A Study of Immediate Requery Behavior in Search

Haotian Zhang
School of Computer Science
University of Waterloo
haotian.zhang@uwaterloo.ca

Mustafa Abualsaud
School of Computer Science
University of Waterloo
m2abuals@uwaterloo.ca

Mark D. Smucker
Department of Management Sciences
University of Waterloo
mark.smucker@uwaterloo.ca

ABSTRACT

When search results fail to satisfy users' information needs, users often reformulate their search query in the hopes of receiving better results. In many cases, users immediately requery without clicking on any search results. In this paper, we report on a user study designed to investigate the rate at which users immediately reformulate at different levels of search quality. We had users search for answers to questions as we manipulated the placement of the only relevant document in a ranked list of search results. We show that as the quality of search results decreases, the probability of immediately requerying increases. We find that users can quickly decide to immediately reformulate, and the time to immediately reformulate appears to be independent of the quality of the search results. Finally, we show that there appears to be two types of users. One group has a high probability of immediately reformulating and the other is unlikely to immediately reformulate unless no relevant documents can be found in the search results. While requerying takes time, it is the group of users who are more likely to immediately requery that are able to able find answers to questions the fastest.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**;

KEYWORDS

Immediate requery; Query abandonment; User study

ACM Reference Format:
Haotian Zhang, Mustafa Abualsaud, and Mark D. Smucker. 2018. A Study of Immediate Requery Behavior in Search. In *CHIIR '18: 2018 Conference on Human Information Interaction & Retrieval, March 11–15, 2018, New Brunswick, NJ, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3176349.3176400

1 INTRODUCTION

Today's search engines are typified by interfaces that allow a search user to issue a text query and then receive a list of search results. The moment the search engine results page (SERP) is displayed, the user begins processing that page with a goal of making one of three decisions:

(1) Click a search result to navigate to its page for viewing.
(2) Abandon the query, but continue the search by reformulating the query to produce a new search results page.
(3) Abandon not only the query but also the search. The next interaction with the search engine will not be a continuation of the current search.

Modern web search engines not only return organic search results, but also advertisements and other possible interaction mechanisms, for example, other suggested queries. In this paper, we limit our discussion to an abstract search engine that only returns organic search results in a ranked list, and where each search result is displayed with a summary to aid the user in deciding on the result's relevance.

To distinguish the abandonment in choice 2 from the abandonment in choice 3 above, we term choice 2 an *immediate requery*, i.e. a query reformulation without any clicks on search results. While a user performing an immediate requery does not click on any search results, the user will spend some time to view the search results and reformulate the query.

An immediate requery means that the user effectively places zero value on the search results. Even if the search results may contain relevant results, the immediate requery means that the user has spent time on the page but remains unsatisfied. If a user found significant value in the search result summaries, we assume the user would either click on a search result or abandon the query satisfied. Given the apparent loss in value to the user that results from an immediate requery, it is important to understand what conditions make immediate requeries likely. In particular, how good do search results need to be to have at least one click and avoid being treated as worthless with an immediate requery?

We conducted a controlled user study to investigate the relationship between search results quality and immediate requeries. In our study, we asked participants to find the answers to a set of questions. The questions were selected to be simple to answer given a good search engine, but unlikely for our study participants to already know the answers. For example, one question was "How long is the Las Vegas monorail in miles?" We varied the quality of the search results by placing one relevant document at varying ranks. We selected the non-relevant search results to appear somewhat plausible as search results for the given question, but to also be clearly non-relevant on inspection.

We found that in our study:

- Users make their decisions to requery or click quickly. The median time from query to immediate requery was 7.7 seconds.
- The probability of an immediate requery increases as the user has to search further down the ranked list to find a relevant document. In particular, the probability of an immediate

CHIIR '18, March 11–15, 2018, New Brunswick, NJ, USA

requery approximately doubles when the topmost relevant document is at rank 2 rather than at rank 1.
- The time it takes users to make a decision of whether to requery or not, appears to be independent of search results quality.

We also found that there may be two classes of user behavior for the examination of search results. One group, the majority, focuses on the top of the ranked list to make their decision about whether to requery or not. The other group appears to be more likely to examine the whole ranked list. The group more likely to immediately requery is able to find answers faster than the group less like to immediately requery.

We next review related work, then detail our experiment, report results, and finally conclude the paper.

2 BACKGROUND AND RELATED WORK

Hearst [12, Section 3.5.5] synthesizes research on web search behavior to suggest that a common strategy for users to follow is to "issue general queries, get information about the results, reformulate based on information seen in the results, and then navigate to promising-looking links or else give up." As part of this common strategy, it has long been recognized that some search users will decide to immediately reformulate their query without clicking on any search results. While terminology describing this behavior varies, commonly it is referred to as *query abandonment*. Joachims and Radlinski [16] termed "abandonment" to be "the user's decision to not click on any of the results." Likewise, Radlinski et al. [24] defined *abandonment rate* to be "the fraction of queries for which no results were clicked on." Unfortunately, "query abandonment" also sounds similar to what a user does after clicking on a result and deciding to reformulate a query. Indeed, Wu and Kelly [31] defined query abandonment to be "the point at which a person decides to stop his/her current query and enter a new one."

To avoid confusion, we define an *immediate requery* to be when a user enters a query, and then without clicking on any of the search results, the user reformulates the query to continue their search. For our notion of a modern search engine, clicking on a search result is the user interface action that allows the user to navigate to the result and view it in its entirety. We expect that users will, in most cases, view some search result summaries/snippets even if they do not click on them.

Li et al. [18] highlight that there is both good and bad query abandonment. Good query abandonment occurs when users find, for example, the answer they were looking for in the search results summaries or located somewhere on the SERP. Bad abandonment is associated with the user being dissatisfied with the search results. In our study, we are focused on bad abandonment caused by poor quality in search results.

There are a host of reasons why users may abandon queries. Stamou and Efthimiadis [29] classified query abandonment reasons into two categories: intentional causes and unintentional causes. The intentional causes include, for example, spelling or syntax changes to the query, checking whether search results have changed since the last time they issued the same query, and understanding the meaning of the query by looking at its results. The unintentional causes include, for example, no results returned, results are irrelevant, repetition of previously seen results, and interrupted search.

