Patterns of Search Result Examination: Query to First Action

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

To determine key factors that affect a user's behavior with search results, we conducted a controlled eye-tracking study of users completing search tasks using both desktop and mobile devices. We focus our investigation on users' behavior from their query to the first action they take with the search engine results page (SERP): either a click on a search result or a reformulation of their query. We found that a user deciding to reformulate a query rather than click on a result is best understood as being caused by the user's examination pattern not including a relevant search result. If a user sees a relevant result, they are very likely to click it. Of note, users do not look at all search results and their examination may be influenced by other factors. The key factors we found to explain a user's examination pattern are: the rank of search results, the user type, and the query quality. While existing research has identified rank and user types as important factors affecting examination patterns, to our knowledge, query quality is a new discovery. We found that user queries can be understood as either of *weak* or *strong* quality. Weak queries are those that the user may believe are more likely to fail compared to a strong query, and as a result, we find that users modify their examination patterns to view fewer documents when they issue a weak query, i.e. they give up sooner.

KEYWORDS

Query abandonment; Requery behavior; User behavior; User study

ACM Reference Format:
Mustafa Abualsaud and Mark D. Smucker. 2019. Patterns of Search Result Examination: Query to First Action. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM '19), November 3–7, 2019, Beijing, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3357384.3358041

1 INTRODUCTION

Given a set of search results, we know that as the rank of the topmost relevant result increases, the probability increases that a user will not click on the relevant result and will instead reformulate and requery to get fresh search results [21]. In this paper, we use eye tracking to better understand the underlying causes of these requeries without clicks and direct our study to user behavior from the query to the user's first action: either a click on a search result or a requery.

Our study is motivated by Zhang et al. [21]. Like Zhang et al., we allow users to freely query our search engine and control the search results to allow either only one relevant result at ranks 1-10 or no relevant results in response to a user's first query. If a user requeries, the search engine defaults to a commercial search engine's results. We include eye-tracking in our study to be able to know what users do and do not examine. We also investigate an important part that was missing from Zhang et al. and others, that is, user queries and their influence to examination behavior and decisions to requery.

While it is well known that users are less likely to examine lower ranked search results, we show that regardless of rank, if a user sees a relevant result, the user will click it with high probability. We confirm Zhang et al.'s hypothesis that the *exhaustive* and *economic* user types as characterized by Aula et al. [1] play a significant role in understanding requeries without clicks. What drives a user's examination to end their search process at certain ranks in the search result? We find that certain ranks and display issues affect user examination patterns, but most interestingly we found that the quality of a user's query appears to be known to the user and the user will modify their examination pattern based on query quality. This gives us an understanding of how likely are people going to examine certain ranks under different types of query quality and can be seen as motivation to design effectiveness measures that include factors other than the relevance of search results.

In particular, we show that:

- The first three search results are special. If a user issues a query unlikely to produce good results, the user is more likely to requery after finding the top three results to be non-relevant than if the user had issued a query expected to produce good results. If a relevant document is in the first three search results, the user will click on it.
- If the user is an exhaustive user, they are less influenced by the quality of their queries and are more persistent than economic users. Rank has much less effect on their likelihood of viewing a relevant result than it does for economic users.
- Economic users, are unlikely to scroll and view search results off of the page, and thus are likely to requery when the topmost relevant result is below the page fold.

In addition to these findings, we also show that for mobile search, users are likely to scroll to view the first five results, but if a relevant result is not seen, they will then requery. We also show a decision tree model that uses the factors of rank, user type, and query quality and demonstrates the importance of these factors to understanding a user's decision to click or requery as their first action. The decision tree provides a holistic view of users interactions with search engines.

CIKM '19, November 3–7, 2019, Beijing, China

ACM ISBN 978-1-4503-6976-3/19/11…$15.00
https://doi.org/10.1145/3357384.3358041

2 RELATED WORK

Query abandonment, or the decision to not click at any search result, can be seen as negative signals that indicate failure or dissatisfaction [18]. In an effort to understand why people abandon their queries, Stamou and Efthimiadis [17] employed a survey to study search tasks without clickthroughs. The authors categorized the causes of abandonment as *intentional* and *unintentional.* Intentional causes are encountered with a predetermined intention to look for answers in the search results' snippets and unintentional causes can be due to irrelevant results, already seen result or interrupted search. The authors conducted a follow up study where they collected queries from users conducting their daily searches in the web [18]. Their study focused on determining the impact of query abandonment to users clicking behaviour and satisfaction. Stamou and Efthimiadis show that approximately 50% of the queries that did not trigger any clicks are queries with non-relevant results negatively influenced users.

Diriye et al. [6] extended previous work by conducting a much larger user study that collected abandonment rationals at abandonment time by using a browser plugin that prompts participants with survey questions right after a query is abandoned. The authors reported that 7% of abandonments were due to unintentional causes and around 5% of the abandonments were due to the participants deciding to reformulate their query to a better query i.e. a better query came to their mind. Both Diriye et al.'s and Stamou and Efthimiadis' work is focused on providing reasons why users abandon their queries. Our paper is focused connecting different patterns of user examination with query abandonment.

In another body of work, Wu et al. [20] and others [4, 14, 15] used information forging theory to better understand how users seek information in the web. Wu et al. [20] manipulated the number of the relevant documents in the search results of users first 3 queries', and asked users to search for relevant documents to open-ended question. The authors have found that the number of relevant documents in the SERP can affect the rate of search abandonment and the number of query reformulations users perform. This was due to the nature of their search tasks, as some of the topics are opinion-based and require multiple queries to complete. Ong et al. [15] used the same tasks and manipulation technique as in [20] to study differences between desktop and mobile search behavior. Information forging theory, as others have shown, can be a useful technique to study search behavior.

Several studies have used eye-tracking as a tool to better understand behavior [1, 5, 7–10, 12, 13]. Klöckner et al. [12], Dumais et al. [7] and Aula et al. [1] looked at individual differences in user examination patterns. Based on eye-tracking data, Aula et al. classified users as either *economic*: users who examine few items and are quick to make decisions or *exhaustive*: users who examine more items and even scroll below the page fold to view more items.

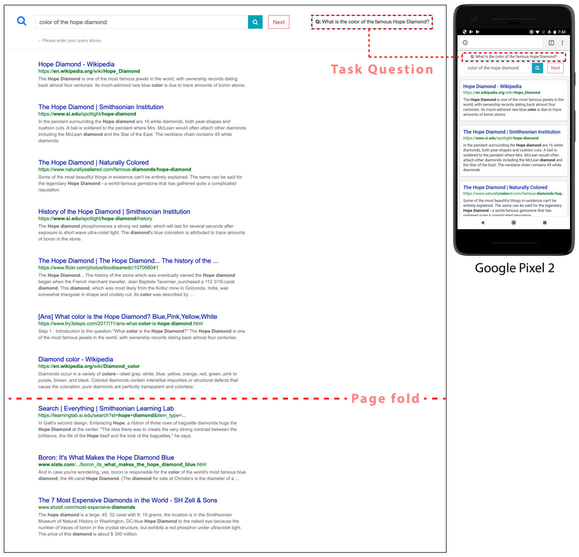
Cutrell and Guan [5] looked at how varying the amount of information in the search snippet affects user examination in both informational and navigational tasks. In particular, they found that increasing the amount of information in the snippets helps with informational queries but can hurt performance for navigational tasks. They also manipulated the search results to include what the authors describe as "best" search result item and looked at the fraction of times participants looked at it [9]. The placement of the "best" search result was either at the top, mid or bottom of the list. They report that as the rank of the target "best" search result decreases from the top to the bottom of the list, the chances of users clicking at it decreases and may be related to their probability of examining it.

