the_**Provost**
The Office of the **Provost** Department at **Simmons College** on Academia.edu

Searches related to Simmons college provost

smith college provost
Simmons College **Louisville**
lesley university provost
Simmons College **Log** in
wheaton college provost
College Provost **Salary**
harvard university provost
McDaniel College Provost
northeastern university provost

Figure 11: User Study 1: Ephemeral Relevance – a screenshot for the search interface.



Task Description

Who is the provost of Simmons College? Please find out the answer.

Simmons college provost

Provost - Simmons College Provost Katie Conboy Institutional R...
www.**simmons**.edu/about-**simmons**/leadership/**provost**
The **Provost's** Office is responsible for **Simmons'** academic operations. We work closely with faculty and staff throughout the **College** on matters of curriculum, ...

Searches related to Simmons college provost

smith college provost
Simmons College **Louisville**
lesley university provost
Simmons College **Log** in
wheaton college provost
College Provost **Salary**
harvard university provost
McDaniel College Provost
northeastern university provost

You just visited the following webpage:

Provost - Simmons College Provost Katie Conboy Institutional...
www.simmons.edu/about-simmons/leadership/provost
The Provost's Office is responsible for Simmons' academic operations. We work closely with faculty and staff throughout the College on matters of curriculum, ...

Could you please answer the following questions?

How much useful information did you get from this web page?

none — a lot of

1 2 [✓ 3] 4 5 6 7

How much new information did you get from this web page?

none — a lot of

1 2 3 [✓ 4] 5 6 7

How much effort did you spend on this web page?

none — a lot of

1 2 [✓ 3] 4 5 6 7

How trustworthy is the information in the web page?

not at all trustworthy — very trustworthy

1 2 3 [✓ 4] 5 6 7

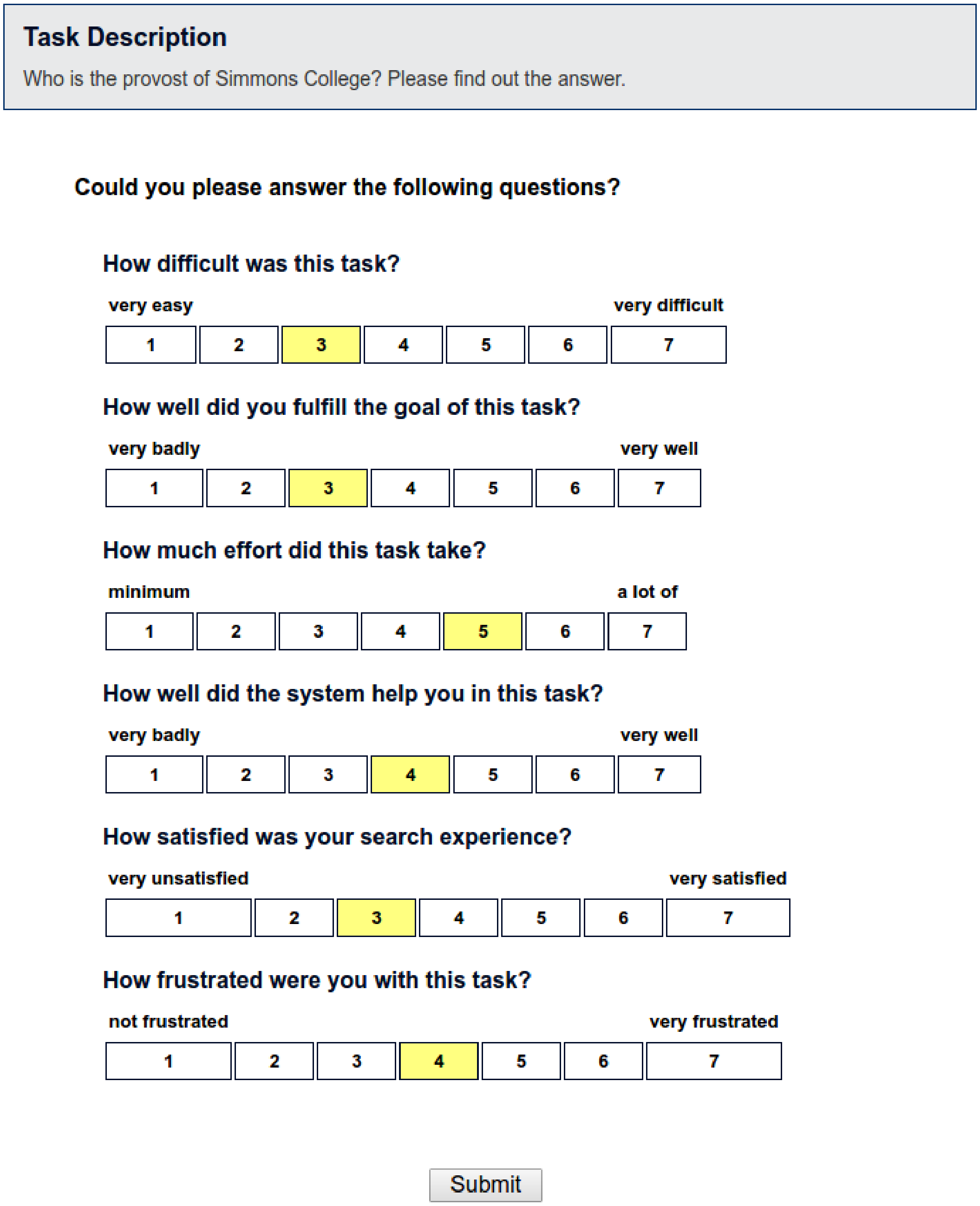
How difficult was it for you to follow the content of the web page?

very difficult — very easy

1 2 [✓ 3] 4 5 6 7

simmons.academia.edu/Departments/Office_of_the_Provost
The Office of the **Provost** Department at **Simmons College** on Academia.edu

Figure 12: User Study 1: Ephemeral Relevance – a screenshot for the contextual judgment after clicking on a result.



Task Description

Who is the provost of Simmons College? Please find out the answer.

Could you please answer the following questions?

How difficult was this task?

very easy — very difficult

1 2 [✓ 3] 4 5 6 7

How well did you fulfill the goal of this task?

very badly — very well

1 2 [✓ 3] 4 5 6 7

How much effort did this task take?

minimum — a lot of

1 2 3 4 [✓ 5] 6 7

How well did the system help you in this task?

very badly — very well

1 2 3 [✓ 4] 5 6 7

How satisfied was your search experience?