In an another study, Stamou and Efthimiadis [28] examined two types of post-query search abandonment: 1) pre-determined (when the user plans to find answers from the result snippets without clicking at any result), and 2) post-determined (when the user plans to click on a result but decides not to after viewing the SERP). They found that 27% of queries were abandoned due to a pre-determined intentions, and nearly half of the post-determined queries were abandoned due to dissatisfaction with the SERP.

Diriye et al. [9] found that 27% of SERP abandonment is not due to satisfaction nor dissatisfaction with results. The reasons of abandonment were: users came up with a better query before they viewed the SERP (13%), users found search results not sufficiently important (3%), and the user got interrupted by some factor (1%) (e.g., network failed and tab closed). Some 10% of the reasons fell into a catch-all "other" category.

Wu and Kelly [31] found three factors that may influence query abandonment. The first factor was the properties of search results. The proportion and relative location of relevant results determines the quality of SERPs and further affect query abandonment. The second factor was the properties of query. Users can learn new vocabulary from current query result and as a result they issue a new query. The last factor was the properties of the search task. Some users requery each time a subtask is fulfilled.

Several researchers have used eye-tracking as part of their studies on how users interact with search results. Granka et al. [11] showed that users spend more time and attention to top ranked results, and that they generally work top to bottom when looking for relevant documents. In addition to spending more time on top ranked results, researchers have also found that users are biased towards clicking on top results [14, 15, 19].

Klöckner et al. [17] classified users into two groups based on how they processed search results. One group followed a "strictly depth-first" strategy where they work down the ranked list one result at a time. The remaining participants followed either "partially breadth-first" or "extreme breadth-first" strategies. A partial breadth-first strategy is reflected by looking ahead a few results and making comparisons between the results to determine what to click on. The extreme breadth-first approach involves studying all of the search results before deciding which to click on.

Like Klöckner et al., Aula et al. [1] found users to follow either an "economic" or "exhaustive" strategy for processing search results. In Aula et al.'s study, about 6-7 summaries were visible at a time on the computer screen, and *economic* users would scan at most the first three results before acting. The *exhaustive* users would examine more than half of the visible summaries and sometimes even scroll to see the remaining summaries before acting. Aula et al. [1] found that the *economic* searchers had more computer experience and would fixate for shorter periods on each result.

Dumais et al. [10] found three groups of users and following the convention of Aula et al. [1], named the groups: "economic-results", "economic-ads", and "exhaustive". Dumais et al.'s study involved a commercial search engine and the two *economic* groups differed in how they examined advertisements. A significant difference between the economic and exhaustive groups was the amount of time spent examining result summaries. The *economic* users spent

between 8.7 and 9.9 seconds while the *exhaustive* users spent 14.6 seconds on average. Some users may display exhaustive behavior as a result of being dyslexic, for MacFarlane et al. [21] have found that dyslexic users are more likely to backtrack and reread material.

Lorigo et al. [20] investigated how a SERP's components such as result summaries can affect clicking behavior. In their study, participants used a web search engine to find answers to various short questions. Eye tracking data of participants using the web search was collected and analyzed. Using this data, the authors found that the relevance of the top 3 documents in the list can be a useful indicator to whether users will further explore the rest of the list. If the first 3 documents of the SERP are non-relevant, most users will end their exploration of the list. Eye tracking data provided insight on participants reading behavior of the SERP. In particular, they found that users generally tend to skim document summaries.

A large-scale cursor/mouse tracking study by Huang et al. [13] found that users tend to hover over 4 documents before deciding to requery. In contrast to Lorigo et al. [20] and Huang et al. [13], Cutrell and Guan [7] report that users view the first 8 results before deciding to requery.

Wu et al. [32] conducted a user study in which participants had to complete several search tasks. Each task consisted of using a web search engine to find an answer to exploration-type questions that often require multiple queries and multiple page visits. They manipulated their SERP according to two within-subject variables: Information Scent Level (ISL) and Information Scent Pattern (ISP). ISL was defined as the number of relevant documents appearing in the first SERP of the task, and ISP as the distribution of four relevant documents in the SERP. Both ISL and ISP included 3 categories. Low, medium, and high for ISL and persistent, disrupted, and bursting for ISP. These categories addressed different qualities of the SERPs. The authors found that around 42% of users abandoned their queries without any click on low ISL SERPs (where only the first document is relevant), and 13% of users requery on medium ISL SERPs (where only the top 3 documents are relevant). Only 1.6% of uses requery on high ISL SERP (where only the top 5 documents are relevant). For tasks under ISP, they found no big difference in SERP abandonment between persistent ISP (relevant documents at rank 1, 2, 5, and 8) and disrupted ISP (relevant documents at rank 1, 2, 3, and 4). Persistent ISP and disrupted ISP had 10% and 12% SERP abandonment rate respectively. Bursting ISP (relevant documents at rank 4, 5, 6, and 7) had 20% rate of SERP abandonment.

Finally, there has been considerable recent work in the simulation of user behavior for information retrieval evaluation [3, 6]. Much of this work attempts to model user behavior with search engines so that the models can be used to make accurate predictions of user behavior and gain received from the search engine [4, 5, 25, 27, 30].

Of particular note, many researchers have looked at modeling a user's decision to either stop processing search results, or when to stop and reformulate the query to get new, and hopefully better search results [8, 22, 23, 26].

3 METHODS AND MATERIALS

In this section, we describe the details of our experiment. To measure the effect of search results quality on users' requery behavior, we created a controlled within-subjects laboratory user study. After giving their consent to participate, each participant in our study was asked to find the answer to 12 questions using our custom search engine. We designed the search engine to manipulate the search results and control the placement of one relevant document in the 10 displayed search results. We carefully instrumented the search engine to allow us to record detailed user interaction data. We next describe the search tasks, how we manipulated the search results to vary their quality, how we measured user behavior, and the study design.

3.1 Search Tasks

We asked each participant to search for answers to 12 questions. Table 1 shows the 12 questions including a practice question. For each search task, we provided participants with a single question and asked them to use our search engine to find an answer to the question. Participants could enter as many queries as they wanted and spend as much time as needed to find the correct answer. We designed the questions to meet the following requirements: (i) Most participants should not already know the answer, and thus, participants would be forced to search to find an answer. (ii) The question should be straightforward, non-confusing, and be able to be answered easily with the help of a modern search engine. (iii) Each question should only have one standard answer. (iv) The question should make it easy for us to find plausible non-relevant search results as well as a relevant web page that contains the answer.