Of particular interest is the study of Joachims et al. [10], where users were provided Google results to answer informational and navigational questions. Subjects were assigned to one of three experiment conditions. Either the results were not manipulated, manipulated by swapping the first two results, or by reversing the results order. They found that users are likely to click on higher ranking items irrespective of relevance and the performance of the search engine. The number of relevant results in the search list, however, is not controlled and could contain multiple relevant documents, which could be a possible influence to which document a user clicks.

Our work differs from previous work as follows:

- Unlike Joachims et al. [10] and others [14, 15, 20] where many of their search tasks include multiple relevant documents in the SERP, our focus is not in investigating which document among those that are relevant should the user click at, but on understanding possible causes to how far are users are willing to examine SERPs with either no relevant document or one relevant document placed at different ranks and what motivates users to continue or stop their examination. We aim to understand how different reasons to query abandonment can affect examination and vice-versa, i.e. how examination patterns can influence a person to abandon their query.
- Guan and Cutrell [9] concluded that the low click probability on what the authors described as the "best" search result is caused by their probability of examining it. We investigate factors that influence their examination and cause them not to look at the "best" search result when it is placed in any of the 10 ranks of the search result, as opposed to two ranks from the top, mid and bottom areas of the search results as in their study.
- Our work extends the work of Diriye et al. and Aula et al. by not only integrating different types of users' examination behavior, but also integrating queries, an essential part to the search process, into its influence on users' examination behavior and decisions to requery.
- We provide a holistic view of the search process and abandonment, encompassing three important parts, users, queries and search results, and show the influence of users and queries to each other at specific ranks in the search result. This has important implications on designing a more comprehensive effectiveness measures that also include users and queries into the evaluation.

We focus on negative search abandonment, where users abandon search results without achieving their intended search goal and use eyetracking interaction data to help us understand the behavior exhibited during query abandonment.



color of the hope diamond

Next

Q: What is the color of the famous Hope Diamond?

Please enter your query above.

Hope Diamond - Wikipedia
https://en.wikipedia.org/wiki/Hope_Diamond
The Hope Diamond is one of the most famous jewels in the world, with ownership records dating back almost four centuries. Its much-admired rare blue color is due to trace amounts of boron atoms.

The Hope Diamond | Smithsonian Institution
https://www.si.edu/spotlight/hope-diamond
In the pendant surrounding the Hope diamond are 16 white diamonds, both pear-shapes and cushion cuts. A bail is soldered to the pendant where Mrs. McLean would often attach other diamonds including the McLean diamond and the Star of the East. The necklace chain contains 45 white diamonds

The Hope Diamond | Naturally Colored
https://www.naturallycolored.com/famous-diamonds/hope-diamond
Some of the most beautiful things in existence can't be entirely explained. The same can be said for the legendary Hope Diamond - a world-famous gemstone that has gathered quite a complicated reputation.

History of the Hope Diamond | Smithsonian Institution
https://www.si.edu/spotlight/hope-diamond/history
The Hope diamond phosphoresces a strong red color, which will last for several seconds after exposure to short wave ultra-violet light. The diamond's blue coloration is attributed to trace amounts of boron in the stone.

The Hope Diamond | The Hope Diamond... The history of the ...
https://www.flickr.com/photos/bootbearwdc/107006041
The Hope Diamond... The history of the stone which was eventually named the Hope diamond began when the French merchant traveller, Jean Baptiste Tavernier, purchased a 112 3/16-carat diamond. This diamond, which was most likely from the Kollur mine in Golconda, India, was somewhat triangular in shape and crudely cut. Its color was described by ...

[Ans] What color is the Hope Diamond? Blue,Pink,Yellow,White
https://www.try3steps.com/2017/11/ans-what-color-is-hope-diamond.html
Step 1 : Introduction to the question "What color is the Hope Diamond?" The Hope Diamond is one of the most famous jewels in the world, with ownership records dating back almost four centuries.

Diamond color - Wikipedia
https://en.wikipedia.org/wiki/Diamond_color
Diamonds occur in a variety of colors—steel gray, white, blue, yellow, orange, red, green, pink to purple, brown, and black. Colored diamonds contain interstitial impurities or structural defects that cause the coloration, pure diamonds are perfectly transparent and colorless.

Search | Everything | Smithsonian Learning Lab
https://learninglab.si.edu/search?st=hope+diamond&item_type=...
In Galli's second design, Embracing Hope, a ribbon of three rows of baguette diamonds hugs the Hope Diamond at the center. "The idea there was to create the very strong contrast between the brilliance, the life of the Hope itself and the look of the baguettes," he says.

Boron: It's What Makes the Hope Diamond Blue
www.slate.com/.../boron_its_what_makes_the_hope_diamond_blue.html
And in case you're wondering, yes, boron is responsible for the color of the world's most famous blue diamond, the 46-carat Hope Diamond. (The diamond for sale at Christie's is the diameter of a ...

The 7 Most Expensive Diamonds in the World - SH Zell & Sons
www.shzell.com/most-expensive-diamonds
The hope diamond is a large, 45. 52 carat with 9. 10 grams, the location is in the Smithsonian Museum of Natural History in Washington, DC-blue Hope Diamond to the naked eye because the number of traces of boron in the crystal structure, but exhibits a red phosphor under ultraviolet light. The price of this diamond is about $ 350 million.

Figure 1: The search interface fits seven results on the desktop monitor and three on our mobile device, a Pixel 2. The Pixel 2's actual size relative to the desktop's size is as shown.

3 METHODS AND MATERIALS

In this section, we describe the study procedure, search tasks, search interface, the method of controlling the quality of search results, and other details of our experiment.

3.1 Search tasks

Participants were asked to complete 12 search tasks each comprising of a single factoid question. To complete a search task, a user needed to use our custom search engine to find an answer to a question. Participants were allowed to enter as many queries as they wished and were told to search as long as they needed. We instructed participants to stop once they were confident they had found an answer and to say the answer out loud to the researcher. Each task ended once a participant said their answer.

Table 1 shows the complete list of search task questions and their answers. We used the same questions as in [21] except for one question which we replaced with Q12. Many participants were not able to provide the correct answer for that particular question in [21].

