The Effect of Queries and Search Result Quality on the Rate of Query Abandonment in Interactive Information Retrieval

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

ABSTRACT

When a search result does not satisfy a user's needs, the user often abandons their query and submits a reformulated query in the hopes of receiving better search results. The action of abandoning search results is termed "query abandonment", and previous research has indicated possible reasons for this action, such as dissatisfaction with the result or coming up with a better query. Query abandonment can be seen as negative or positive signals. As we move closer to understanding when and what causes a user to abandon their query under different qualities of search results, we move forward in the development of overall understanding of user behaviour with search engines. This can be helpful for developing more accurate evaluation measures and better methods of collecting relevance judgments. In our dissertation project, we plan to study how the quality of both queries and search results can affect the rate and time users abandon their queries. Specifically, we discuss experiments to investigate the rate at which users abandon their query at different levels of search quality, and whether user examination is affected by the type of non-relevant results. User behaviour in our studies will be analyzed with a combination of eye-tracking and questionnaires. The use of eye-tracking will accurately measure user attention to different elements in the search engine result page (SERP) and how far in the search results a user examines before abandoning the query. Questionnaires will provide more insights from users prospective into the form and quality of queries they submit.

KEYWORDS

Query Abandonment, Interactive Information Retrieval, User Study

1 INTRODUCTION & MOTIVATION

After having examined search results of a particular query, a user may decide to abandon their query without clicking at any of the search results. This action of abandoning the search result is called query abandonment and can be due to various reasons, some of which can indicate satisfaction or dissatisfaction with the search engine. Good abandonment is an abandonment that indicates satisfaction, with one common reason behind it is that a user has found what they are looking for in the summary of search results [6, 10, 14]. Bad abandonment, on the other hand, indicates dissatisfaction and can be due to a broader range of reasons [8]. One reason is due to dissatisfaction with the quality of search results (e.g. a user not finding any document worth clicking). Understanding query abandonment is important to estimate search satisfaction [8] with many previous research investigating and predicting the reasons behind it [7, 8, 13].

Understanding how and when people abandon their query also has important implications for better evaluation measures of search engines. Traditional evaluation measures primarily focus on the quality of the search result list of a single query. The performance of the retrieval method is then aggregated over multiple queries. Unfortunately, this method of evaluation focuses attention on the population of users that have a low probability of abandoning their queries and ignores different aspects of interactions that can influence a user to stop or continue their examination, such as the type of user or query [1]. In our previous work [1, 16], we found that in many cases, it does not matter if the top-most relevant document is ranked 5 or 10, as some users are likely to abandon their query without examining anything below the third search result. Understanding when people will abandon their query allows us to evaluate search results by considering only the results that are likely to be examined by users, thus moving forward to a measure that accurately reflects actual behaviour exhibited by users.

Our previous work on query abandonment [1, 16] provided us with valuable insight on user behaviour. It also introduced further research questions, which we hope to address in this dissertation project.

2 RESEARCH QUESTIONS

Specifically, the dissertation will seek to address the following research questions (RQs):

- **RQ1:** *What is the relationship between the quality of search results and the rate of query abandonment?*
- **RQ2:** *What examination pattern/s do users exhibit when confronted with search results with different quality?*

RQ1 and RQ2 examine how users will react to search results with varying levels of quality (e.g. by placing a single relevant document at different ranks). In particular, RQ1 and RQ2 address how much time is consumed before users decide to abandon their query, how far do users examine the search result page before they decide to abandon and what may influence their decisions to continue or stop their examination.

Our previous work has addressed RQ1 and RQ2 [1, 16] and provided us with insight to understand how good a search result needs to be for users to click on a search result and avoid the negative effects of abandoning queries. The experiments in [1] also provided us with insights that are the motivation behind our next research questions RQ3-5. In particular, while observing users complete their tasks using an eye-tracking device, we noticed that in many cases a user would enter a query and examine a few of the top search results before they make a decision. If a relevant item was not in their sequence of examinations, they decide to abandon their query. After further analysis of the queries users have entered, we learned that there are two types of queries and both can affect users' examination and query abandonment differently. The type of queries in our study was categorized into two types, an under-specified query to the task (i.e an ambiguous query that does not fully address the information need), and a strong query. These observations motivated us to study the next research questions.