After completing our study, we found that question 12 failed to meet the requirement that participants be able to easily answer it, for only 28% answered it correctly. In hindsight, we see that question 12 was tricky. Michael Jordan was selected to play in the All-Star Game 14 times but only played 13 games in total due to an injury. Many participants gave an answer of 14 games instead of 13. In addition, some participants had trouble with question 6 (78% accuracy), and this was because they entered the start of the lyrics as the answer rather than the song's title. All other questions had greater than 90% accuracy.

3.2 Search Interface

The search interface used for all study tasks is shown in Figure 1. The interface design was similar to that of common commercial search engines, except it did not include any means to get more than 10 results per query. Participants could enter their search queries using the search bar and trigger the query by either clicking on "Search" button or pressing "Enter" keystroke. The question of the current task that participants need to search an answer for was always visible and shown next to the search bar. The question was also shown during the pre-task. Clicking on the help button would trigger a pop-up showing the help information on how to use the interface. Subjects were asked to use this search interface to find an answer for each question and were allowed to submit multiple queries and click on multiple documents if they wished. To accurately measure clicks and time spent in the SERP and in the documents, we disabled right-clicks and opening documents in new tabs. Participants needed to use the back button on the browser to return back to the SERP after clicking and viewing a document.

ID	Question	Answer	Triggered Query Words
P	What is the weight of Hope Diamond in carats?	45.52	N/A (practice question)
1	How long is the Las Vegas monorail in miles?	3.9/4 miles.	Las, Vegas, monorail
2	Find out the name of the album that the Mountain Goats band released in 2004.	We Shall All Be Healed	Mountain, Mountian, Goats, Goat, album
3	Which year was the first Earth Day held?	1970	Earth, Day
4	Which year was the Holes (novel) written by Louis Sachar first published?	1998	Holes, hole, louis, sachar, Novel
5	Find the phone number of Rocky Mountain Chocolate Factory located in Ottawa, ON?	(613) 241-1091	Rocky, Mountain, Chocolate, Factory, Ottawa
6	What is the name of opening theme song for Mister Rogers' Neighbourhood?	Won't You Be My Neighbor?	Mister, Rogers, Roger, Roger's, Neighbourhood, opening, theme, song
7	Which album is the song Rain Man by Eminem from?	Encore	Rain, Man, Eminem
8	How many chapters are in The Art of War book written by Sun Tzu?	13	Art, War, Sun, Tzu
9	What is the scientific name of Mad cow disease?	Bovine Spongiform Encephalopathy (BSE)	Mad, Cow, Disease
10	How many campuses does the University of North Carolina have?	17	University, North, Carolina, Campus, campuses, UNC
11	Which Canadian site was selected as one of United Nations World Heritage Sites in 1999?	Miguasha National Park	United, Nations, World, Heritage, UN
12	How many times did Michael Jordan play the NBA All-Star Games?	13	Michael, Jordan, NBA, All-Star, Star

Table 1: The study's 1 practice and 12 search task questions and their corresponding answers and trigger query words.

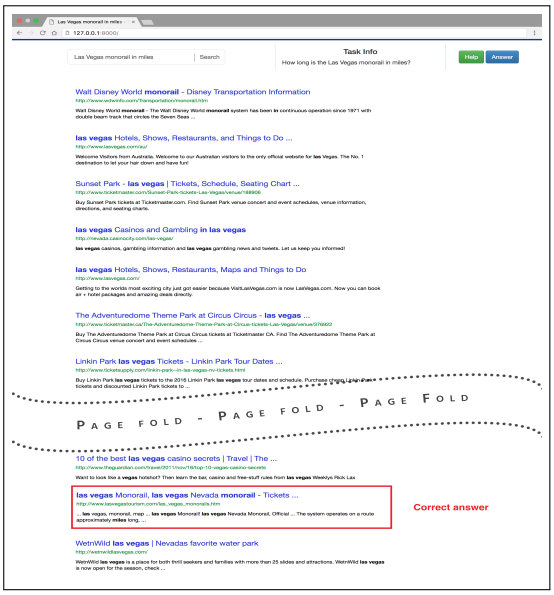

Figure 1: The search interface for all tasks. The interface has a search bar, help button and answer button. The SERP shows a maximum of 10 documents with no further results available. Here, a manipulated SERP is presented and the correct document is placed at the rank 9. In general, the results at ranks 8-10 were not visible without scrolling.

3.3 Quality of Search Results

Our search engine only provided 10 search results in response to a query. With 10 search results and simple binary relevance, there are 1024 (2^{10}) possible ways to construct search results to vary their quality. In this paper, a relevant document contains the answer to the user's question and a non-relevant document does not contain the answer. To simplify our study, we decided to focus on the placement of a single relevant document in a ranked list of 10 search results. Placing the single relevant document at ranks 1 through 10 gave us 10 different rankings where the assumption was that as the relevant document was placed lower in the ranking, the lower the search quality would be to the user. We also produced a ranking where all 10 documents were non-relevant, which we call an "All Bad" SERP. Finally, we also had a control condition where the search results were the actual results produced by the Bing search API[1] in response to the user's query.

All the results shown in manipulated SERPs contained at least one keyword from the question. Relevant, or *correct*, documents provided a straightforward answer to the user's question that should be easy for the user to find. Non-relevant, or *incorrect*, documents contain keywords from the question, and may be related to the question in some way, but their overall topic is clearly non-relevant. A non-relevant document does not contain the answer.

We found all documents and their snippets by issuing queries to the Bing search API. For documents with the correct answer in their snippets, we manually removed the answer from the snippets to force the user to click on the document and find the answer from its content. We only controlled the snippet content for the

[1] https://azure.microsoft.com/en-ca/services/cognitive-services/bing-web-search-api/

manipulated SERPs. The control SERP (Bing) used snippets directly from the Bing API.

In order to make the manipulated SERPs look realistic and reasonable, and to prevent participants from having any suspicion or confusion regarding the SERP, the incorrect documents were selected from queries related to the corresponding factoid question. Take, for example, the "Las Vegas Monorail" question shown in Figure 1 (ID 1 in Table 1). For this question, a somewhat realistic but unrelated query would be "Las Vegas Casino" or "Las Vegas Hotel". Both queries have the phrases "Las Vegas" but are not relevant to "Las Vegas monorail". We used such queries to retrieve incorrect documents for all 12 questions.