3.2 Search interface

Figure 1 shows the search interface for desktop and mobile. For both interfaces, the search task question is shown at the top of the page and a search box is provided to allow users to query the search engine. The search box does not provide query suggestions. After a user submits a query, both interfaces show 10 results with no pagination, i.e. users cannot click to view a second page of results. For the desktop interface, the page fold line is after the seventh SERP result, and for the mobile interface, the page fold is after the third result. The page fold line represents the point below which the search results are off the screen and the user must scroll to see those results.

3.3 Controlling Search Results Quality

For each search task a participant performed, we returned a manipulated SERP, i.e. treatment. Each treatment consists of a different manipulation of SERP quality:

- For ten of the treatments, the SERP contained 1 relevant result and 9 non-relevant results. A relevant result contains the correct answer in the corresponding web page. We placed the relevant results at ranks 1-10 and denote these tasks as **Correct@1, ... Correct@10**.
- For one treatment, the SERP contained 10 non-relevant results and we denote this task as **NoCorrect** (NC for short).
- For one treatment, the SERP result contained results returned by the Bing API[1] without any manipulation, denoted as **Bing**.

3.3.1 *When are manipulated SERPs shown?*

We returned a manipulated SERP result to a participant once they submitted a query with any terms that we deemed to be relevant to the current search task's question. For example, Q8 in Table 1 asks for the number of chapters in the Art of War book by Sun Tzu. The relevant query terms for this question Q8 are: Art, War, Sun, Tzu. Similarly, for question Q3, the query terms are: Earth and Day. We constructed relevant terms for each question prior to the study. A manipulated SERP was shown only once per search task and all other submitted queries returned results from the Bing search API.

3.3.2 *How are manipulated SERPs constructed?*

The manipulated SERP results for each question were constructed prior to the study. For each question, we manually used the Bing API to search for a document that contains the correct answer and is easy for readers to extract. We selected this document as our relevant document for the question, but we made sure not to show the correct answer in the snippet. We did the same to find 10 non-relevant documents. For non-relevant documents, we looked for documents that contain relevant words but their content is obviously not relevant to the question. For example, for Q8 (the Art of War chapters), non-relevant documents can be about books with similar titles and by different authors. Such documents contain relevant words but their content is not relevant to the question.

With our single relevant document and our set of non-relevant documents, we constructed manipulated SERPs as follows:

- For treatments **Correct@1, ... Correct@10**, we placed the relevant document at the corresponding rank, and randomly filled the rest of the results with our non-relevant documents.
- For the **NoCorrect** treatment, we randomly positioned the 10 non-relevant documents.

For the **Bing** treatment, we did not manipulate the results and directly returned the Bing API results. Throughout this paper, we use the term relevant and correct SERP result interchangeably to indicate the relevant document with the correct answer.

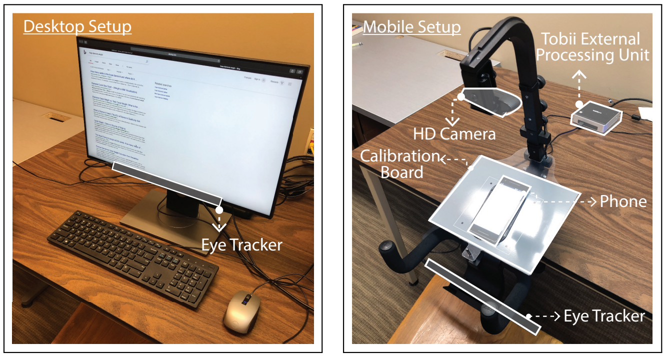
3.4 Study Design, Methods and Procedure

The study was conducted in a private office with a desktop computer and a mobile phone (Figure 2). We used a standard 23.5" desktop monitor and a Google Pixel 2 phone.

[1] azure.microsoft.com/services/cognitive-services/bing-web-search-api/

Table 1: Search task questions and their answers.

QID	Question	Answer
P	What is the colour of Hope Diamond?	Fancy Dark Grayish Blue.
Q1	How long is the Las Vegas monorail in miles?	3.9 Miles.
Q2	Find out the name of the album that the Mountain Goats band released in 2004.	We Shall All Be Healed.
Q3	Which year was the first Earth Day held?	1970.
Q4	Which year was the Holes (novel) written by Louis Sachar first published?	1998.
Q5	Find the phone number of Rocky Mountain Chocolate Factory located in Ottawa, ON.	(613) 241-1091
Q6	What is the name of opening theme song for Mister Rogers' Neighborhood?	Won't You Be My Neighbor?
Q7	Which album is the song Rain Man by Eminem from?	Encore.
Q8	How many chapters are in The Art of War book written by Sun Tzu?	13 Chapters.
Q9	What is the scientific name of Mad cow disease?	Bovine Spongiform Encephalopathy, or BSE
Q10	How many campuses does the University of North Carolina have?	17 Campuses.
Q11	Which Canadian site was selected as one of United Nations World Heritage Sites in 1999?	Miguasha National Park.
Q12	What is the first studio album Rihanna has released?	Music of the Sun.



(a) Desktop eye tracking. (b) Mobile phone eye tracking.

Figure 2: Desktop and Mobile phone eye tracking setup.

3.4.1 Balanced design. In total, there are 12 different treatments and 12 different topics. We used a 12×12 Graeco-Latin square to balance search topics and treatments across task order. The 12×12 Graeco-Latin square forms a single block where each row represents the order of tasks a participant undergo. By randomizing the rows and columns of the block, we created three separate blocks for our participants. Each study participant saw each search topic and treatment once.

3.4.2 Eye tracking. The eye tracker used was Tobii Pro X3-120. The eye tracker is screen-based (Figure 2a) but can be mounted in a custom Tobii Mobile Device Stand as shown in Figure 2b. The sampling rate of the eye tracker is 120 Hz, which allows detailed research into the timing and duration of fixations. The eye tracker also comes with Tobii Studio[2]/Pro Lab[3] software that enables the use of the eye tracker in the Internet Explorer web browser, which we used for the desktop. We used the Google Chrome browser on the mobile device. We used Tobii Studio Lab and Tobii Pro software to calibrate the eye tracks for each participant and for analyzing fixation data.

3.4.3 Implementation. The web application used in the study was implemented in Python and JavaScript. JavaScript was used to record various user behavior such as clicks and mouse moves. The web server was hosted locally and accessed with a web browser.

3.4.4 Procedure. Before the participants started the study, we asked them to sign an informed consent form. We then began by calibrating the eye tracker. If the calibration was not successful, the participant was given $5 for their time. We started the study by collecting demographics and general information on participants' experience with search engines. We provided participants with a tutorial on the study and how to use our search engine. Participants were asked to turn off their phone, sit comfortably and try not to make strong body movements as it would prevent the eye tracker from capturing their eyes. Each participant began with the practice question (question P in Table 1). The SERPs during the practice task are not manipulated. After the practice task, participants continue with the main study of twelve tasks. A study task is comprised of a pre-task questionnaire, a search task, and a post-task questionnaire. We provided the current search task question to the participant during the pre-task and asked the participant their perceived difficulty and familiarity of the question and whether they already know the correct answer. We also showed the question to the participant in the search interface (Figure 1). We repeated this process 12 times, each for a different search task question. After completing the twelve tasks, each participant filled out an exit questionnaire about their experience.