very unsatisfied — very satisfied

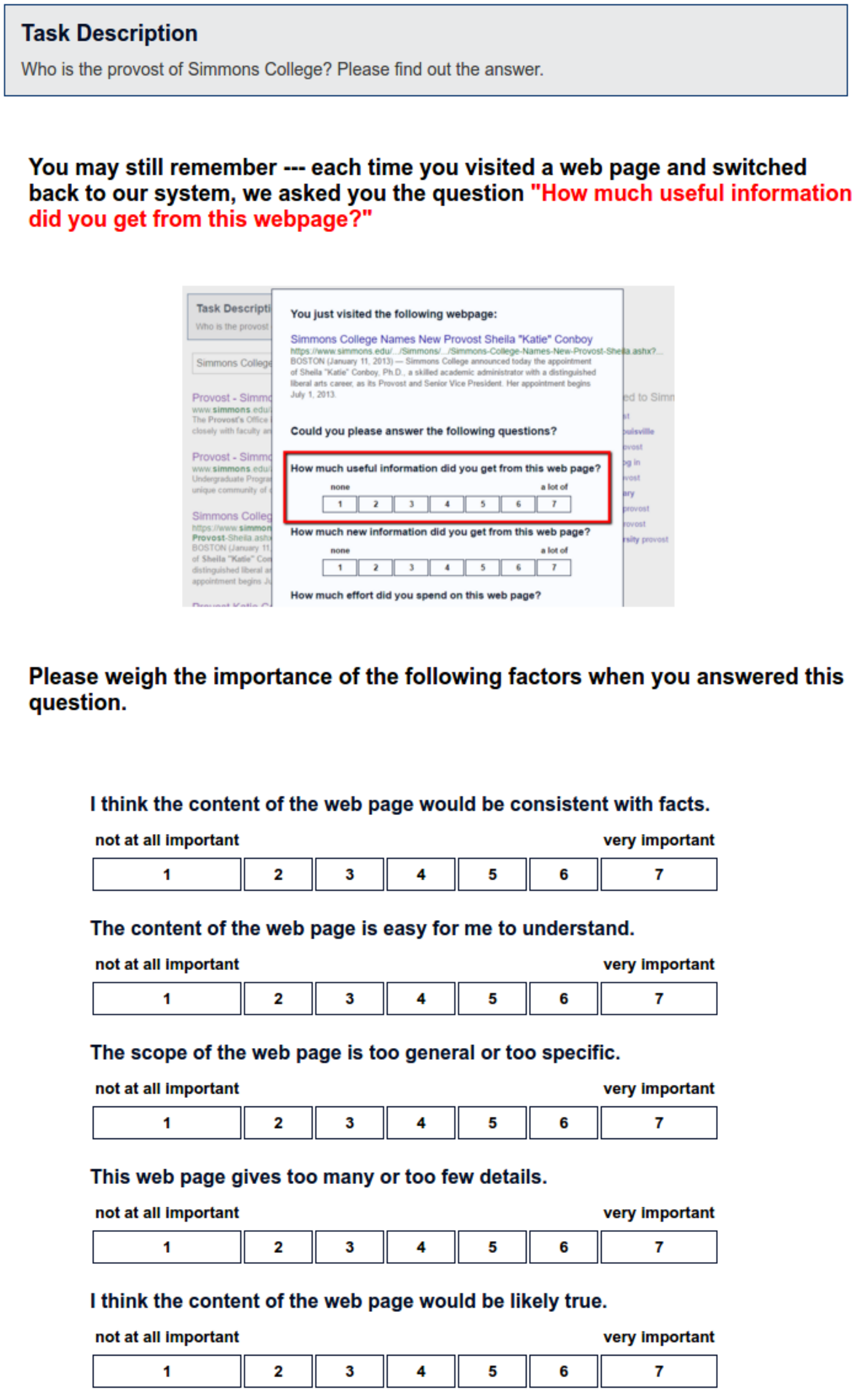
1 2 [✓ 3] 4 5 6 7

How frustrated were you with this task?

not frustrated — very frustrated

1 2 3 [✓ 4] 5 6 7

Figure 13: User Study 1: Ephemeral Relevance – a screenshot for the post-task survey.



Task Description

Who is the provost of Simmons College? Please find out the answer.

You may still remember --- each time you visited a web page and switched back to our system, we asked you the question "How much useful information did you get from this webpage?"

You just visited the following webpage:

Simmons College Names New Provost Sheila "Katie" Conboy
https://www.simmons.edu/.../Simmons/.../Simmons-College-Names-New-Provost-Sheila.ashx?...
BOSTON (January 11, 2013) — Simmons College announced today the appointment of Sheila "Katie" Conboy, Ph.D., a skilled academic administrator with a distinguished liberal arts career, as its Provost and Senior Vice President. Her appointment begins July 1, 2013.

Could you please answer the following questions?

How much useful information did you get from this web page?

none — a lot of

1 2 3 4 5 6 7

How much new information did you get from this web page?

none — a lot of

1 2 3 4 5 6 7

How much effort did you spend on this web page?

Please weigh the importance of the following factors when you answered this question.

I think the content of the web page would be consistent with facts.

not at all important — very important

1 2 3 4 5 6 7

The content of the web page is easy for me to understand.

not at all important — very important

1 2 3 4 5 6 7

The scope of the web page is too general or too specific.

not at all important — very important

1 2 3 4 5 6 7

This web page gives too many or too few details.

not at all important — very important

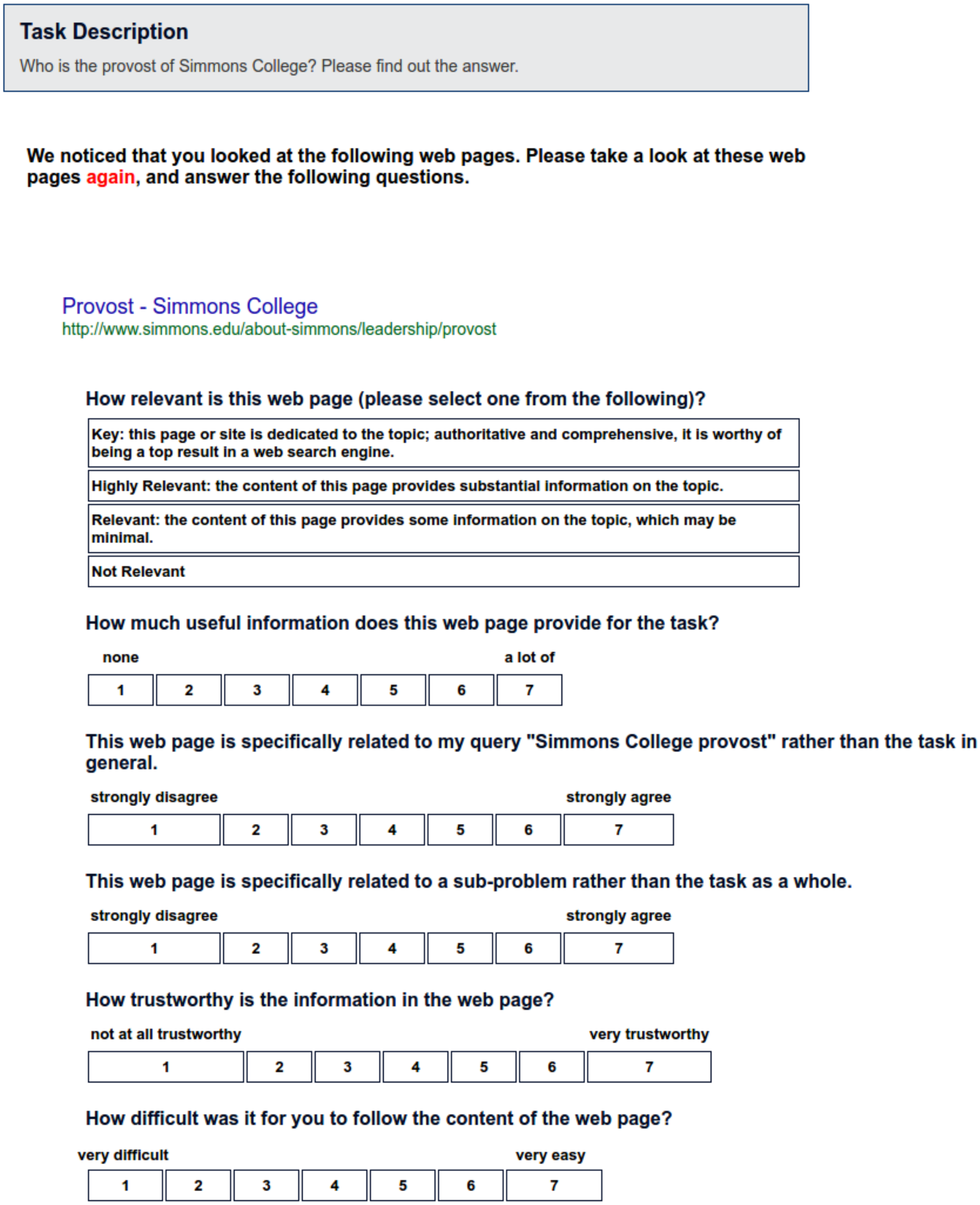
1 2 3 4 5 6 7

I think the content of the web page would be likely true.