We constructed the SERPs in this fashion for participants to think the results were real, but to also make it clear there were many non-relevant documents in the results. We wanted to trigger immediate requeries while also studying the effect of rank on immediate requery behavior.

3.4 Triggering Manipulated SERPs

As described above, we had 11 manipulated SERP tasks (one relevant document at ranks 1-10, and all non-relevant documents). We wanted to be careful to only show the study participant the manipulated SERP if the participant entered a query that could reasonably be an attempt to use a search engine to find an answer to the given question. For each question, we constructed a list of keywords that if any of them were entered by the participant as part of their query, they would trigger the manipulated SERP. If the participant entered a query lacking all of the selected keywords, we would send the query to the Bing search API and return organic results. Table 1 shows the trigger keywords for each question.

For each search task, the participant can only trigger the manipulated SERP once. All further queries will not trigger the manipulated SERP, regardless of what the query terms are. All queries following the display of a manipulated SERP produce live, organic results from the Bing search API.

After analyzing the search logs for manipulated SERP tasks, we found that only 2 participants on 2 different tasks failed to trigger the manipulated SERP with their first query. The first user entered "canadian heratige site 1999" as their first query for task # 11, with the wrong spelling of the word "heritage". None of the query terms are triggers. The second user entered an empty query for task # 3 and our system returned an empty SERP. Both of these two users successfully triggered a manipulated SERP with their second query. For both of these two users, we skip their first query and analyze their data from the query that triggered a manipulated SERP.

For the control search task, all queries are sent to the Bing search API, and its results are shown to the participant.

3.5 Measurements

Our goal is to investigate the requery behavior of users given SERPs with different quality. More specifically, how much time users spend before they abandon a SERP and issue a new query and probability of an immediate requery on SERPs of different quality. To collect all necessary data to achieve these goals, we designed our interface to record all user actions and system responses (clicks, keystrokes, query submission, SERP appearance, etc.) and their corresponding timestamps in the client time to allow us to compute time spent between any two actions. Any time that a participant spent on the help page, was excluded from all measurements.

The time spent before an immediate requery is measured from the time the SERP loads following a query to the time they select the query box to reformulate their query. In 5 cases, a participant clicked the search button without reformulating their query, and in these cases, the time from query to the time they clicked the search button is counted as the length of time for the immediate requery.

The time to submit an answer is measured from the display of the search interface to the moment the participant submits their answer.

The time of a participant's "first click" on a result in a SERP is measured from the time the SERP loads following a query to the time of the first click on one of a SERP's results.

The time to formulate or reformulate a query was measured from the time of a participant's first keystroke in the query box to the time they submit the query.

3.5.1 Time on documents. For the manipulated SERPs, we already know the rank of correct (relevant) and incorrect (non-relevant) documents in the list. We can easily measure how long the participants spend on correct documents and incorrect documents.

For the organic SERPs returned by Bing, we performed a post-study analysis to manually check every web page clicked by participants. We classified these clicked documents into three categories.

- **Correct**: The document contains the correct answer.
- **Incorrect**: The document is non-relevant or does not contain the correct answer in linked to pages.
- **Not sure**: The document content does not include the correct answer, but the answer can be found by navigating the page links. To get to the answer, participants would need to click on some links from the document to get to the page where the answer is written.

After manually checking clicked documents from Bing SERPs and categorizing each document, we measure the time users spend on documents of each category.

3.6 Study Procedure and Design

The study was run in a closed computer laboratory using desktop machines with the same monitor size and specifications. The computer monitors had a screen resolution of 1680×1050 pixels. The Google Chrome browser was used to access the study.

After receiving participants' informed consent, we collected participants' demographics and information on their search engine usage and experience before the start of the study. Instructions on the study tasks and expectations were provided before the study. We mention that "You can enter as many queries as many times as you want." in the instructions to encourage participants to query. Cell phone usage was prohibited and complete attention during the study was expected of participants. We explained to participants that they were not allowed to use other search engines to find answers. A short quiz was used to ensure participants read and understood the study's tasks and instructions. Participants were not allowed to proceed to the study until all quiz questions were answered correctly.

We provided a practice page of the search interface and asked all participants to familiarize themselves with the interface by searching for an answer to a practice question (Table 1, ID "**P**"). All search results returned by the system during the practice phase were organic Bing results. Participants proceeded to their first task after completing the practice task. Completion of the practice task and all further tasks were done by providing a written answer to the task's question.

Each search task included a pre-task and a post-task questionnaire. During the pre-task, we showed the current question and asked participants about their prior knowledge of the current question topic. The post-task questionnaire asks the participants about their confidence in their answer. We asked participants on their feedback and overall experience with an end-of-study questionnaire.

3.6.1 Balanced Design. The study involved 12 tasks and 60 participants. We used a 12×12 Graeco-Latin square to create a fully balanced design and randomize SERP quality treatments and experimental conditions. As mentioned before, there are 12 different SERPs including 11 manipulated SERPs and one control Bing SERP. The 12 different SERPs composed one block. Each block balanced the order of tasks and the rank positions of correct documents. By randomizing the columns and rows, this process creates five separate 12×12 Graeco-Latin Squares - for all the 60 participants.

3.7 Participants

After receiving ethics approval from our university's office of research ethics, we recruited participants through posters placed in different departments of the university. The study involved 73 participants in total, but only 60 participants' data was used for our analysis. We removed data of 13 participants due to pilot testing and technical issues. After careful examination of the 60 participants' data, we did not find any irregularities and thus did not clean or modify their data before the analysis. Each of the 60 participants completed their 12 tasks in a balanced order, yielding a total of 720 tasks, 660 were manipulated SERP tasks, and 60 were non-manipulated organic Bing SERP tasks (control).

Participants' age ranged between 18 and 48 years old (mean = 23.6). There were 34 male and 26 female participants. Of these participants, 54 of them were from science, technology, engineering, or math, 1 from arts, and 5 did not specify their major.

Each participant was compensated $15 with an advertised payment of $10 for participation and a $5 bonus for answering at least 10 out of 12 questions correctly. However, regardless of participant performance, we paid all participants the full $15. This payment structure was designed to motivate good performance while not harming any person who might not have been able to answer 10 questions correctly. 58 participants answered 10 or more questions correctly. One participant answered 9 questions correctly, and one participant only answered 8 questions correctly.