3.5 Participants

After receiving ethics approval from our university's Office of Research Ethics, we recruited people through posters posted across our university. We began by collecting data for the mobile search setup. We recruited 22 people, 3 of whom were for pilot testing. We successfully calibrated and ran the study for 11 people on the mobile device. We did not use the remaining participants because of poor eye tracker calibration for tall participants, participants with eyeglasses, or participants with some eye condition or disorder. After noticing these issues and to prevent such scenarios, we added extra requirements to participate in the study including an overlooked requirement that participants should be fluent English

[2] tobiipro.com/product-listing/tobii-pro-studio
[3] tobiipro.com/product-listing/tobii-pro-lab

speakers. To avoid calibration problems, we required future participants to not wear eyeglasses, to not have long eye lashes, to not wear mascara, and to not have any eye condition or disorder.

For desktop search, we recruited 30 participants, but we used only 24 participants' data in our study. We were unable to calibrate the eye tracker for 5 participants, and one participant was for pilot testing the setup.

Our participants were university students: 15 females, 19 males and 1 who prefers another term. 27 students were enrolled in an undergraduate program and 8 in a graduate program. Their average age is 20.48, with a minimum age of 17 and a maximum age of 30. Their majors are 3 in art, 1 in environment, and 31 in a STEM major.

We advertised that participants would be remunerated $10 for their time and $15 if they were able to answer 10 out of the 12 questions correctly. However, each participant was given $15 dollars regardless of how many correct answers they have provided. This was done to add some incentive to participants to engage more in the study. After analyzing the participants' data, 29 participants answered the 12 questions correctly and the lowest score was 10 correct answers.

3.6 Collected Measurements

Submitted queries: All queries submitted to the search engine by the participants during their 12 tasks.
Action: The action made by the user once they are shown the manipulated/Bing SERP. An action could be a requery, a document click, or a *snippet answer*. For document clicks during manipulated SERPs, we record whether or not the clicked search result was relevant. A snippet answer indicates a participant has announced their answer to the question by reading the snippet of a search result without clicking on the result. The items in our manipulated SERPs do not contain the correct answer in their snippet, but the Bing search results can directly contain answers.
Time to action: Time to action is measured from the moment the result is shown to the user to the moment the action is triggered (e.g. clicking a document, clicking the search bar, or time of announcing the answer from a snippet). In few cases, participants clicked the search bar then started looking at SERP results. In these cases, the end time of the action is their first keystroke in the search bar.

The time period between time to action is important as it involves the decision making process by the user. Measurements described below are recorded within this time period.
Mouse moves: The number of mouse moves the participant has made. Two consecutive mouse moves include a ≥ 200 ms idle period between each other.
Number of Fixations: The total number of fixations made by a participant during a task.
Fixation Duration at SERP items: We create 10 areas of interest (AOI) using the Tobii software, one for each search result in the SERP. We record the total fixation duration a user looked at a search result using the fixation data.
Eye Fixation Sequence: The complete sequence of fixations at each search result. An example sequence is 1→2→1, which indicates a user has looked at the first search result, the second result then back to the first result. We define *Unique Fixation Sequence* as the sequence of unique search results fixated by the user.

4 RESULTS AND DISCUSSION

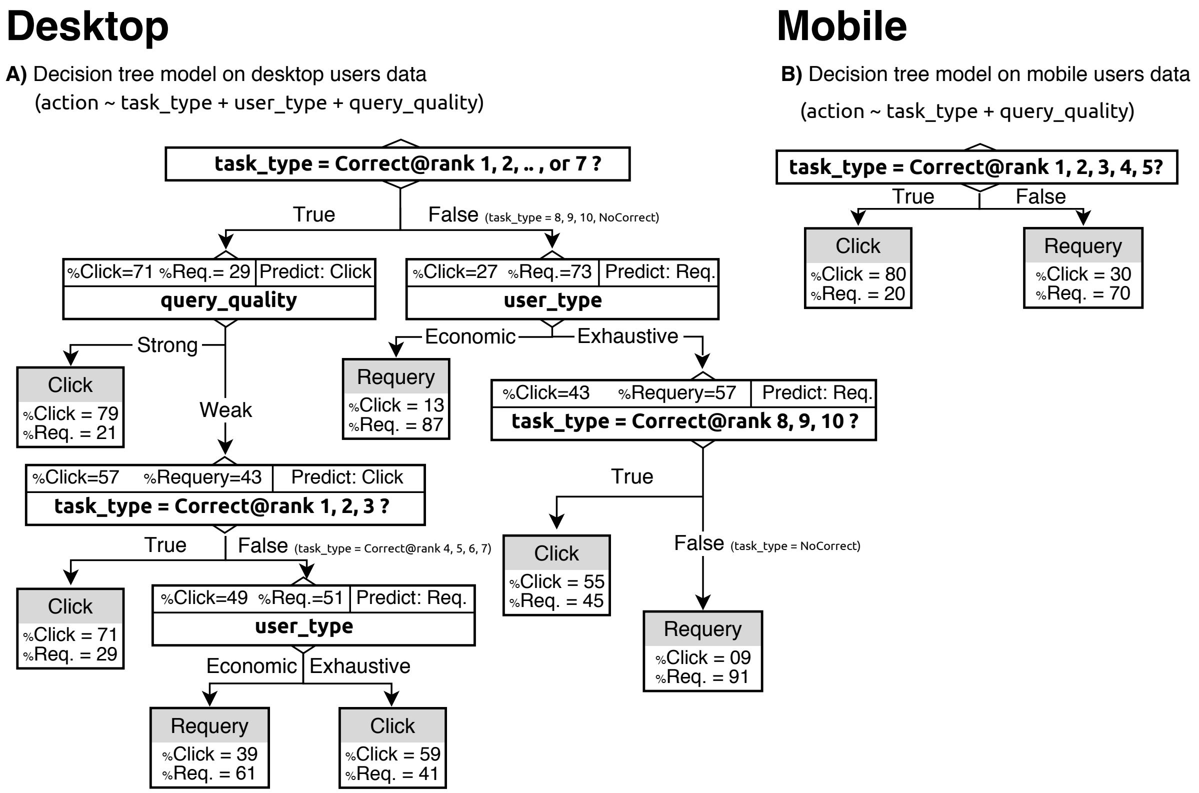
We have learned from Zhang et al. [21] that user type is important to understand probability of an immediate, zero-click requery. Zhang et al. [21] also hypothesized that their study participants could possibly be *economic* and *exhaustive* users as described in previous eye-tracking studies [1, 7], but they lacked eye-tracking data to confirm their hypothesis. While observing our users complete their tasks, we indeed noticed *economic* and *exhaustive* behavior described in previous eye-tracking studies [1, 7]. *Economic* users tend to examine fewer results and give up quicker than *exhaustive* users. This observation motivated us to classify users as either economic or exhaustive using our eye tracking fixation data, as we describe in Section 4.2.