not at all important — very important

1 2 3 4 5 6 7

Figure 14: User Study 1: Ephemeral Relevance – a screenshot for the survey of ERel judgments criteria.



Task Description

Who is the provost of Simmons College? Please find out the answer.

We noticed that you looked at the following web pages. Please take a look at these web pages again, and answer the following questions.

Provost - Simmons College
http://www.simmons.edu/about-simmons/leadership/provost

How relevant is this web page (please select one from the following)?

Key: this page or site is dedicated to the topic; authoritative and comprehensive, it is worthy of being a top result in a web search engine.
Highly Relevant: the content of this page provides substantial information on the topic.
Relevant: the content of this page provides some information on the topic, which may be minimal.
Not Relevant

How much useful information does this web page provide for the task?

none						a lot of
1	2	3	4	5	6	7

This web page is specifically related to my query "Simmons College provost" rather than the task in general.

strongly disagree						strongly agree
1	2	3	4	5	6	7

This web page is specifically related to a sub-problem rather than the task as a whole.

strongly disagree						strongly agree
1	2	3	4	5	6	7

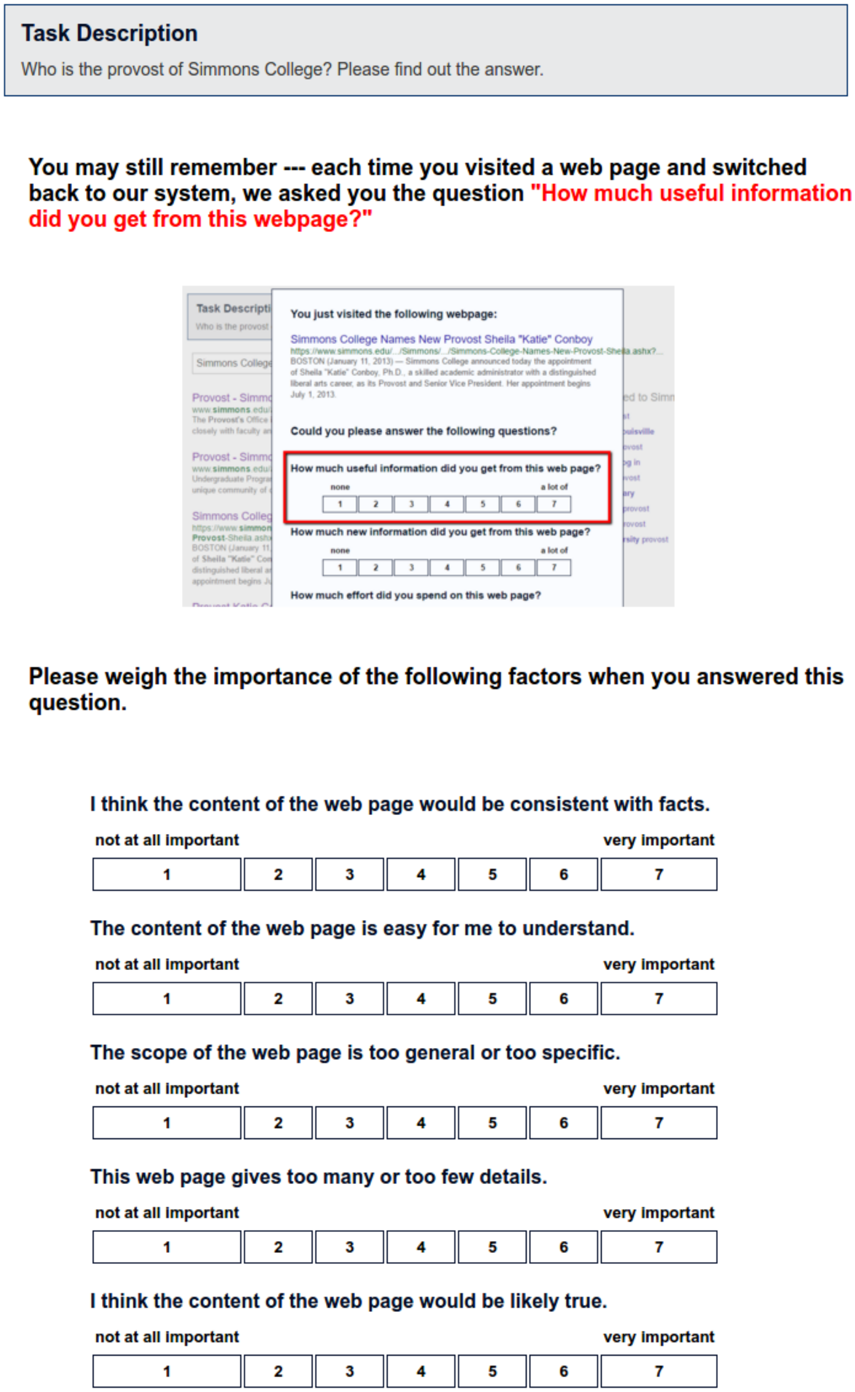
How trustworthy is the information in the web page?

not at all trustworthy						very trustworthy
1	2	3	4	5	6	7

How difficult was it for you to follow the content of the web page?

very difficult						very easy
1	2	3	4	5	6	7

Figure 15: User Study 1: Ephemeral Relevance – a screenshot for the required topical relevance judgments.



Task Description

Who is the provost of Simmons College? Please find out the answer.

You may still remember --- each time you visited a web page and switched back to our system, we asked you the question "How much useful information did you get from this webpage?"

You just visited the following webpage:

Simmons College Names New Provost Sheila "Katie" Conboy
https://www.simmons.edu/.../Simmons/.../Simmons-College-Names-New-Provost-Sheila.ashx?...
BOSTON (January 11, 2013) — Simmons College announced today the appointment of Sheila "Katie" Conboy, Ph.D., a skilled academic administrator with a distinguished liberal arts career, as its Provost and Senior Vice President. Her appointment begins July 1, 2013.

Could you please answer the following questions?

How much useful information did you get from this web page?

none — a lot of

1 2 3 4 5 6 7

How much new information did you get from this web page?

none — a lot of

1 2 3 4 5 6 7

How much effort did you spend on this web page?

Please weigh the importance of the following factors when you answered this question.