4 RESULTS AND DISCUSSION

In our study, participants used a search engine to find answers to 12 questions. For 11 search tasks, we manipulated the search results quality. For one of the search tasks, which acted as a control, participants received results directly from the Bing search API. For the manipulated SERPs, any queries that followed the manipulated SERP provided results from the Bing search API. As explained in Section 3.3, the manipulated SERPs included 1 single correct document, placed in different ranks from 1 to 10, or 0 correct documents.

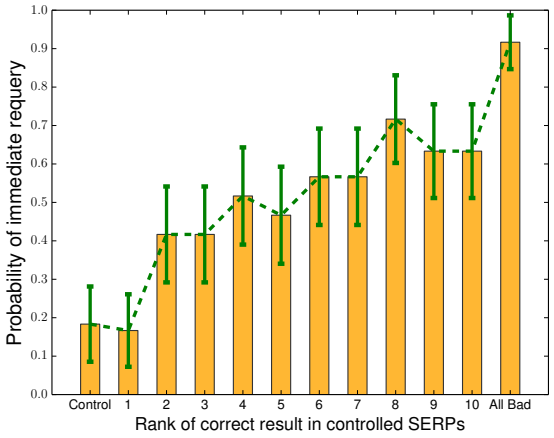

Figure 2: The probability of immediate requery for the 12 different SERP conditions. The control condition's search results are from the Bing search API. The "All Bad" condition means that all 10 search results are non-relevant. The error bars are 95% confidence intervals.

Figures 2, 3, 4, and Table 2 show our main results. In Figure 2, we see that as the rank of the relevant document goes from rank 1 (top of page) to rank 10 (bottom of page), the probability of an immediate requery increases. The highest probability for an immediate requery, 0.92, occurs when all of the search results are non-relevant, and this rate is a statistically significant difference from the other conditions. The control condition's search results, which are Bing API search results, have a probability of immediate requery of only 0.18, which is, for all purposes, the same as we saw for a relevant result at rank 1. The probability of immediate requery at rank 1, 0.17, is significantly different than at rank 2, 0.42.

Figure 3 shows that the time it takes a user to decide to immediately requery appears to be independent of the search results quality. Figure 4 shows the distribution of all times to immediate requery. The median time for an immediate requery is a fast 7.7 seconds, and the average time is 9.2 seconds. A log-normal distribution fitted to this data has a mean of 2.0 and a standard deviance of 0.68.

We also measured the time from a query to a participant's first click on the search results. Figure 5 and Table 3 show the time from a query to the first result click for ranks 1-10. We can see a very linear increase in the time it takes participants to scan the ranked list of results from rank 1 to rank 4. The median time from query to a click on rank 1 is only 3.1 seconds, and then it takes approximately 2 seconds more for each rank up to rank 4, which takes 10.4 seconds to reach. Participant's behavior on ranks 5-7 is different with these median times taking 8.5, 11.4, and 11.3 seconds. Finally, for the ranks that require the participant to scroll to reach, ranks 8-10, we see that participants appear to scan these upward

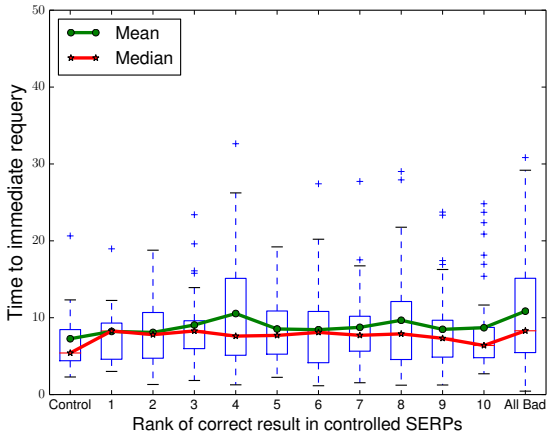

Figure 3: Time to immediate requery without any document clicks on the manipulated SERPs and organic Bing SERPs using the triggered query.

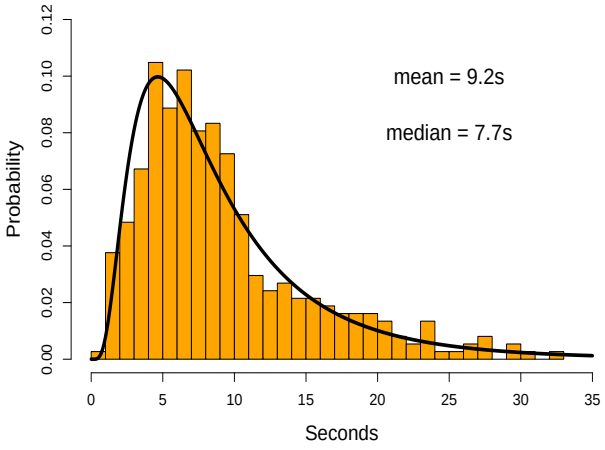

Figure 4: The distribution of time to immediate requery on the manipulated SERPs and organic Bing SERPs. A log normal curve fit to the data is also shown.

from rank 10 to 9 to 8 with median times of 14.4, 16.6, and 17.5 seconds, respectively.

As reviewed in the related work section (Section 2), past eye-tracking research also shows that users tend to linearly scan search results. But if users are linearly scanning the results, why does the time to immediately requery appear to be independent of the rank of the relevant document? One possible explanation is found in the other eye-tracking research that largely shows that users scan the first 3 or 4 results before deciding to requery. Our participants appear to be able to scan ranks 1-3 in 7.3 median seconds and our median time to immediately requery is 7.7 seconds. Unfortunately,

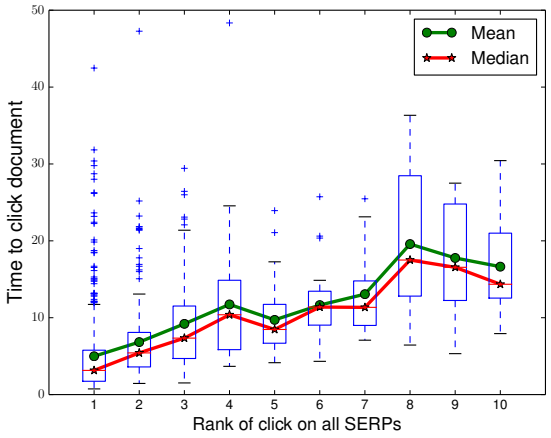

Figure 5: Time from query to the first result click at different ranks on all SERPs.