During our observation of users completing their tasks, we also noticed unique behavior users seem to follow. In many cases, we saw participants enter a query, examine few of the top results but not click on any, and then make a decision to requery. We hypothesize that a possible reason behind the requery decision is due to the quality of their queries. To illustrate, in one case during question Q4 (Publication date of Holes by Louis Sachar), a participant entered the query "holes novel" and examined, without clicking, the first three search results and then reformulated their query. We think that after the user examined a few search results, the user might have realized their query is under-specified and likely to fail and therefore decided to stop their SERP examination. The user's reformulated query included the author's name. This motivated us to assess query quality in terms of specificity to the question and check whether it affects user behavior (see Section 4.3).

From our observations, we understood that user type, query quality, and rank of the topmost relevant result are factors that likely affect how people examine search results and their decision to click or requery. The question then arises: given a search result and our knowledge of users and queries, when and where do these factors start to matter?

We used decision trees to aid in understanding the overall behavior of users. Decision trees are known for capturing interactions between variables while providing a simple interpretation of the data [3]. We model users' first action, i.e. whether users click a search result or requery. Input to the decision tree consists of the rank of the topmost relevant result (1-10, No Correct), user type (economic or exhaustive, see Section 4.2), and query quality (weak or strong, see Section 4.3). We built decision trees using the recursive partitioning algorithm [3] as implemented in the rpart [19] package in R. The rpart algorithm works by recursively partitioning the data into multiple nodes and selecting splits based on node impurity. We used information gain as our splitting index. *Bing* treatments were excluded from the decision tree modeling for we do not control the quality of the SERP and thus do not know the rank of the topmost correct result.

Figure 3A shows the decision tree produced for desktop search. The model selects whether or not a relevant result is above or below the page fold (task type = Correct@1, . . . , or 7) as the root of the decision tree, which means this is the most important information to predict whether a user will click or requery. When the topmost relevant result is below the page fold, economic users will requery


Desktop
A) Decision tree model on desktop users data
(action ~ task_type + user_type + query_quality)
task_type = Correct@rank 1, 2, .. , or 7 ?
True
False (task_type = 8, 9, 10, NoCorrect)
%Click=71 %Req.= 29 Predict: Click
query_quality
%Click=27 %Req.=73 Predict: Req.
user_type
Strong
Click
%Click = 79
%Req. = 21
Weak
Economic
Exhaustive
Requery
%Click = 13
%Req. = 87
%Click=43 %Requery=57 Predict: Req.
task_type = Correct@rank 8, 9, 10 ?
%Click=57 %Requery=43 Predict: Click
task_type = Correct@rank 1, 2, 3 ?
True
False (task_type = Correct@rank 4, 5, 6, 7)
Click
%Click = 71
%Req. = 29
%Click=49 %Req.=51 Predict: Req.
user_type
Economic
Exhaustive
Requery
%Click = 39
%Req. = 61
Click
%Click = 59
%Req. = 41
True
Click
%Click = 55
%Req. = 45
False (task_type = NoCorrect)
Requery
%Click = 09
%Req. = 91
Mobile
B) Decision tree model on mobile users data
(action ~ task_type + query_quality)
task_type = Correct@rank 1, 2, 3, 4, 5?
True
False
Click
%Click = 80
%Req. = 20
Requery
%Click = 30
%Req. = 70

action: Whether a user clicks at a document or requery without any clicks.
task_type: Determines the position of the correct result in the SERP. NoCorrect means there are no correct result in the SERP.
user_type: A user is classified is either economic or exhaustive based on fixation data (Section 4.2).
query_quality: A query can be either weak or strong. Weak queries are under-specified to the task question (Section 4.3).

Figure 3: Decision tree models on desktop and mobile.

87% of the time, while exhaustive users are more likely to click on a result.

When the topmost relevant result is above the page fold, query quality becomes important to predicting a click or requery. A strong query means that 79% of the time a user will click. In contrast, a user's behavior changes for a weak query based on whether or not the topmost relevant result is found in the top 3 ranks. With weak queries, if the topmost result in the top 3 ranks, users are likely to click (71%). For weak queries and results at ranks 4-7, behavior again depends on user type. Economic users are more likely to requery (61%) than click on a result at ranks 4-7 if they issue a weak query, while exhaustive users are more like to click (59%) than requery.

Because query quality affects user behavior, researchers should consider whether it is appropriate to supply queries to users in studies as opposed to allowing users to interactively query the search engine. Many controlled user studies make use of a fixed set of results produced from a fixed "query", but such fake queries hide that users appear to have an internal sense of the quality of their query and thus modify their behavior appropriately.

Figure 3B shows the decision tree for mobile search. In mobile, the position of the correct SERP items is the only important factor determining a user's action, with rank 1 to 5 being most important. Our smaller amount of user data for mobile search may limit the usefulness of the decision tree for understanding mobile search behavior.

Next, we look at requeries and examination, how we classified users and queries, and how users and queries influence examination.

4.1 Requeries and Examination

Figure 4 and Table 2 show the requery probability across ranks and during *NoCorrect* tasks, where there is no correct item in the SERP and *Bing* tasks, where the results are not manipulated. The probability of requery is high when we place the correct item at ranks 8 to 10 in desktop and in 6 to 10 in mobile. While increasing rank means an increasing probability of requerying rather than clicking, the question arises whether or not users are viewing results at higher ranks, viewing but ignoring documents at higher ranks, or some combination of the two. This question was raised by Guan and Cutrell [9] who concluded that in some cases users fail to look at results and in some cases they discount lower ranked results. An important difference between Guan and Cutrell's work and ours is that we more tightly controlled the search results to only have non-relevant and relevant results while they manipulated the rank of the "best" result while allowing the search engine to provide other results, which we presume may have also appeared somewhat relevant to the users.

To address this question, we measured how many times a user decided to requery when they have not seen the correct item. If it is the case that users requery often, this serves as a suggestive piece of evidence that an examination sequence, that does not include the correct item, causes users to requery. We used our eye tracking data to determine how much time users spent fixating at correct items and what their resulting action was.

Table 3 shows the frequency of requeries, clicks, and snippet answers (answers provided by reading the snippet alone) grouped by the duration of fixation at the correct item. The table also shows the frequencies during *NoCorrect* tasks and *Bing* tasks. We first notice that only two people clicked on wrong documents when there is no relevant document in the SERP and that the majority of users decided to requery. We also notice that 85% of the time a user would requery if they have not seen the correct item or quickly glanced over it, and that 88% of the time, a user would click at the correct document if they have examined it for ≥1 seconds. In summary, if a user sees the correct item, they click it, if they have not seen it, they requery.

To what extent does rank matter when they see the correct item? Table 4 shows the probability of a requery or clicking at a correct result when it has been seen (≥1 second), across the 10 ranks. The table shows that no matter where the correct item is, if the user sees it, they are more likely to click on it than requery. The results are also similar when we set the threshold to ≥200ms. Our results indicate that if a user sees the correct answer, rank does not seem to have an influence on their clicking decision.