- **RQ3:** *To what extent, if any, can users predict the quality of their queries? Do users inherently know whether a query would get them to the right information quicker than another set of queries?*

That is, do users have a perception of the quality of their queries before they enter it to a search engine? Can users predict which query would get them to their information need quicker? Do users have a preference for one query over the other? If so, what motivates their preference?

- **RQ4:** *To what extent, if any, does the quality of non-relevant documents affect users examination and therefor their abandonment rate?*

Having observed users abandon their query after examining a few of the top non-relevant and off-topic search results in a SERP, the question then arises, does the quality of the non-relevant results serve as a stimulus to examining more or less search results? For example, if users were shown non-relevant but encouraging search results (e.g. results that are on the same topic as that of the information need, but does not particularly address or contain the answers of the information need), would that affect their decision to continue or stop examining further result? We believe that addressing this question has direct implications on how we should collect relevance judgment, which is the bases of our next research question.

- **RQ5:** *How do we collect relevance judgment in a way that aligns with our findings of how people abandon search result?*

Relevance judgment is often collected by asking assessors to assess documents in a 3-levels relevance scale [9, Section 2.4.3, page 39], with two items indicating relevance and one for non-relevant. This method of relevance judgment, however, does not take into consideration the variability of user preference and quality of documents, which may be of influence to a user's decision to whether or not abandon a search result.

3 METHODOLOGY

For RQ3-4, we plan to conduct a two-stage user study. The first stage is a controlled within-subjects eye-tracking study that addresses the effect of different types of irrelevant results to examination behaviour (RQ4). After completing the first stage, participants will move forward to the second stage of the study which will involve the interview with questionnaires addressing RQ3 specifically. The eye-tracking study will involve asking users to complete search tasks as they normally do when using search engines. To understand the effect of non-relevant results, we plan to control the quality of search results of their first query of each task, similar to previous research aiming to understand user behaviour with search results [1, 11, 15, 16]. By conducting an interview task, we hope to better understand the differences in how people issue their queries, the reasoning behind the queries they have entered, and whether or not users believe some queries are *better* than other queries. By decomposing the study into two stages, with the first being an eye-tracking study, we also hope to be able to understand differences in query strategies between economic and exhaustive users [4], for which they can be identified using eye-tracking data as demonstrated by Aula et al. [4].

3.1 Eye-tracking Study

Eye-tracking can be a helpful method to determine what users are looking at and in what order. To determine whether different types of non-relevant documents can affect examination, we plan to do an eye-tracking study of participants completing search tasks with a custom search engine. The search tasks in our proposed study will involve asking users to search for information for specific topics, but the search results of their queries will be carefully manipulated to include different types of non-relevant documents positioned at specific ranks. By analyzing how long people spend examining the results and how far in the list they examine, we can determine whether indeed users are affected by the different types of non-relevant documents.

3.2 Interview

In the interview, we will show participants recorded segments of their interactions with the search engine and ask questions addressing RQ3. By interviewing people, we hope to gain insight from a user perspective as to why people formulate their queries a certain way and whether users have a sense of the quality of their queries prior to submitting them.

3.3 Relevance Scale

The relevance scale we propose are based on what we feel may affect users from continuing or stopping their examination of the search result. For example, let us imagine a search task where the user wants to find information about the United Nations' World Heritage Sites in Canada. In Figure 2, we show three different types of documents with what we believe have different levels of relevance. We categorized relevance into three types:

- **Relevant**: a document that directly contains relevant information about the topic of interest. For the example topic above, the Wikipedia page of the list of World Heritage Sites is considered a relevant document.
- **Non-relevant within-topic**: a document that mentions the topic of interest but does not directly address the information need. For example, a document about Canada and the United Nations is within the topic of interest but does not directly provide information about World Heritage Sites in Canada.
- **Non-relevant off-topic**: a document with an erroneous topic, and only shown to the user because it contains matching terms. For example, a document mentioning heritage properties in a Canadian city.