Rank of Correct Document	Freq.	Prob. Immediate Requery [95% CI]	Seconds to Immediate Requery [95% CI]
Control (Bing API)	11	0.18 [0.09, 0.28]	7.3 [4.1, 10.5]
1	10	0.17 [0.07, 0.26]	8.2 [5.2, 11.2]
2	25	0.42 [0.29, 0.54]	8.1 [6.2, 10]
3	25	0.42 [0.29, 0.54]	9.1 [7.0, 11.1]
4	31	0.52 [0.39, 0.64]	10.5 [7.9, 13.2]
5	28	0.47 [0.34, 0.59]	8.5 [6.9, 10.2]
6	34	0.57 [0.44, 0.69]	8.4 [6.5, 10.4]
7	34	0.57 [0.44, 0.69]	8.7 [7.0, 10.5]
8	43	0.72 [0.60, 0.83]	9.7 [7.7, 11.6]
9	38	0.63 [0.51, 0.76]	8.5 [6.8, 10.2]
10	38	0.63 [0.51, 0.76]	8.7 [6.8, 10.6]
All Bad Results	55	0.92 [0.85, 0.99]	10.9 [8.9, 12.8]

Table 2: The frequency, probability and time to immediate requery with corresponding 95% Confidence Interval on the controlled SERPs (cf. Figures 2 and 3).

Rank of Correct Document	Median Time To Click	Mean Time To Click [95% CI]
1	3.1	5.0 [4.5, 5.5]
2	5.4	6.8 [6.0, 7.7]
3	7.3	9.2 [8.0, 10.4]
4	10.4	11.7 [9.5, 13.9]
5	8.5	9.7 [8.6, 10.9]
6	11.4	11.6 [10.1, 13.2]
7	11.3	13.1 [10.8, 15.3]
8	17.5	19.6 [14.6, 24.5]
9	16.6	17.8 [14.5, 21.0]
10	14.4	16.6 [14.0, 19.3]

Table 3: Time in seconds to first click on a result at different ranks (cf. Figure 5).

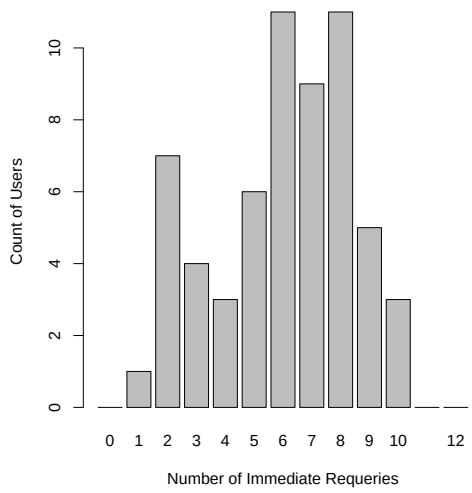

Figure 6: Distribution of immediate requeries for all participants.

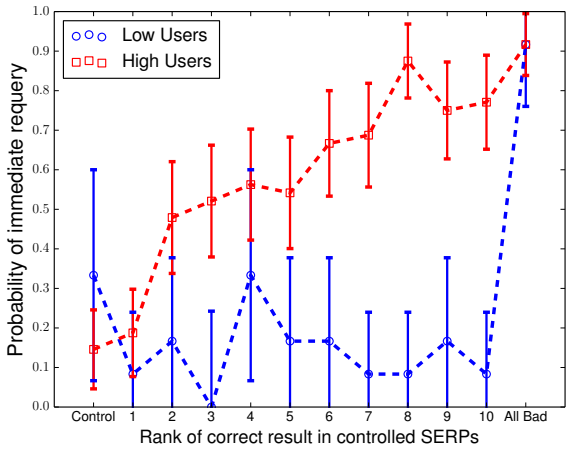

Figure 7: The probability of immediate requery for the 12 different SERP conditions for two different groups of participants (cf. Figure 2). The "Low Users" issued immediate requeries for 3 or fewer of the 12 search tasks. The "High Users" each had 4 or more immediate requeries. The error bars are 95% confidence intervals.

this does not explain why we see participants clicking at ranks 5-10, for if participants stopped their scans at ranks 3 or 4, they should never see the relevant documents at lower ranks to click on them.

Given past eye-tracking research that has shown there to be two different classes of searchers, i.e. economic and exhaustive searchers (see Section 2), we looked closer at the individual behavior of the study participants.

Figure 6 shows the distribution of number of immediate requeries per participant. While our analysis is limited by the number of participants and the number of search tasks, it appears that we have one group of participants who have a low rate of immediately requerying ($\leq$ 3 immediate requeries), and other group that immediately requery much more frequently ($\geq$ 4 immediate requeries). As such, we label each participant as either having a low or high probability of immediately requerying and looked at the behavior of each group.

Figure 7 shows the probability of immediately requerying for the *low* vs. *high* groups. As can be seen, the *low* group's probability of immediately requerying stays low until they are faced with search results that are all non-relevant. In contrast, the *high* group's probability of immediately requerying grows quickly as the rank of the relevant document goes from 1 to 10. It appears that the *low* group are *exhaustive* searchers while the *high* group are likely *economic* searchers.

If we believe that search users optimize their search behavior to find answers as quickly as possible [2], then the majority of participants who appear to be *economic* in their search behavior, should find answers faster in spite of their higher probability for immediately requerying. Indeed, we find that participants who are more likely to immediately requery to be able to find answers faster. The mean time to answer for the participants likely to immediately requery (*high*) is 85.9 seconds and the mean time for

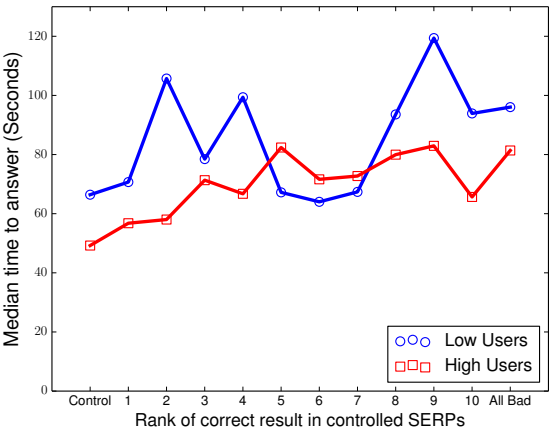

Figure 8: The median time from starting a task to answer a question for different groups of users.

the participants with *low* probability of immediately requerying is 111.6 seconds, and this difference is statistically significant by a two-tailed, Student's t-test (p=0.0005). While this difference is significant, it is possible that the *high* group's performance is the result of many additional factors that correlate with a higher probability for immediately requerying.