I think the content of the web page would be consistent with facts.

not at all important — very important

1 2 3 4 5 6 7

The content of the web page is easy for me to understand.

not at all important — very important

1 2 3 4 5 6 7

The scope of the web page is too general or too specific.

not at all important — very important

1 2 3 4 5 6 7

This web page gives too many or too few details.

not at all important — very important

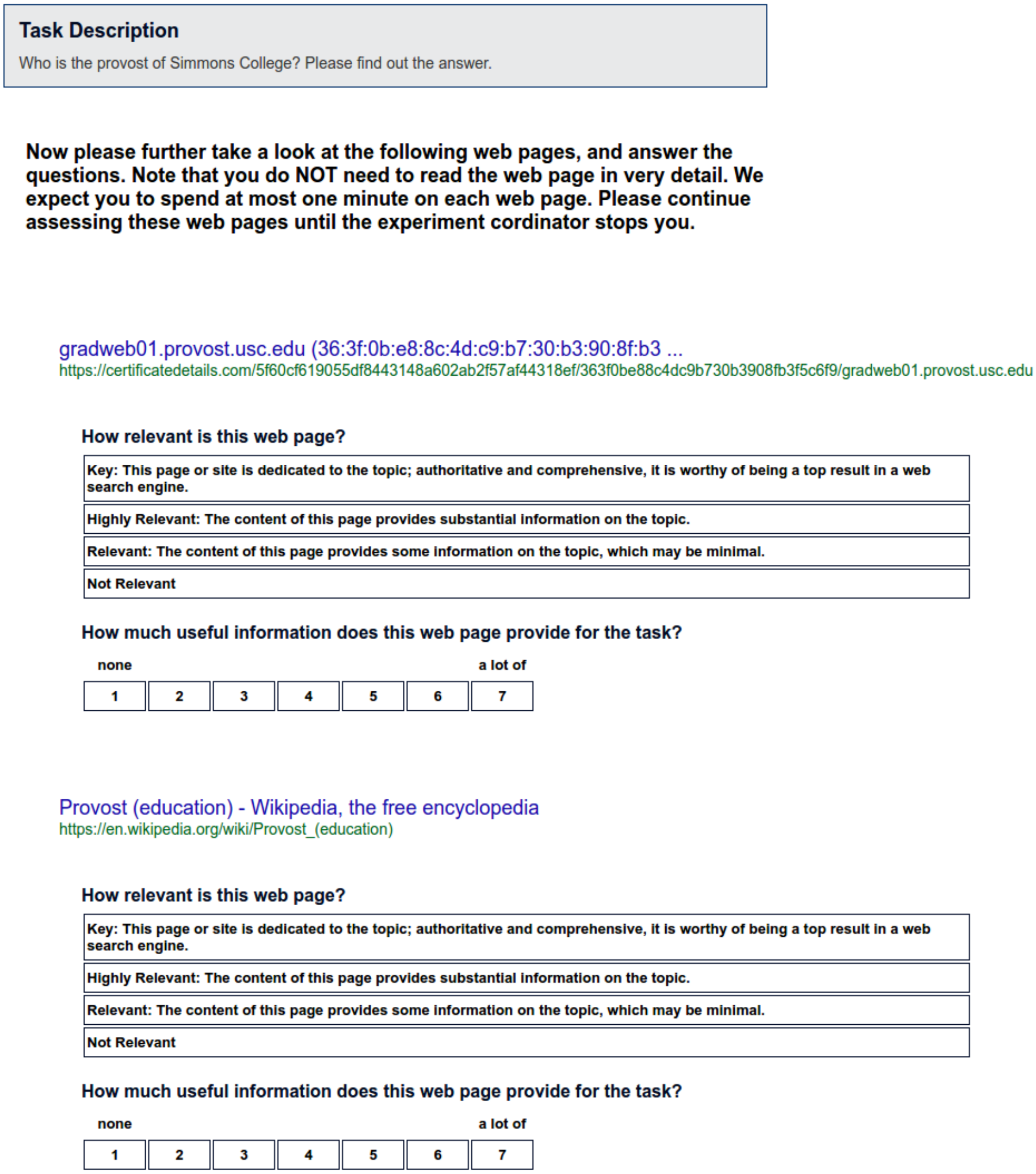
1 2 3 4 5 6 7

I think the content of the web page would be likely true.

not at all important — very important

1 2 3 4 5 6 7

Figure 16: User Study 1: Ephemeral Relevance – a screenshot for the survey of topical relevance judgments criteria.



Task Description

Who is the provost of Simmons College? Please find out the answer.

Now please further take a look at the following web pages, and answer the questions. Note that you do NOT need to read the web page in very detail. We expect you to spend at most one minute on each web page. Please continue assessing these web pages until the experiment cordinator stops you.

gradweb01.provost.usc.edu (36:3f:0b:e8:8c:4d:c9:b7:30:b3:90:8f:b3 ...
https://certificatedetails.com/5f60cf619055df8443148a602ab2f57af44318ef/363f0be88c4dc9b730b3908fb3f5c6f9/gradweb01.provost.usc.edu

How relevant is this web page?

Key: This page or site is dedicated to the topic; authoritative and comprehensive, it is worthy of being a top result in a web search engine.

Highly Relevant: The content of this page provides substantial information on the topic.

Relevant: The content of this page provides some information on the topic, which may be minimal.

Not Relevant

How much useful information does this web page provide for the task?

none						a lot of
1	2	3	4	5	6	7

Provost (education) - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Provost_(education)

How relevant is this web page?

Key: This page or site is dedicated to the topic; authoritative and comprehensive, it is worthy of being a top result in a web search engine.

Highly Relevant: The content of this page provides substantial information on the topic.

Relevant: The content of this page provides some information on the topic, which may be minimal.

Not Relevant

How much useful information does this web page provide for the task?

none						a lot of
1	2	3	4	5	6	7

Figure 17: User Study 1: Ephemeral Relevance – a screenshot for the optional topical relevance judgments.

APPENDIX D

REFERENCES

Agichtein, E., Brill, E., & Dumais, S. (2006). Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)* (pp. 19–26).

Agichtein, E., Brill, E., Dumais, S., & Ragno, R. (2006). Learning user interaction models for predicting web search result preferences. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)* (pp. 3–10).

Agrawal, R., Gollapudi, S., Halverson, A., & Ieong, S. (2009). Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM '09)* (pp. 5–14).

Allan, J., Aslam, J., Belkin, N., Buckley, C., Callan, J., Croft, B., ... Zhai, C. (2003). Challenges in information retrieval and language modeling: Report of a workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, September 2002. *SIGIR Forum*, *37*(1), 31–47.

Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* Pearson.

Anick, P. (2003). Using terminological feedback for web search refinement: A log-based study. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR '03)* (pp. 88–95).

Armstrong, T. G., Moffat, A., Webber, W., & Zobel, J. (2009). Improvements that don't add up: Ad-hoc retrieval results since 1998. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09)* (pp. 601–610).

Azzopardi, L. (2014). Modelling interaction with economic models of search. In *Proceedings of the 37th International ACM SIGIR Conference on Research &#38; Development in Information Retrieval (SIGIR '14)* (pp. 3–12).

Baeza-Yates, R., Hurtado, C., & Mendoza, M. (2004). Query recommendation using query logs in search engines. In *Proceedings of the 2004 International Conference on Current Trends in Database Technology (EDBT'04)* (pp. 588–596).

Bar-Yossef, Z., & Kraus, N. (2011). Context-sensitive query auto-completion. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)* (pp. 107–116).