3.4 Search Task and Topics

We plan to ask participants to complete search tasks that involve finding information about a particular topic. For each search task, participants will be provided with a custom search engine with an interface design similar to that of common commercial search

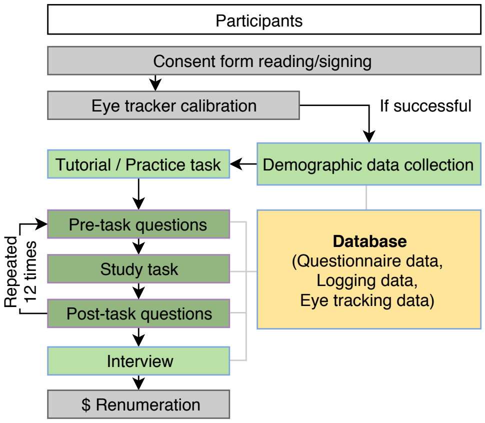

Figure 1: A summary of our proposed study procedure.

engines. We will ask participants to use the search engine to find answer(s) to the information need of the task's topic. Once a participant enters their first query of the task, we will show a carefully manipulated SERP and collect various behaviour data while the participant interacts with the search result. This method of manipulation has been conducted in previous research and is an effective way to understand the effect of different conditions on user behaviour [1, 11, 15, 16].

In terms of topics, we are considering following a similar procedure to Sanderson [12] that utilizes Wikipedia's disambiguation pages[1] to find topics associated with a single term but with different meanings. Many previous research that involves studying user behaviour has also employed the same procedure [2, 3, 5]. Alternatively, we may choose to select a subset of the topics used in Bota et al. [5], which are all considered ambiguous by the authors. Because we plan to manually construct the manipulated SERPs which will be shown to users, ambiguous topics offer easier ways to find a plausible set of documents that match with our definition of relevance. For example, the topic "Doom (video game)" can have documents on the movie version of the Doom video game as non-relevant within-topic documents, and "Doom metal"[2](a type of music genre) as possible non-relevant off-topic documents.

3.5 Study Design and Procedure

The study will involve 6 different conditions representing different search result qualities, as shown in Figure 2. These conditions represent our purposed manipulation of the SERPs which will be shown to users while completing their tasks. Three of the conditions will have the same manipulation in terms of positioning the relevant and non-relevant documents, but will only differ in the type of non-relevant documents shown in the list. By manipulating the position of the relevant and non-relevant results, and by having two conditions with the same structure but with a different type of non-relevant documents, we hope to investigate and compare how far in the ranked list users examine before deciding to abandon their query and whether the type of non-relevant documents affect their examination.

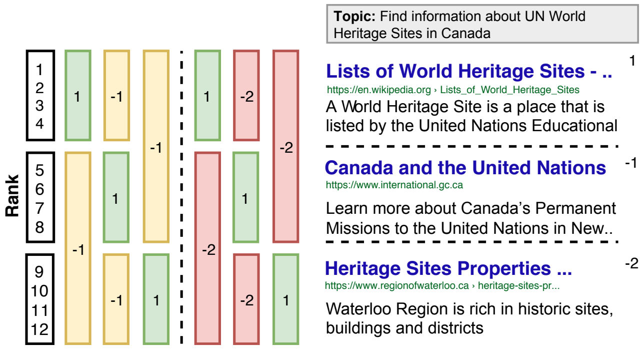
Rank
1
2
3
4
5
6
7
8
9
10
11
12
1
-1
-1
1
-1
-1
1
1
1
-2
-2
-2
1
-2
-2
1
Topic: Find information about UN World Heritage Sites in Canada
Lists of World Heritage Sites - ..
https://en.wikipedia.org › Lists_of_World_Heritage_Sites
A World Heritage Site is a place that is listed by the United Nations Educational
1
Canada and the United Nations
https://www.international.gc.ca
Learn more about Canada's Permanent Missions to the United Nations in New..
-1
Heritage Sites Properties ...
https://www.regionofwaterloo.ca › heritage-sites-pr...
Waterloo Region is rich in historic sites, buildings and districts
-2