Of note, we do not dispute previous research claiming that there exists a position bias or that searchers trust the ranking of search engines as Joachims et al. [10] have shown. When there are multiple relevant documents, as in Joachims et al. [10] study, the authors show that users trust and click higher ranking items irrespective of actual relevance. Users seem to trust the ranking presented by the search engine and click on what the search engine has chosen to be higher in the list. When the number and position of relevant documents in the SERP are controlled, as in our study, users are most likely to stop their examination and click on the relevant document once they see it. Figure 5 shows the mean number of items a user examines after seeing the correct item, clearly showing how examination stops.

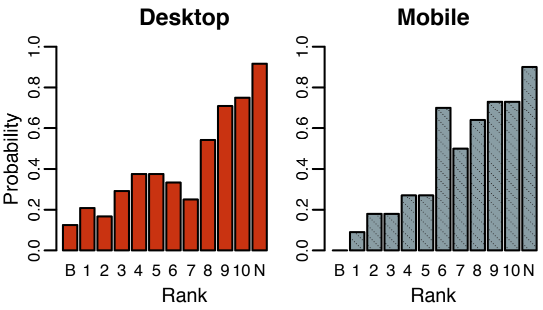


Figure 4: Prob. of a requery in desktop and mobile. X-axis indicates the position of the relevant item in the manipulated SERP. B indicates Bing and N indicates NoCorrect tasks.

Table 2: Probabilities of a requery action, click on wrong/correct SERP items, and average time to requery.

Correct Doc. Rank	Prob. of Requery		Prob. of Wrong Click		Prob. of Correct Click		Avg. Seconds to Requery	
	Desk	Mob	Desk	Mob	Desk	Mob	Desk	Mob
Bing	0.12	0.00	–	–	–	–	4.73	–
1	0.21	0.09	0.00	0.00	0.79	0.91	9.44	7.28
2	0.17	0.18	0.04	0.00	0.79	0.82	6.81	7.58
3	0.29	0.18	0.08	0.00	0.62	0.82	6.32	3.21
4	0.38	0.27	0.00	0.00	0.62	0.73	9.65	8.23
5	0.38	0.27	0.00	0.09	0.62	0.64	4.68	4.37
6	0.33	0.70	0.08	0.00	0.58	0.30	4.37	8.13
7	0.25	0.50	0.12	0.00	0.62	0.50	6.90	5.24
8	0.54	0.64	0.08	0.00	0.38	0.36	5.63	6.66
9	0.71	0.73	0.00	0.09	0.29	0.18	6.16	7.96
10	0.75	0.73	0.04	0.09	0.21	0.18	6.73	5.16
NC	0.92	0.90	0.08	0.10	0.00	0.00	8.66	6.65

Table 3: Frequency table of actions grouped by duration of fixation at the correct item. Data is for desktop users.

Time fixating at correct document	Requery	Click		Snippet Answer
		Wrong	Correct	
<200ms	64	10	1	0
≥200ms, <1sec	18	1	26	0
≥1sec	14	0	106	0
Frequencies in NoCorrect and Bing tasks				
NoCorrect	22	2	0	0
Bing	3	Total Clicks: 16		5

4.2 User Types

Motivated by our observations of users and previous research [1, 7, 12, 21], we investigated the possibility of different user types in our study. Previous research [1] has indicated that exhaustive users tend to be slower and spend more time analyzing search results than economic users. Using the fixation data we collected, we plot the distribution of the average total number of fixations during the search tasks. If there exist two types of users, we should be able to see a bimodal distribution of total fixations. While a strong bimodal distribution is missing, Figure 6 does show that the distribution of total fixations for desktop users has a large spread, and we selected those users with more than 650 fixations to be labeled exhaustive users while those with fewer than 650 fixations to be economic users. In total, we have 11 exhaustive users and 13 economic users. For mobile, we were unable to see a difference in the distribution, and we treat mobile users as equal in our analysis.

Table 4: Probability of requery or click on a correct (corr.) item when it has been seen (≥1 sec). SE reported in brackets.

Probability of	Rank of Correct Item									
	1	2	3	4	5	6	7	8	9	10
Desktop Users										
Corr. Click	.7[.2]	.9[.1]	.8[.2]	.9[.2]	1[0]	1[0]	.9[.1]	.9[.2]	.9[.2]	1[0]
Requery	.3[.2]	.1[.1]	.2[.2]	.1[.2]	0[0]	0[0]	.1[.1]	.1[.2]	.1[.2]	0[0]
Mobile Users										
Corr. Click	.9[.3]	.9[.3]	1[0]	.8[.4]	1[0]	0.8[.4]	1[0]	1[0]	1[0]	1[0]
Requery	.1[.3]	.1[.3]	0[0]	.2[.4]	0[0]	0.2[.4]	0[0]	0[0]	0[0]	0[0]

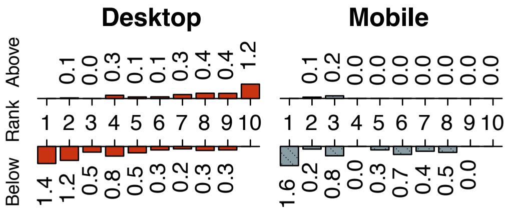


Figure 5: Mean number of unique SERP items looked at (fixated ≥200ms) after the user has seen the correct result and before they clicked on it. Bars above/below the rank indicate the mean number of items whose rank is above/below the correct result's rank. Users tend to stop their examination after seeing the correct result, and in the first two ranks in desktop, users stop their examination after examining about 1 more non-relevant document.

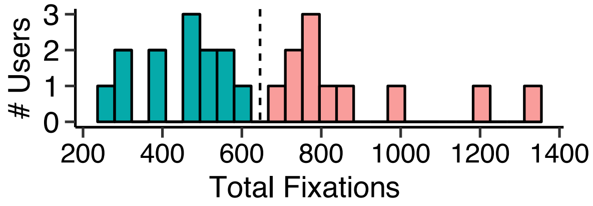


Figure 6: Total fixations histogram on desktop users.

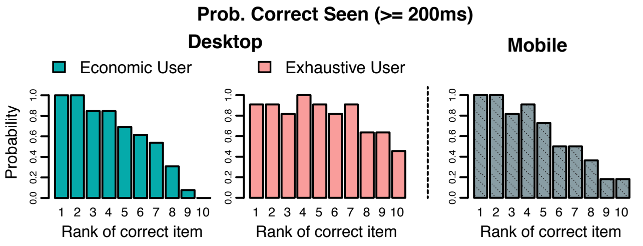


Figure 7: Probability of correct result seen.