Baskaya, F., Keskustalo, H., & Järvelin, K. (2012). Time drives interaction: Simulating sessions in diverse searching environments. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)* (pp. 105–114).

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

Belkin, N. J., Cole, M. J., & Liu, J. (2009). A model for evaluation of interactive information retrieval. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation.*

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

Bendersky, M., Metzler, D., & Croft, W. B. (2010). Learning concept importance using a weighted dependence model. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM '10)* (pp. 31–40).

Bennett, P. N., Collins-Thompson, K., Kelly, D., White, R. W., & Zhang, Y. (Eds.). (2015). Overview of the special issue on contextual search and recommendation. *ACM Transactions on Information Systems*, *33*(1), 1e:1–1e:7.

Bennett, P. N., Radlinski, F., White, R. W., & Yilmaz, E. (2011). Inferring and using location metadata to personalize web search. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR*

'11) (pp. 135–144).

Bennett, P. N., White, R. W., Chu, W., Dumais, S. T., Bailey, P., Borisyuk, F., & Cui, X. (2012). 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 (SIGIR '12)* (pp. 185–194).

Bernstein, M. S., Teevan, J., Dumais, S., Liebling, D., & Horvitz, E. (2012). Direct answers for search queries in the long tail. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)* (pp. 237–246).

Björnsson, C.-H. (1983). Readability of newspapers in 11 languages. *Reading Research Quarterly*, 480–497.

Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, *54*(10), 913–925.

Bruza, P., & Dennis, S. (1997). Query reformulation on the internet: Empirical data and the hyperindex search engine. In *Proceedings of the RIAO '97 Conference: Computer-Assisted Information Searching on Internet (RIAO '97)* (pp. 488–499).

Cao, H., Jiang, D., Pei, J., He, Q., Liao, Z., Chen, E., & Li, H. (2008). Context-aware query suggestion by mining click-through and session data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08)* (pp. 875–883).

Carmel, D., Zwerdling, N., Guy, I., Ofek-Koifman, S., Har'el, N., Ronen, I., ... Chernov, S. (2009). Personalized social search based on the user's social network. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09)* (pp. 1227–1236).

Carterette, B., Clough, P., Hall, M., Kanoulas, E., & Sanderson, M. (2016). Evaluating retrieval over sessions: the TREC session track 2011-2014. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)* (pp. 685–688).

Carterette, B., Kanoulas, E., Hall, M., Bah, A., & Clough, P. (2013). Overview of the TREC 2013 session track. In *Proceedings of the Twenty-Second Text REtrieval Conference (TREC 2013).*

Carterette, B., Kanoulas, E., Hall, M., & Clough, P. (2014). Overview of the TREC 2014 session track. In *Proceedings of the Twenty-Third Text REtrieval Conference (TREC 2014)*.

Catledge, L. D., & Pitkow, J. E. (1995). Characterizing browsing strategies in the world-wide web. In *Proceedings of the Third International World-Wide Web Conference on Technology, Tools and Applications* (pp. 1065–1073).

Chao, W., Liu, Y., Wang, M., Zhou, K., Nie, J., & Ma, S. (2015). Incorporating non-sequential behavior into click models. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)* (pp. 283–292).

Chapelle, O., Joachims, T., Radlinski, F., & Yue, Y. (2012). Large-scale validation and analysis of interleaved search evaluation. *ACM Transactions on Information Systems*, *30*(1), 6:1–6:41.

Chapelle, O., Metlzer, D., Zhang, Y., & Grinspan, P. (2009). Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09)* (pp. 621–630).

Chapelle, O., & Zhang, Y. (2009). A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th International Conference on World Wide Web (WWW '09)* (pp. 1–10).

Chilton, L. B., & Teevan, J. (2011). Addressing people's information needs directly in a web search result page. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)* (pp. 27–36).

Chuklin, A., Markov, I., & de Rijke, M. (2015). Click models for web search. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *7*(3), 1-115.

Chuklin, A., Schuth, A., Hofmann, K., Serdyukov, P., & de Rijke, M. (2013). Evaluating aggregated search using interleaving. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management (CIKM '13)* (pp. 669–678).

Clarke, C. L. A., Agichtein, E., Dumais, S., & White, R. W. (2007). The influence of caption features on clickthrough patterns in web search. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information*

Retrieval (SIGIR '07) (pp. 135–142).

Clarke, C. L. A., Craswell, N., & Soboroff, I. (2009). Overview of the TREC 2009 web track. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009).*

Clarke, C. L. A., Craswell, N., & Soboroff, I. (2010). Overview of the TREC 2010 web track. In *Proceedings of the Nineteenth Text REtrieval Conference Proceedings (TREC 2010).*

Clarke, C. L. A., Craswell, N., & Soboroff, I. (2011). Overview of the TREC 2011 web track. In *Proceedings of the Twentieth Text REtrieval Conference Proceedings (TREC 2011).*

Clarke, C. L. A., Craswell, N., & Soboroff, I. (2012). Overview of the TREC 2012 web track. In *Proceedings of the Twenty-First Text REtrieval Conference Proceedings (TREC 2012).*

Clarke, C. L. A., Craswell, N., & Soboroff, I. (2013). Overview of the TREC 2013 web track. In *Proceedings of the Twenty-Second Text REtrieval Conference Proceedings (TREC 2013).*

Clarke, C. L. A., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., & MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)* (pp. 659–666).

Cole, M. J., Gwizdka, J., Liu, C., Bierig, R., Belkin, N. J., & Zhang, X. (2011). Task and user effects on reading patterns in information search. *Interacting with Computers*, *23*(4), 346–362.

Cole, M. J., Hendahewa, C., Belkin, N. J., & Shah, C. (2014). Discrimination between tasks with user activity patterns during information search. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)* (pp. 567–576).

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, *60*(2), 283–284.

Collins-Thompson, K., Bennett, P. N., White, R. W., de la Chica, S., & Sontag, D. (2011). Personalizing web search results by reading level. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11)* (pp. 403–412).

Collins-Thompson, K., Macdonald, C., Bennett, P., Diaz, F., & Voorhees, E. M. (2014). TREC 2014 web track overview. In *Proceedings of the Twenty-Third Text REtrieval Conference Proceedings (TREC 2014).*

Craswell, N., Zoeter, O., Taylor, M., & Ramsey, B. (2008). An experimental comparison of click position-bias models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM '08)* (pp. 87–94).

Cutrell, E., & Guan, Z. (2007). What are you looking for?: An eye-tracking study of information usage in web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)* (pp. 407–416).

Danescu-Niculescu-Mizil, C., Broder, A. Z., Gabrilovich, E., Josifovski, V., & Pang, B. (2010). Competing for users' attention: On the interplay between organic and sponsored search results. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)* (pp. 291–300).

Dang, V., & Croft, B. W. (2010). Query reformulation using anchor text. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM '10)* (pp. 41–50).

Dou, Z., Song, R., & Wen, J.-R. (2007). A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of the 16th International Conference on World Wide Web (WWW '07)* (pp. 581–590).

Dupret, G., & Liao, C. (2010). A model to estimate intrinsic document relevance from the clickthrough logs of a web search engine. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM '10)* (pp. 181–190).

Dupret, G. E., & Piwowarski, B. (2008). A user browsing model to predict search engine click data from past observations. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)* (pp. 331–338).