Figure 2: Proposed study design. In each task, users will be provided with a topic and will be asked to use our custom search engine to search for information on the topic. User behaviour measurements will be recorded during their interaction with the SERP. Each column represents the condition of the SERP that will be shown to the user in a balanced order based on a Graeco-Latin square. Green boxes (labeled 1) indicate relevant documents to the topic. Yellow and red boxes (labeled -1 and -2) indicate two types of non-relevant documents to the task, denoted as non-relevant within-topic and non-relevant off-topic. On the right, we show an example of search results under our proposed levels of relevance.

The interview will involve showing participants recorded segments of their interaction with the search engine and asking them various questions. In particular, we plan to ask what terms they believe are important to provide as an input to the search engine and how they have decided to formulate their query.

3.5.1 Balanced Design. To generate enough data and to allow the study to be completed under reasonable time, each participant will complete each of the 6 conditions twice. Therefore, each participant will complete 12 tasks in total, thus we plan to create 12 different topics for the study. We will also randomize and balance the design of the study using a 12 × 12 Graeco-Latin square. By randomizing and balancing the study, we reduce any ordering bias that may occur and ensure that each topic will occur under each condition an equal number of times, thus also reducing bias that may occur from the topics.

3.5.2 Procedure. The study will run in a private computer-lab in our university. Participants will use a standard 23.5" desktop monitor equipped with Tobii eye-tracking hardware[3] to complete the study. A summary of the study procedure is shown in Figure 1. First, the participant will read and sign a consent form regarding their participation in the study. Then, we will begin the calibration process of the eye-tracker for the participant. If the calibration was not successful, we plan to give partial remuneration to the participant as an appreciation of their time. The next steps will involve collecting demographic questions and undergoing a tutorial and a practice task phase. The participant will then complete each

[1] https://en.wikipedia.org/wiki/Wikipedia:Links_to_disambiguating_pages
[2] Or others from https://en.wikipedia.org/wiki/Doom
[3] https://www.tobiipro.com/product-listing/tobii-pro-x3-120/

of the 12 tasks in a balanced order according to the Graeco-Latin square. Once the study is done, we will move to the interview where participants will be shown segments of their interactions and be asked questions to clarify and gain more insight into the behaviour of query formulation. Finally, the participant will complete an exit questionnaire about their experience of the study.

4 DATA COLLECTION AND ANALYSIS

While participants interact with the search engine, we plan to record the following measurements.

4.0.1 Submitted queries. All queries submitted to the search engine by the participants during their 12 tasks.

4.0.2 Type of Action. The first action made by the user once they are shown the manipulated SERP. The action can be a document click or a query abandonment.

4.0.3 Time to action. Measured from the moment the manipulated SERP is shown to the user to the moment the user makes their action, whether clicking on a document or abandoning their query (e.g. clicking on the search bar).

4.0.4 Fixation Duration at SERP items. The time users spend looking at each of the search results. This can be measured using the eye-tracking software[4] by creating separate areas of interest (AOI) for each search result.

4.0.5 Eye Fixation Sequence. The sequence of fixation at each of the search results.

4.0.6 Data Analysis. To understand whether the the type of non-relevant document has an effect on examination, we plan to compare and check for statistical significant on each of our collected measures across the proposed conditions. In particular, we hope to answer three main questions: When confronted with our manipulated SERPs with the two type of non-relevant documents, how much time do users spend examining the SERP before making a decision to click on a document or abandon the query? How many search results users examine before making their decision? What is the probability of examining the search results at each rank?

5 CURRENT PROGRESS AND FUTURE PLANS

In our previous work [1, 16], we have explored the rate of abandonment under different types of search result quality (e.g. by modifying the rank of the top-most relevant document in the search result) on simple factoid question-answering tasks. In Abualsaud and Smucker [1], we have used eye-tracking to further understand how far in the search result users' examine before abandoning their query, and the effect of the user and query type on examination behaviour.