Figure 8 shows the median time to answer a question for the *low* and *high* groups of users across the 12 search conditions. While the data is noisy because of the limited size of the *low* group, we see that for the control condition, and when the relevant document is at ranks 1-4 and 8-10, the *low* participants take longer than the high

Rank of correct document	Median
Bing	57.6
1	59.7
2	60.7
3	72.2
4	69.6
5	77.9
6	71.1
7	72.5
8	81.5
9	89.5
10	70.2
All Bad	85.9

Table 4: For all participants, the median time in seconds from starting a task to answering a question for the different search conditions.

	Mean	Standard Error	Median
Correct doc time	22.9	0.7	16.1
Incorrect doc time	20.3	1.7	9.7
Not sure doc time	35.2	3.2	27.4

Table 5: Time in seconds to view types of documents.

group. We also see that for the mid-ranks of 5-7, the *low* users have slightly faster times to answer than the *high* group. For comparison, Table 4 reports the median time to answer for all participants.

What seems to be happening is that the *low* group wastes time looking at more results for results at ranks 1-4 than is necessary to select the relevant document. When the relevant document is at ranks 5-7, the group of participants with a *high* probability of immediately requerying has apparently stopped scanning at rank 3 or 4 and immediately requeried. Meanwhile, the *low* group, which is exhaustively scanning results finds the relevant document at ranks 5-7 without needing to incur the cost of an immediate requery. Strikingly, the *high* group appears to be able to keep the time to answer nearly uniform for ranks 5-10 and "All Bad", for while they have to take time to immediately requery, we know from the control condition that participants find the answer quickly with the Bing search results.

The cost of an immediate requery is actually quite low. First there is the sunk cost of examining the current results, which we reported earlier as a median time of 7.7 seconds. The median time for a user to reformulate their query was 3.2 seconds. The median time to reach rank 1 in the search results is only 3.1 seconds. Thus, assuming the reformulated query can find a relevant document at rank 1, users with a high probability of immediately requerying should be able to cap their median cost to reach a relevant document at approximately 14 seconds, which is more than the cost to reach relevant documents at ranks 5-7, but less than the cost to reach ranks 8-10, and this seems to explain their behavior and our results.

4.1 Document Review Time

Table 5 shows the mean, standard error of the mean, and median time users spent reading documents in the three different categories we defined in Section 3.5.1. Given that this sort of user data typically follows a log-normal distribution, the median time is usually more informative than the mean. We see that participants are able to quickly realize their mistake in clicking on an incorrect (non-relevant) document. After clicking on an incorrect document, they return back to the search results in only 9.7 seconds. The median times on correct and "not sure" documents are longer than for incorrect documents.

4.2 Study Limitations

A limitation of our work is that we only studied one type of search task. Our study participants needed to find answers to simple questions. Other search tasks may result in different behavior. For example, when our study participants experienced a SERP with only 1 relevant document at rank 1, we only saw a 17% immediate requery rate, which is considerably different than the 42% that Wu et al. [32] found (see Section 2). Likewise, when our topmost relevant document is at rank 4, we found that 52% of participants would immediately requery while Wu et al.'s "bursting" pattern had only a 20% rate. We think these differences in results are likely the result of the different types of search tasks that our two studies used. Our study had participants search for a single answer to a simple question. On the other hand, Wu et al. had many search tasks that would involve attempting to find many relevant documents. It appears that the search task can change immediate requery behavior.

A potential concern of our study would be if participants noticed the manipulation of search results. Our study provided a means for participants to supply open ended feedback after each search task as well as at the end of the study. Some participants commented that they were surprised that our search engine would not return Wikipedia search results at rank one when they included keywords such as "wiki" in their queries. One participant noted that our search engine seemed to be sensitive to the order of words in the query. Thus, while participants may have noticed some behavior different from commercial search engines, they did not specifically make mention of our manipulated behavior, and we did not notice any behavior that would indicate that they understood how the results were manipulated.

5 CONCLUDING DISCUSSION

There are many reasons for immediate requeries and not all of them are bad. In this paper, we focused our study on immediate requeries caused by poor quality results. Ignoring good reasons for abandonment, conventional wisdom holds that search engines should strive to minimize the fraction of queries that result in an immediate requery.

We expected to find that as the search results quality decreased, that the rate of immediately requerying would increase, and we did find this to be the case. Based on our results and others' eye-tracking studies, it appears that immediately requerying in web search is largely caused by the topmost relevant search results appearing at ranks lower than 3 or 4.

What we did not expect to find were a group of participants who were more likely to immediately requery, and that these participants would find answers faster than participants who stayed with the search results. In other words, being quick to immediately requery may actually be an efficient strategy for use of web search engines, which would explain why we saw a majority of participants employ this technique.

Unfortunately, these results mean that modifications to a search engine that lowers the rate of immediate requeries, may actually hurt user performance if the modification forces users to stick with bad results rather than quickly move to better results.

Indeed, it would seem that an important function of web search engines is to help users quickly find a query that delivers relevant documents at ranks 1 to 3. The faster a search engine can guide a user's query reformulations to the "right query", the faster the user will find relevant results.

Traditional evaluation of search engines focuses on the single list of search results produced by a query. Unfortunately, looking only at the quality of a search engine's ranking, focuses attention primarily on the minority of users who have a low probability of immediately requerying. In our study, it does not seem to matter to the majority of participants if a relevant document is at rank 5 or rank 10, both are considered to be worthless. It is important to keep in mind that for different or more complex search tasks, we might expect user behavior to differ from what we observed.

If only the top 3 or 4 results matter to a majority of users, as information retrieval researchers, we will need to both work to help users zero-in on the right query and to find ways to evaluate a search engine's ability to help users with this process of querying and repeated reformulation.

ACKNOWLEDGMENTS

The idea for this research was in part a result of a sabbatical that Mark Smucker spent at Microsoft, and the conversations he had with Ryen White, Susan Dumais, Paul Bennett, and others during the sabbatical. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (Grants CRDPJ 468812-14 and RGPIN-2014-03642), in part by Google, and in part by the University of Waterloo.