Eisenberg, M., & Barry, C. (1988). Order effects: A study of the possible influence of presentation order on user judgments of document relevance. *Journal of the American Society for Information Science*, *39*(5), 293–300.

Ellis, D. (1989). A behavioural approach to information retrieval system design. *Journal of*

Documentation, *45*(3), 171–212.

Freund, L., & Berzowska, J. (2010). The goldilocks effect: Task-centred assessments of E-government information. In *Proceedings of the 73rd Annual Meeting of the American Society for Information Science and Technology (ASIS&T '10).*

Freund, L., Toms, E. G., & Clarke, C. L. A. (2005). Modeling task-genre relationships for IR in the workplace. In *Proceedings of the 28th annual international acm sigir conference on research and development in information retrieval (sigir '05)* (pp. 441–448).

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378.

Gauch, S., Chaffee, J., & Pretschner, A. (2003). Ontology-based personalized search and browsing. *Web Intelligence and Agent Systems*, *1*(3-4), 219–234.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models.* Cambridge University.

Grice, H. P. (1991). *Studies in the way of words.* Harvard University Press.

Gross, M. (2001). Imposed information seeking in public libraries and school library media centres: a common behaviour. *Information Research*, *6*(2).

Guan, D., Zhang, S., & Yang, H. (2013). Utilizing query change for session search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13)* (pp. 453–462).

Guan, Z., & Cutrell, E. (2007). An eye tracking study of the effect of target rank on web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)* (pp. 417–420).

Guo, F., Liu, C., Kannan, A., Minka, T., Taylor, M., Wang, Y.-M., & Faloutsos, C. (2009). Click chain model in web search. In *Proceedings of the 18th International Conference on World Wide Web (WWW '09)* (pp. 11–20).

Guo, F., Liu, C., & Wang, Y. M. (2009). Efficient multiple-click models in web search. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM '09)* (pp. 124–131).

Hansen, P., & Karlgren, J. (2005). Effects of foreign language and task scenario on relevance assessment. *Journal of Documentation*, *61*(5), 623–639.

Harman, D. (1992a). Overview of the first Text REtrieval Conference (TREC-1). In *Proceedings of the First Text REtrieval Conference (TREC-1)* (pp. 1–20).

Harman, D. (1992b). Relevance feedback revisited. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '92)* (pp. 1–10).

Harman, D. (2002). Overview of the TREC 2002 novelty track. In *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002).*

Hassan, A., Jones, R., & Klinkner, K. L. (2010). Beyond dcg: User behavior as a predictor of a successful search. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM '10)* (pp. 221–230).

Hawking, D., Voorhees, E., Craswell, N., & Bailey, P. (1999). Overview of the TREC-8 web track. In *Proceedings of the Eighth Text REtrieval Conference (TREC 8)* (pp. 131–150).

He, D., Göker, A., & Harper, D. J. (2002). Combining evidence for automatic web session identification. *Information Processing & Management*, *38*(5), 727–742.

Hofmann, K., Behr, F., & Radlinski, F. (2012). On caption bias in interleaving experiments. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)* (pp. 115–124).

Hofmann, K., Whiteson, S., & Rijke, M. D. (2013). Fidelity, soundness, and efficiency of interleaved comparison methods. *ACM Transactions on Information Systems*, *31*(4), 17:1–17:43.

Huang, J., & Efthimiadis, E. N. (2009). Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09)* (pp. 77–86).

Huffman, S. B., & Hochster, M. (2007). How well does result relevance predict session satisfaction? In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07)* (pp. 567–574).

Ingwersen, P., & Järvelin, K. (2005). *The turn: Integration of information seeking and retrieval in context (the information retrieval series).* Springer-Verlag New York, Inc.

Jansen, B. J., Booth, D., & Smith, B. (2009). Using the taxonomy of cognitive learning to model online searching. *Information Processing & Management*, *45*(6), 643–663.

Jansen, B. J., Booth, D. L., & Spink, A. (2009). Patterns of query reformulation during web searching. *Journal of the American Society for Information Science and Technology*, *60*(7), 1358–1371.

Järvelin, K., Price, S. L., Delcambre, L. M. L., & Nielsen, M. L. (2008). Discounted cumulated gain based evaluation of multiple-query IR sessions. In *Proceedings of the 30th annual European Conference on Information Retrieval Research (ECIR '08)* (pp. 4–15).

Jiang, J., & Allan, J. (2016a). Adaptive effort for search evaluation metrics. In *Proceedings of the 38th European Conference on Information Retrieval (ECIR '16)* (pp. 187–199).

Jiang, J., & Allan, J. (2016b). Correlation between system and user metrics in a session. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval (CHIIR '16)* (pp. 285–288).

Jiang, J., & Allan, J. (2016c). Reducing click and skip errors in search result ranking. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM '16)* (pp. 183–192).

Jiang, J., Hassan Awadallah, A., Jones, R., Ozertem, U., Zitouni, I., Gurunath Kulkarni, R., & Khan, O. Z. (2015). Automatic online evaluation of intelligent assistants. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)* (pp. 506–516).

Jiang, J., Hassan Awadallah, A., Shi, X., & White, R. W. (2015). Understanding and predicting graded search satisfaction. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)* (pp. 57–66).

Jiang, J., He, D., & Allan, J. (2014). Searching, browsing, and clicking in a search session: Changes in user behavior by task and over time. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)* (pp. 607–616).

Jiang, J., He, D., & Han, S. (2012). On duplicate results in a search session. In *Proceedings of the Twenty-First Text REtrieval Conference (TREC 2012).*

Jiang, J., He, D., Han, S., Yue, Z., & Ni, C. (2012). Contextual evaluation of query reformulations in a search session by user simulation. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)* (pp. 2635–2638).

Jiang, J., & Ni, C. (2016). What affects word changes in query reformulation during a task-based search session? In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval (CHIIR '16)* (pp. 111–120).

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '02)* (pp. 133–142).

Joachims, T. (2003). Evaluating retrieval performance using clickthrough data. In J. Franke, G. Nakhaeizadeh, & I. Renz (Eds.), *Text Mining* (pp. 79–96). Physica/Springer Verlag.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)* (pp. 154–161).

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

Jones, R., & Fain, D. C. (2003). Query word deletion prediction. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR '03)* (pp. 435–436).

Jones, R., & Klinkner, K. L. (2008). 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 (CIKM '08)* (pp. 699–708).

Jones, R., Rey, B., Madani, O., & Greiner, W. (2006). Generating query substitutions. In *Proceedings of the 15th International Conference on World Wide Web (WWW '06)* (pp. 387–396).

Kanoulas, E., Carterette, B., Clough, P. D., & Sanderson, M. (2011). Evaluating multi-query sessions. In *Proceedings of the 34th International ACM SIGIR Conference on*

Research and Development in Information Retrieval (SIGIR '11) (pp. 1053–1062).

Kanoulas, E., Carterette, B., Hall, M., Clough, P., & Sanderson, M. (2011). Session track 2011 overview. In *Proceedings of the Twentieth Text REtrieval Conference (TREC 2011).*

Kanoulas, E., Carterette, B., Hall, M., Clough, P., & Sanderson, M. (2012). Overview of the TREC 2012 session track. In *Proceedings of the Twenty-First Text REtrieval Conference (TREC 2012).*

Kelly, D., Arguello, J., Edwards, A., & Wu, W.-c. (2015). Development and evaluation of search tasks for iir experiments using a cognitive complexity framework. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval (ICTIR '15)]* (pp. 101–110).