In regards to RQ3-4, we have partially developed the set of topics we plan to use for the eye-tracking study. We are still in the process of refining and exploring the questionnaire to ask during the interview. The platform in which participants are going to use will be the same as in our previous studies [1, 16], but modified to adhere to the changes in our study tasks and procedure. Before conducting the study, we plan to recruit a few pilot participants to reduce the number of shortcomings in our platform or study design.

After conducting the study, we hope to have gained enough insight to investigate different methods of collecting relevance judgment (RQ5) that aligns with our findings on how people examine search results.

[4] https://tobiipro.com/product-listing/tobii-pro-studio/

REFERENCES

[1] Mustafa Abualsaud and Mark Smucker. 2019. Patterns of Search Result Examination: Query to First Action. In *Proceedings of the 28th ACM International Conference on Conference on Information and Knowledge Management (CIKM '19)*. ACM, New York, NY, USA.

[2] Jaime Arguello and Robert Capra. 2014. The Effects of Vertical Rank and Border on Aggregated Search Coherence and Search Behavior. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14)*. ACM, New York, NY, USA, 539–548. https://doi.org/10.1145/2661829.2661930

[3] Jaime Arguello and Rob Capra. 2016. The Effects of Aggregated Search Coherence on Search Behavior. *ACM Trans. Inf. Syst.* 35, 1, Article 2 (Sept. 2016), 30 pages. https://doi.org/10.1145/2935747

[4] 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.

[5] Horatiu Bota, Ke Zhou, and Joemon M. Jose. 2016. Playing Your Cards Right: The Effect of Entity Cards on Search Behaviour and Workload. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval (CHIIR '16)*. ACM, New York, NY, USA, 131–140. https://doi.org/10.1145/2854946.2854967

[6] Aleksandr Chuklin and Pavel Serdyukov. 2012. Good Abandonments in Factoid Queries. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12 Companion)*. ACM, New York, NY, USA, 483–484. https://doi.org/10.1145/2187980.2188088

[7] Atish Das Sarma, Sreenivas Gollapudi, and Samuel Ieong. 2008. Bypass Rates: Reducing Query Abandonment Using Negative Inferences. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08)*. ACM, New York, NY, USA, 177–185. https://doi.org/10.1145/1401890.1401916

[8] Abdigani Diriye, Ryen White, Georg Buscher, and Susan Dumais. 2012. Leaving So Soon?: Understanding and Predicting Web Search Abandonment Rationales. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*. ACM, New York, NY, USA, 1025–1034. https://doi.org/10.1145/2396761.2398399

[9] Donna Harman. 2011. *Information Retrieval Evaluation* (1st ed.). Morgan & Claypool Publishers.

[10] Jane Li, Scott Huffman, and Akihito Tokuda. 2009. Good Abandonment in Mobile and PC Internet Search. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*. ACM, New York, NY, USA, 43–50. https://doi.org/10.1145/1571941.1571951

[11] Kevin Ong, Kalervo Järvelin, Mark Sanderson, and Falk Scholer. 2017. Using Information Scent to Understand Mobile and Desktop Web Search Behavior. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. ACM, New York, NY, USA, 295–304. https://doi.org/10.1145/3077136.3080817

[12] Mark Sanderson. 2008. Ambiguous Queries: Test Collections Need More Sense. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*. ACM, New York, NY, USA, 499–506. https://doi.org/10.1145/1390334.1390420

[13] 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.

[14] Kyle Williams, Julia Kiseleva, Aidan C. Crook, Imed Zitouni, Ahmed Hassan Awadallah, and Madian Khabsa. 2016. Detecting Good Abandonment in Mobile Search. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 495–505. https://doi.org/10.1145/2872427.2883074

[15] Wan-Ching Wu, Diane Kelly, and Avneesh Sud. 2014. Using information scent and need for cognition to understand online search behavior. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 557–566.

[16] Haotian Zhang, Mustafa Abualsaud, and Mark D. Smucker. 2018. A Study of Immediate Requery Behavior in Search. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18)*. ACM, New York, NY, USA, 181–190. https://doi.org/10.1145/3176349.3176400