REFERENCES

[1] Anne Aula, Päivi Majaranta, and Kari-Jouko Räihä. 2005. Eye-Tracking Reveals the Personal Styles for Search Result Evaluation. In *Human-Computer Interaction – INTERACT 2005 (LNCS)*, Vol. 3585, 1058–1061.

[2] Leif Azzopardi. 2011. The economics in interactive information retrieval. In *SIGIR*, 15–24.

[3] Leif Azzopardi, Kalervo Järvelin, Jaap Kamps, and Mark D. Smucker. 2011. Report on the SIGIR 2010 workshop on the simulation of interaction. *SIGIR Forum* 44 (January 2011), 35–47. Issue 2.

[4] Feza Baskaya, Heikki Keskustalo, and Kalervo Järvelin. 2012. Time Drives Interaction: Simulating Sessions in Diverse Searching Environments. In *SIGIR*, 105–114.

[5] Ben Carterette, Evangelos Kanoulas, and Emine Yilmaz. 2011. Simulating simple user behavior for system effectiveness evaluation. In *CIKM*, 611–620.

[6] Charles L.A. Clarke, Luanne Freund, Mark D. Smucker, and Emine Yilmaz. 2013. Report on the SIGIR 2013 Workshop on Modeling User Behavior for Information Retrieval Evaluation (MUBE 2013). *SIGIR Forum* 47, 2 (Jan. 2013), 84–95.

[7] Edward Cutrell and Zhiwei Guan. 2007. What are you looking for?: an eye-tracking study of information usage in web search. In *SIGCHI*, 407–416.

[8] Arjen P. de Vries, Gabriella Kazai, and Mounia Lalmas. 2004. Tolerance to Irrelevance: A User-effort Oriented Evaluation of Retrieval Systems without Predefined Retrieval Unit. In *RIAO*, 463–473.

[9] Abdigani Diriye, Ryen White, Georg Buscher, and Susan Dumais. 2012. Leaving so soon?: understanding and predicting web search abandonment rationales. In *CIKM*, 1025–1034.

[10] Susan T. Dumais, Georg Buscher, and Edward Cutrell. 2010. Individual differences in gaze patterns for web search. In *IIiX*, 185–194.

[11] Laura A. Granka, Thorsten Joachims, and Geri Gay. 2004. Eye-tracking analysis of user behavior in WWW search. In *SIGIR*, 478–479.

[12] Marti A. Hearst. 2009. *Search User Interfaces* (1st ed.). Cambridge University Press.

[13] Jeff Huang, Ryen W White, and Susan Dumais. 2011. No clicks, no problem: using cursor movements to understand and improve search. In *SIGCHI*, 1225–1234.

[14] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR*, 154–161.

[15] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. 2007. Evaluating the Accuracy of Implicit Feedback from Clicks and Query Reformulations in Web Search. *ACM TOIS* 25, 2 (2007), 1–27.

[16] Thorsten Joachims and Filip Radlinski. 2007. Search Engines That Learn from Implicit Feedback. *Computer* 40, 8 (Aug. 2007), 34–40.

[17] Kerstin Klöckner, Nadine Wirschum, and Anthony Jameson. 2004. Depth- and breadth-first processing of search result lists. In *SIGCHI extended abstracts*, 1539–1539.

[18] Jane Li, Scott Huffman, and Akihito Tokuda. 2009. Good Abandonment in Mobile and PC Internet Search. In *SIGIR*, 43–50.

[19] Lori Lorigo, Maya Haridasan, Hrönn Brynjarsdóttir, Ling Xia, Thorsten Joachims, Geri Gay, Laura Granka, Fabio Pellacini, and Bing Pan. 2008. Eye tracking and online search: Lessons learned and challenges ahead. *JASIS* 59, 7 (2008), 1041–1052.

[20] Lori Lorigo, Maya Haridasan, Hrönn Brynjarsdóttir, Ling Xia, Thorsten Joachims, Geri Gay, Laura Granka, Fabio Pellacini, and Bing Pan. 2008. Eye tracking and online search: Lessons learned and challenges ahead. *JAIST* 59, 7 (2008), 1041–1052.

[21] Andrew MacFarlane, George Buchanan, Areej Al-Wabil, Gennady Andrienko, and Natalia Andrienko. 2017. Visual Analysis of Dyslexia on Search. In *CHIIR*, 285–288.

[22] David Maxwell, Leif Azzopardi, Kalervo Järvelin, and Heikki Keskustalo. 2015. Searching and Stopping: An Analysis of Stopping Rules and Strategies. In *CIKM*, 313–322.

[23] Teemu Pääkkönen, Jaana Kekäläinen, Heikki Keskustalo, Leif Azzopardi, David Maxwell, and Kalervo Järvelin. 2017. Validating simulated interaction for retrieval evaluation. *Information Retrieval* (2017), 1–25.

[24] Filip Radlinski, Madhu Kurup, and Thorsten Joachims. 2008. How Does Clickthrough Data Reflect Retrieval Quality?. In *CIKM*, 43–52.

[25] Mark D. Smucker and Charles L.A. Clarke. 2012. Time-based calibration of effectiveness measures. In *SIGIR*, 95–104.

[26] Mark D. Smucker and Charles L.A. Clarke. 2016. Modeling Optimal Switching Behavior. In *CHIIR*, 317–320.

[27] Mark D. Smucker and Charles L. A. Clarke. 2012. Modeling User Variance in Time-Biased Gain. In *HCIR*. ACM, 1–10.

[28] Sofia Stamou and Efthimis N Efthimiadis. 2009. Queries without clicks: Successful or failed searches. In *SIGIR 2009 Workshop on the Future of IR Evaluation*. 13–14.

[29] Sofia Stamou and Efthimis N. Efthimiadis. 2010. Interpreting User Inactivity on Search Results. In *ECIR*, 100–113.

[30] Paul Thomas, Alistair Moffat, Peter Bailey, and Falk Scholer. 2014. Modeling Decision Points in User Search Behavior. In *IIiX*, 239–242.

[31] Wan-Ching Wu and Diane Kelly. 2014. Online search stopping behaviors: An investigation of query abandonment and task stopping. *Proceedings of the Association for Information Science and Technology* 51, 1 (2014), 1–10.

[32] Wan-Ching Wu, Diane Kelly, and Avneesh Sud. 2014. Using information scent and need for cognition to understand online search behavior. In *SIGIR*, 557–566.