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

Kelly, D., Gyllstrom, K., & Bailey, E. W. (2009). A comparison of query and term suggestion features for interactive searching. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)* (pp. 371–378).

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

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

Kim, K.-S., & Allen, B. (2002). Cognitive and task influences on web searching behavior. *Journal of the American Society for Information Science and Technology*, *53*(2), 109–119.

Kim, Y., Hassan, A., White, R. W., & Zitouni, I. (2014a). Comparing client and server dwell time estimates for click-level satisfaction prediction. In *Proceedings of the 37th International ACM SIGIR Conference on Research &#38; Development in Information Retrieval (SIGIR '14)* (pp. 895–898).

Kim, Y., Hassan, A., White, R. W., & Zitouni, I. (2014b). Modeling dwell time to predict click-level satisfaction. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM '14)* (pp. 193–202).

Kiseleva, J., Williams, K., Jiang, J., Hassan Awadallah, A., Crook, A. C., Zitouni, I., & Anastasakos, T. (2016). Understanding user satisfaction with intelligent assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval (CHIIR '16)* (pp. 121–130).

Koenemann, J., & Belkin, N. J. (1996). A case for interaction: A study of interactive information retrieval behavior and effectiveness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '96)* (pp. 205–212).

Kotov, A., Bennett, P. N., White, R. W., Dumais, S. T., & Teevan, J. (2011). Modeling and analysis of cross-session search tasks. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)* (pp. 5–14).

Krovetz, R. (1993). Viewing morphology as an inference process. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '93)* (pp. 191–202).

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

Li, J., Huffman, S., & Tokuda, A. (2009). Good abandonment in mobile and PC internet search. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)* (pp. 43–50).

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

Liu, C. (2012). *Personalizing information retrieval using interaction behaviors in search sessions in different types of tasks* (Ph.D. dissertation). School of Communication and Information, Rutgers, the State University of New Jersey.

Liu, C., Belkin, N. J., & Cole, M. J. (2012). Personalization of search results using interaction behaviors in search sessions. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)* (pp. 205–214).

Liu, C., Liu, J., & Belkin, N. J. (2014). Predicting search task difficulty at different search stages. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14)* (pp. 569–578).

Liu, J., & Belkin, N. J. (2010). 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 (SIGIR '10)* (pp. 26–33).

Liu, J., & Belkin, N. J. (2014). Multi-aspect information use task performance: The roles of topic knowledge, task structure, and task stage. In *Proceedings of the 77th Annual Meeting of the American Society for Information Science and Technology (ASIS&T '14).*

Liu, J., & Belkin, N. J. (2015). Personalizing information retrieval for multi-session tasks: Examining the roles of task stage, task type, and topic knowledge on the interpretation of dwell time as an indicator of document usefulness. *Journal of the Association for Information Science and Technology*, *66*(1), 58–81.

Liu, J., Belkin, N. J., Zhang, X., & Yuan, X. (2013). Examining users' knowledge change in the task completion process. *Information Processing & Management*, *49*(5), 1058–1074.

Liu, J., Cole, M. J., Liu, C., Bierig, R., Gwizdka, J., Belkin, N. J., ... Zhang, X. (2010). Search behaviors in different task types. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries (JCDL '10)* (pp. 69–78).

Liu, Z., Liu, Y., Zhou, K., Zhang, M., & Ma, S. (2015). Influence of vertical result in web search examination. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)* (pp. 193–202).

Lorigo, L., Pan, B., Hembrooke, H., Joachims, T., Granka, L., & Gay, G. (2006). The influence of task and gender on search and evaluation behavior using Google. *Information Processing & Management*, *42*(4), 1123–1131.

Luo, J., Zhang, S., & Yang, H. (2014). 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 (SIGIR '14)* (pp. 587–596).

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

Mei, Q., Zhou, D., & Church, K. (2008). Query suggestion using hitting time. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM '08)* (pp. 469–478).

Menard, S. (2002). *Applied logistic regression analysis* (Second Edition ed.). Sage Publications.

Metzler, D., & Croft, W. B. (2005). A markov random field model for term dependencies. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)* (pp. 472–479).

Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science*, *48*(9), 810–832.

Mizzaro, S. (1998). How many relevances in information retrieval? *Interacting with Computers*, *10*(3), 303–320.

Montanez, G. D., White, R. W., & Huang, X. (2014). Cross-device search. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14)* (pp. 1669–1678).

Over, P. (2001). The TREC interactive track: an annotated bibliography. *Information Processing & Management*, *37*(3), 369–381.

Pitkow, J., Schütze, H., Cass, T., Cooley, R., Turnbull, D., Edmonds, A., ... Breuel, T. (2002). Personalized search. *Communications of the ACM*, *45*(9), 50–55.

Radlinski, F., & Craswell, N. (2013). Optimized interleaving for online retrieval evaluation. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (WSDM '13)* (pp. 245–254).

Radlinski, F., & Joachims, T. (2005). Query chains: Learning to rank from implicit feedback. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD '05)* (pp. 239–248).

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, *124*(3), 372–422.

Reid, J. (2000). A task-oriented non-interactive evaluation methodologyfor information retrieval systems. *Information Retrieval*, *2*(1), 115–129.

Richardson, M., Dominowska, E., & Ragno, R. (2007). Predicting clicks: Estimating the click-through rate for new ads. In *Proceedings of the 16th International Conference on World Wide Web (WWW '07)* (pp. 521–530).

Rieh, S. Y., & Xie, H. (2006). Analysis of multiple query reformulations on the Web: The interactive information retrieval context. *Information Processing & Management*, *42*(3), 751–768.

Robertson, S., Zaragoza, H., & Taylor, M. (2004). Simple bm25 extension to multiple weighted fields. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management (CIKM '04)* (pp. 42–49).

Robertson, S. E., & Spärck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, *27*(3), 129–146.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., & Gatford, M. (1994). Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC 3)* (pp. 109–126).

Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, *41*(4), 288–297.

Sanderson, M. (2010). Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, *4*(4), 247–375.

Sanderson, M., Paramita, M. L., Clough, P., & Kanoulas, E. (2010). Do user preferences and evaluation measures line up? In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10)* (pp. 555–562).

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, *26*(6), 321–343.

Saracevic, T. (1996). Relevance reconsidered. In *Proceedings of the Second Conference on Conceptions of Library and Information Science (CoLIS 2)* (pp. 201–218).

Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology*, *58*(13), 1915–

1933.

Schuth, A., Sietsma, F., Whiteson, S., Lefortier, D., & de Rijke, M. (2014). Multileaved comparisons for fast online evaluation. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14)* (pp. 71–80).

Senter, R., & Smith, E. (1967). *Automated readability index* (Tech. Rep. No. AMRL-TR-66-220). Aerospace Medical Research Laboratories.

Shen, S., Hu, B., Chen, W., & Yang, Q. (2012). Personalized click model through collaborative filtering. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM '12)* (pp. 323–332).

Shen, X., Tan, B., & Zhai, C. (2005). Context-sensitive information retrieval using implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)* (pp. 43–50).

Shokouhi, M. (2013). Learning to personalize query auto-completion. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13)* (pp. 103–112).