Given the two types of user in desktop search, how different are their examination patterns when presented with our controlled SERPs? Given that exhaustive users tend to spend more time analyzing the SERP, we expect that they are more likely to see the correct result in the SERP. Figure 7 shows the probability of seeing the correct result when placed at different ranks for the different user types and for mobile. Indeed, exhaustive users' probability of seeing the correct result is higher than that of economic users. Their probability stays high and decreases a little when the correct result is placed below the fold (ranks 8 to 10). Economic users, on the other hand, are less likely to see the correct result as its position decreases in the list. Previous research has shown that people tend to examine search results in a linear fashion and the time required to reach a specific rank increases linearly [5, 10]. As economic users spend more and more of their time examining non-relevant items, they are more likely to stop their examination and thus not see the correct result. The result also explains why the page fold is an important factor: economic users rarely examine beyond the page fold but exhaustive users do. In mobile, it seems that people are willing to scroll beyond the page fold to look at more items, and the decrease of probability after the 5th result explains why the decision model picked *Correct@1..5* tasks as the decision criterion.

We investigated other differences of behavior in the two groups. Table 5:left shows the averages of different recorded measures and their statistical significance. Statistical significance was computed using generalized linear-mixed models as implemented in the lme4 package [2] for R [16], with study subjects and search questions treated as random effects. Total number of fixations, sequence length, and duration of fixation on top, middle, and bottom ranks are statistically significant. Exhaustive users tend to read 1.5 more unique results than economic users and produce 2.96 longer sequence lengths. The time spent fixating at different ranking areas is significantly less for economic users. The time to action is statistically significant with economic users showing a faster time to make a decision.

4.3 Query Analysis

As we explained earlier, we noticed that some users would examine only a few results and then requery with a more specific query. This motivated us to look into user queries and evaluate them based on specificity to the search task questions. We looked at all users' queries and noticed that there exists a number of queries we consider to be under-specified and that could possibly return non-relevant results if entered in a real search engine. For example, in Q4, one participant entered "holes novel" as their first query without any mention of the author name and possibly assumed that the search engine would know. Similarly, "art of war chapters" on Q8, "mad cow" on Q9, "mountain goats" on Q2, or "UN world heritage sites", "rain man album" on Q7, "north carolina campus" on Q10, and "canada united nations 1999" on Q11. All of these queries are missing important terms such as author name or location. We do not consider these queries to be completely bad, but we believe that they are of weak quality in terms of specificity and can be improved by including some of the most important or specific terms.

Based on preliminary analysis of queries, we decided to assess all queries based on their specificity/generality to the question. We consider a query to be either weak or strong. We wrote a short description and a tutorial of what we considered to be a weak or strong query and provided it to assessors that we hired for query

Table 5: Left: Averages and significant testing of different measures between the two user groups in the desktop setup (*/**/*** indicates statistical significance at p<0.05/0.01/0.001). Right: Averages grouped by user type and under different query types.

	Economic	**Exhaustive**	***p*-val**	**All Queries**	**Weak Queries**			**Strong Queries**		
	Users	**Users**		Both Users	Economic	Exhaustive	Both	Economic	Exhaustive	Both
# of fixations	448.9	869.7	***	641.7	391.5	774.3	552.0	479.3	909.7	684.6
Seq. length	5.2	8.2	***	6.6	4.2	7.0	5.4	5.8	8.7	7.2
Unique Seq. length$_{>0ms}$	4.3	5.7	***	4.9	3.5	5.5	4.3	4.7	5.7	5.2
Unique Seq. length$_{>=200ms}$	3.8	5.3	***	4.5	3.1	5.0	3.9	4.1	5.3	4.7
top_ranks_duration$_{1,2,3}$	1.7	2.9	***	2.2	1.6	2.4	2.0	1.7	3.1	2.4
mid_ranks_duration$_{4,5,6,7}$	1.2	2.5	***	1.4	0.7	2.3	1.4	1.4	2.5	2.0
bottom_ranks_duration$_{8,9,10}$	0.2	1.4	***	0.8	0.1	1.2	0.6	0.3	1.5	0.9
correct_item_duration$_{>0ms}$	1.3	2.2	***	1.8	1.1	2.2	1.7	1.4	2.3	1.8
Time to action	5.1	9.5	***	7.1	4.5	8.8	6.3	5.4	9.9	7.5
Time to requery	5.0	10.2	***	6.8	4.3	8.1	5.5	5.6	11.4	7.7
Time to click	5.1	9.2	***	7.3	4.8	9.2	7.1	5.3	9.3	7.4
# Mouse moves	2.3	4.7	**	3.4	2.1	4.4	3.1	2.4	4.8	3.6
# Query terms	3.8	4.6	**	–	–	–	–	–	–	–

assessment. In the description, we mentioned that a strong query *"includes important terms in the question and you consider specific enough to the question. The query is not ambiguous."* And a weak query is *"is not specific enough, can lead to off topic results, can be considered ambiguous or contain any misspellings."* The tutorial included examples of different questions and what would be considered weak and strong. After the first assessor completed their judging, we hired a second assessor for verification and to test agreement. The two assessors were fluent in English and had good experience in using search engines. Both assessors had not participated in the study nor were aware of the purpose of the study. We asked assessors to be careful assessing the queries and take as much time as needed. The assessors were given $20 for their time.

The total number of queries is 12 × 24 (tasks × participants) for desktop and 12 × 11 for mobile. The two assessors took about 1 hour to finish assessing all queries. The Krippendorff's alpha [11] for inter-rater reliability of the two assessors is 0.80, which indicates substantial agreement [11]. Our two assessors judged 27% and 28% of the submitted queries as weak, accordingly. Since both assessors have substantial agreements, we use the first assessor's judgments as the final judgment for the query type.

The percentages of weak queries made by economic and exhaustive users are 52% and 48%, accordingly. Only 18% of the weak queries were queries with a misspelling. Table 6 shows the percentages of weak and strong queries among different groups. Exhaustive users are less likely to requery regardless of query type.

Figure 8 shows the probability of seeing the correct result under weak and strong query quality for both user types. Interestingly, query quality seems to have a stronger influence on economic users than exhaustive users, as we have shown in Figure 3. Under weak queries, if the correct result is placed at the top of the list, economic users' probability of seeing the correct result is high. Their probability of seeing the correct result quickly drops as the correct result is placed lower in the list. Exhaustive users, on the other hand, are willing to examine more results in weak queries. In strong quality queries, economic users are willing to go further in the list than if they have entered a weak query. If an economic user enters a weak query and the correct result happens to be not in the top of the list, they stop their examination and requery.

Table 6: Percentages of query types for queries ended with a requery, and for diff. user types. Data is for desktop users.

Query Type %	In requeries where correct not seen and task type != {NC, Bing}	In user type	
		Eco.	Exh.
Weak	47% (80% are by Economic users)	35%	30%
Strong	53% (71% are by Economic users)	65%	70%

Table 5:right shows the averages of our measures grouped by the type of query and user. We do not compute the number of query terms by query type as different questions may require different number of terms. Although query type is not a controlled variable, we notice that weak queries have fewer fixations and shorter time to action on average. The average time to requery for weak queries (excluding *NoCorrect* tasks) is 5.16 seconds and 7.42 for strong queries, and the result is statistically significant using unpaired two sided Student's t-test ($p = 0.017$).