Shokouhi, M., & Radinsky, K. (2012). Time-sensitive query auto-completion. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)* (pp. 601–610).

Shokouhi, M., White, R., & Yilmaz, E. (2015). Anchoring and adjustment in relevance estimation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)* (pp. 963–966).

Shokouhi, M., White, R. W., Bennett, P., & Radlinski, F. (2013). Fighting search engine amnesia: Reranking repeated results. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13)* (pp. 273–282).

Smith, C. L., & Kantor, P. B. (2008). User adaptation: Good results from poor systems. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)* (pp. 147–154).

Smucker, M. D., & Clarke, C. L. (2012). Time-based calibration of effectiveness measures.

In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)* (pp. 95–104).

Soboroff, I. (2004). Overview of the TREC 2004 novelty track. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004).*

Soboroff, I., & Harman, D. (2003). Overview of the TREC 2003 novelty track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003).*

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, *28*(1), 11–21.

Spink, A., Bateman, J., & Jansen, B. J. (1998). Searching heterogeneous collections on the Web: behaviour of Excite users. *Information Research*, *4*(2).

Spink, A., Jansen, B. J., Wolfram, D., & Saracevic, T. (2002). From E-sex to E-commerce: Web search changes. *Computer*, *35*(3), 107–109.

Spink, A., Ozmutlu, H. C., & Ozmutlu, S. (2002). Multitasking information seeking and searching processes. *Journal of the American Society for Information Science and Technology*, *53*(8), 639–652.

Spink, A., Park, M., Jansen, B. J., & Pedersen, J. (2006). Multitasking during web search sessions. *Information Processing & Management*, *42*(1), 264–275.

Spink, A., & Saracevic, T. (1997). Interaction in information retrieval: Selection and effectiveness of search terms. *Journal of the American Society for Information Science*, *48*(8), 741–761.

Srikant, R., Basu, S., Wang, N., & Pregibon, D. (2010). User browsing models: Relevance versus examination. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10)* (pp. 223–232).

Tan, B., Shen, X., & Zhai, C. (2006). Mining long-term search history to improve search accuracy. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06)* (pp. 718–723).

Taylor, A. R., Cool, C., Belkin, N. J., & Amadio, W. J. (2007). Relationships between categories of relevance criteria and stage in task completion. *Information Processing & Management*, *43*(4), 1071–1084.

Teevan, J., Adar, E., Jones, R., & Potts, M. A. S. (2007). 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 (SIGIR '07)* (pp. 151–158).

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

Tombros, A., & Sanderson, M. (1998). Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)* (pp. 2–10).

Trotman, A., & Keeler, D. (2011). Ad hoc IR: Not much room for improvement. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)* (pp. 1095–1096).

Turpin, A., & Scholer, F. (2006). User performance versus precision measures for simple search tasks. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)* (pp. 11–18).

Vargas, S., & Castells, P. (2011). Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys '11)* (pp. 109–116).

Verberne, S., Sappelli, M., Järvelin, K., & Kraaij, W. (2015). User simulations for interactive search: Evaluating personalized query suggestion. In *Proceedings of the 37th annual European Conference on Information Retrieval Research (ECIR '15)* (pp. 678–690).

Verma, M., Yilmaz, E., & Craswell, N. (2016). On obtaining effort based judgements for information retrieval. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM '16)* (pp. 277–286).

Voorhees, E. M. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)* (pp. 315–323).

Wang, C., Liu, Y., Zhang, M., Ma, S., Zheng, M., Qian, J., & Zhang, K. (2013). Incorporating vertical results into search click models. In *Proceedings of the 36th International*

ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13) (pp. 503–512).

Wang, H., Song, Y., Chang, M.-W., He, X., White, R. W., & Chu, W. (2013). Learning to extract cross-session search tasks. In *Proceedings of the 22Nd International Conference on World Wide Web (WWW '13)* (pp. 1353–1364).

Wang, X., & Zhai, C. (2008). Mining term association patterns from search logs for effective query reformulation. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM '08)* (pp. 479–488).

White, R. W., & Awadallah, A. H. (2015). Personalizing search on shared devices. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)* (pp. 523–532).

White, R. W., Bailey, P., & Chen, L. (2009). Predicting user interests from contextual information. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)* (pp. 363–370).

White, R. W., & Drucker, S. M. (2007). Investigating behavioral variability in web search. In *Proceedings of the 16th International Conference on World Wide Web (WWW '07)* (pp. 21–30).

White, R. W., Dumais, S. T., & Teevan, J. (2009). Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM '09)* (pp. 132–141).

White, R. W., & Horvitz, E. (2013). Captions and biases in diagnostic search. *ACM Transactions on the Web*, *7*(4), 23:1–23:28.

White, R. W., Jose, J. M., & Ruthven, I. (2003). A task-oriented study on the influencing effects of query-biased summarisation in web searching. *Information Processing & Management*, *39*(5), 707–733.

White, R. W., & Kelly, D. (2006). A study on the effects of personalization and task information on implicit feedback performance. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM '06)* (pp. 297–306).

White, R. W., Ruthven, I., & Jose, J. M. (2005). A study of factors affecting the utility of implicit relevance feedback. In *Proceedings of the 28th Annual International ACM*

SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05) (pp. 35–42).

Wilson, T. D. (1981). On user studies and information needs. *Journal of Documentation*, *37*(1), 3–15.

Wilson, T. D. (1997). Information behaviour: an interdisciplinary perspective. *Information Processing & Management*, *33*(4), 551–572.

Wu, W.-C., Kelly, D., Edwards, A., & Arguello, J. (2012). Grannies, tanning beds, tattoos and NASCAR: Evaluation of search tasks with varying levels of cognitive complexity. In *Proceedings of the 4th Information Interaction in Context Symposium (IIIX '12)* (pp. 254–257).

Xu, Y. (2007). Relevance judgment in epistemic and hedonic information searches. *Journal of the American Society for Information Science and Technology*, *58*(2), 179–189.

Xu, Y., & Chen, Z. (2006). Relevance judgment: What do information users consider beyond topicality? *Journal of the American Society for Information Science and Technology*, *57*(7), 961–973.

Xu, Y., & Wang, D. (2008). Order effect in relevance judgment. *Journal of the American Society for Information Science and Technology*, *59*(8), 1264–1275.

Yilmaz, E., Shokouhi, M., Craswell, N., & Robertson, S. E. (2010). Expected browsing utility for web search evaluation. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM '10)* (pp. 1561–1564).

Yilmaz, E., Verma, M., Craswell, N., Radlinski, F., & Bailey, P. (2014). Relevance and effort: An analysis of document utility. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14)* (pp. 91–100).

Yue, Y., Patel, R., & Roehrig, H. (2010). Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)* (pp. 1011–1018).

Yue, Z., Han, S., He, D., & Jiang, J. (2014). Influences on query reformulation in collaborative web search. *Computer*, *47*(3), 46–53.

Zhai, C., & Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01)* (pp. 334–342).

Zhang, X., Liu, J., Cole, M., & Belkin, N. (2015). 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.

Zhang, Y., Zhang, J., Lease, M., & Gwizdka, J. (2014). Multidimensional relevance modeling via psychometrics and crowdsourcing. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)* (pp. 435–444).

Zhong, F., Wang, D., Wang, G., Chen, W., Zhang, Y., Chen, Z., & Wang, H. (2010). Incorporating post-click behaviors into a click model. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10)* (pp. 355–362).