4.4 Limitations

One of the limitations of our study is that we only investigated a single type of search task. Our participants were asked to search for simple factoid questions and were presented with tightly controlled SERPs. Our findings are based on such search tasks, but we do acknowledge that different types of search tasks, such as searching for answers to more complex questions [20], or having multiple relevant documents in the SERP as in Wu et al. [20] might result in different examination behavior. We choose simple factoid questions because we wanted to capture search behavior of participants searching for answers as they would normally do for topics they are familiar with. Asking participants to search for answers to more complex questions or questions they are unfamiliar with may result in different patterns of SERP examination, which could be interesting to further study but is not in the scope of our experiment. We purposely controlled our SERP to have a single relevant result for a

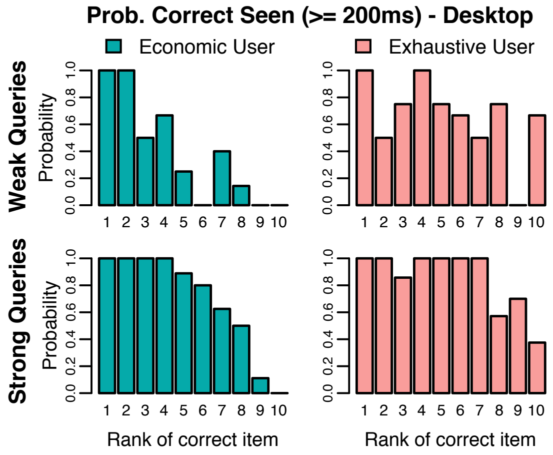


Figure 8: Probability of correct result seen under weak and strong queries for different user types.

reason: we wanted to understand how far people look before they decide to requery, and having multiple relevant documents in the SERP could be a complicating factor that we wanted to avoid. Our participant are young university students. Certainly older people might be different in their behavior in some way.

5 CONCLUSION

Our work investigates how rank, user type, and query quality influence a user's first action to be a click on a search result or to abandon the search results and requery. We controlled the position of a single relevant result in the search results and asked participants to search for answers to simple factoid questions. Doing so allowed us to understand how people examine search results and when they decide to requery.

With regard to user type, we confirm previous research showing different examination behavior among people [1]. The difference in user type allowed us to better understand how people examine and whether or not they see the correct result. Looking through how participants completed their task, we were motivated to understand how the quality of query can influence their examinations. We assessed queries based on specificity to the search question and investigated how query quality affects users' decisions and examination patterns. A new discovery from analyzing user data is unfolded. We noticed that the quality of the query appears to influence examination pattern and therefore query abandonment. We found that the first three search results are special. A weak quality query, a query that is under-specified to the question, influences users' examination patterns and particularly economic users. After entering a weak quality query, economic users examine the first three results and if a relevant document is not there, they reformulate their query to a stronger query that is not under-specified.

The use of a decision tree model (Figure 3) with user type, query quality and position of the correct result as variables to our model, helped us to understand which factors are important to users' click and requery decisions, and where each factor starts to matter in the SERP.

From our analysis, we understood that a strong factor influencing users' examination is the page fold, though its influence differs between user types, with economic users mostly examining what is shown above the fold. Most importantly, users seem to have an internal sense of the quality of their query, which starts to matter when the correct result is placed above the fold. The first 3 results are important to economic users when they submit a weak query. If the correct result is not there, they are more likely to abandon the search results.

ACKNOWLEDGMENTS

We thank Charles Clarke for his feedback. 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*, Maria Francesca Costabile and Fabio Paternò (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1058–1061.

[2] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, Articles* 67, 1 (2015), 1–48. https://www.jstatsoft.org/v067/i01

[3] Leo Breiman. 2017. *Classification and regression trees.* Routledge.

[4] Ed H. Chi, Peter Pirolli, Kim Chen, and James Pitkow. 2001. Using Information Scent to Model User Information Needs and Actions and the Web *(SIGCHI)*. ACM, New York, NY, USA, 490–497.

[5] Edward Cutrell and Zhiwei Guan. 2007. What Are You Looking for?: An Eye-tracking Study of Information Usage in Web Search *(CHI)*. ACM, New York, NY, USA, 407–416.

[6] Abdigani Diriye, Ryen White, Georg Buscher, and Susan Dumais. 2012. Leaving So Soon?: Understanding and Predicting Web Search Abandonment Rationales *(CIKM)*. ACM, New York, NY, USA, 1025–1034.

[7] Susan T. Dumais, Georg Buscher, and Edward Cutrell. 2010. Individual Differences in Gaze Patterns for Web Search *(IIiX)*. ACM, New York, NY, USA, 185–194.

[8] Carsten Eickhoff, Sebastian Dungs, and Vu Tran. 2015. An Eye-Tracking Study of Query Reformulation *(SIGIR)*. ACM, New York, NY, USA, 13–22.

[9] Zhiwei Guan and Edward Cutrell. 2007. An Eye Tracking Study of the Effect of Target Rank on Web Search *(SIGCHI)*. ACM, New York, NY, USA, 417–420.

[10] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately Interpreting Clickthrough Data As Implicit Feedback *(SIGIR)*. ACM, New York, NY, USA, 154–161.

[11] Krippendorff Klaus. 2004. Content analysis: An introduction to its methodology.

[12] Kerstin Klöckner, Nadine Wirschum, and Anthony Jameson. 2004. Depth- and Breadth-first Processing of Search Result Lists *(CHI EA)*. ACM, New York, NY, USA, 1539–1539.

[13] Yiqun Liu, Chao Wang, Ke Zhou, Jianyun Nie, Min Zhang, and Shaoping Ma. 2014. From Skimming to Reading: A Two-stage Examination Model for Web Search *(CIKM)*. ACM, New York, NY, USA, 849–858.

[14] David Maxwell and Leif Azzopardi. 2018. Information Scent, Searching and Stopping. In *Advances in Information Retrieval*, Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury (Eds.). Springer International Publishing, Cham, 210–222.

[15] Kevin Ong, Kalervo Järvelin, Mark Sanderson, and Falk Scholer. 2017. Using Information Scent to Understand Mobile and Desktop Web Search Behavior *(SIGIR)*. ACM, New York, NY, USA, 295–304.

[16] R Core Team. 2014. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org

[17] 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.

[18] Sofia Stamou and Efthimis N. Efthimiadis. 2010. Interpreting User Inactivity on Search Results. In *Advances in Information Retrieval*, Cathal Gurrin, Yulan He, Gabriella Kazai, Udo Kruschwitz, Suzanne Little, Thomas Roelleke, Stefan Rüger, and Keith van Rijsbergen (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 100–113.

[19] Terry M Therneau, Elizabeth J Atkinson, et al. 1997. An introduction to recursive partitioning using the RPART routines.

[20] Wan-Ching Wu, Diane Kelly, and Avneesh Sud. 2014. Using Information Scent and Need for Cognition to Understand Online Search Behavior *(SIGIR)*. ACM, New York, NY, USA, 557–566.

[21] Haotian Zhang, Mustafa Abualsaud, and Mark D. Smucker. 2018. A Study of Immediate Requery Behavior in Search *(CHIIR)*. ACM, New York, NY, USA, 181–190.