Struggling and Success in Web Search

Odijk, D.; White, R.W.; Awadallah, A.H.; Dumais, S.T.

DOI
10.1145/2806416.2806488

Publication date
2015

Document Version
Author accepted manuscript

Published in
CIKM'15

Struggling and Success in Web Search

Daan Odijk
University of Amsterdam
Amsterdam, The Netherlands
d.odijk@uva.nl

Ryen W. White,
Ahmed Hassan Awadallah,
Susan T. Dumais
Microsoft Research, Redmond, WA, USA
{ryenw,hassanam,sdumais}@microsoft.com

ABSTRACT

Web searchers sometimes struggle to find relevant information. Struggling leads to frustrating and dissatisfying search experiences, even if searchers ultimately meet their search objectives. Better understanding of search tasks where people struggle is important in improving search systems. We address this important issue using a mixed methods study using large-scale logs, crowd-sourced labeling, and predictive modeling. We analyze anonymized search logs from the Microsoft Bing Web search engine to characterize aspects of struggling searches and better explain the relationship between struggling and search success. To broaden our understanding of the struggling process beyond the behavioral signals in log data, we develop and utilize a crowd-sourced labeling methodology. We collect third-party judgments about why searchers appear to struggle and, if appropriate, where in the search task it became clear to the judges that searches would succeed (i.e., the pivotal query). We use our findings to propose ways in which systems can help searchers reduce struggling. Key components of such support are algorithms that accurately predict the nature of future actions and their anticipated impact on search outcomes. Our findings have implications for the design of search systems that help searchers struggle less and succeed more.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Storage and Retrieval: *search process, selection process.*

Keywords

Information retrieval; Struggling

1. INTRODUCTION

When searchers experience difficulty in finding information, their struggle may be apparent in search behaviors such as issuing numerous search queries or visiting many results within a search session [5]. Such long sessions are prevalent and time consuming (e.g., around half of Web search sessions contain multiple queries [37]). Long sessions occur when searchers are exploring or learning a new area, or when they are struggling to find relevant information [31, 43]. Methods have recently been developed to distinguish between struggling and exploring in long sessions using only behavioral signals [20]. However, little attention has been paid to *how* and *why* searchers struggle. This is particularly important since struggling is prevalent in long tasks, e.g., Hassan et al. [20] found that in 60% of long sessions, searchers' actions suggested that they were struggling.

Before proceeding, let us present an example of a struggling task. Figure 1 presents a session in which a searcher is interested in watching live streaming video of the U.S. Open golf tournament.

9:13:11 AM	**Query**	us open
9:13:24 AM	**Query**	us open golf
9:13:36 AM	**Query**	us open golf 2013 live
9:13:59 AM	**Query**	watch us open live streaming
9:14:02 AM	**Click**	http://inquisitr.com/1300340/watch-2014-u-s-open-live-online-final-round-free-streaming-video
9:31:55 AM	**END**	

Figure 1: Example of a *struggling* task from June 2014.

The first two queries yield generic results about U.S. Open sporting events and the specific tournament. The third query might have provided the correct results but it included the previous year rather than the current year. At this stage, the searcher appears to be struggling. The fourth query is the so-called *pivotal query* where the searcher drops the year and adds the terms "watch" and "streaming". This decision to add these terms alters the course of the search task and leads to a seemingly successful outcome (a click on a page that serves the streaming content sought). Understanding transitions between queries in a struggling session (e.g., the addition of "golf" and the wrong year), and transitions between struggling and successful queries (e.g., the addition of terminology pertaining to the desired action and content), can inform the development of strategies and algorithms to help reduce struggling.

Related research has targeted key aspects of the search process such as satisfaction, frustration, and search success, using a variety of experimental methods, including laboratory studies [5, 12], search log analysis [17], in-situ explicit feedback from searchers [13], and crowd-sourced games [2]. Such studies are valuable in understanding these important concepts, and yield insights that can directly improve search systems and their evaluation. However, they do not offer insights into how people who initially struggle, in some cases, ultimately succeed. Such insights can have value to both searchers and search providers in detecting problems and designing systems that mitigate them. We address these shortcomings with the search described in this paper.

CIKM'15, October 19–23, 2015, Melbourne, Australia.

ACM 978-1-4503-3794-6/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2806416.2806488.

We make the following specific contributions:

- We use large-scale search log analysis to characterize aspects of struggling search tasks and to understand how some tasks result in success, while others result in failure.
- We propose and apply a crowd-sourced labeling methodology to better understand the nature of the struggling process (beyond the behavioral signals present in log data), focusing on why searchers struggled and where it became clear that their search task would succeed (i.e., the pivotal query).
- We develop a classifier to predict query reformulation strategies during struggling search tasks. We show that we can accurately classify query reformulations according to an intent-based schema that can help select among different system actions. We also show that we can accurately identify pivotal (turning point) queries within search tasks in which searchers are struggling.
- We propose some application scenarios in which such a classifier, and insights from our characterization of struggling more broadly, could help searchers struggle less.

The remainder of this paper is structured as follows. Section 2 describes related work in areas such as satisfaction, success, and query reformulation. Section 3 characterizes struggling based on our analysis of large-scale search log data. In Section 4, we describe our crowd-sourced annotation experiment and the results of the analysis. Section 5 describes predictive models and associated experiments (namely, predicting query transitions and identifying the pivotal query), and the evaluation results. In Section 6 we discuss our findings, their implications and limitations, and conclude.

2. RELATED WORK

Characterizing the behavior of searchers has been subject of study from different perspectives using a range experimental methods. Of particular interest to our research is the extensive body of work on (i) satisfaction and search success, (ii) searcher frustration and difficulty, and (iii) query reformulation and refinement.

Satisfaction and Success. The concepts of satisfaction and search success are related, but they are not equivalent. Success is a measure of goal completion and searchers can complete their goals even when they are struggling to meet them [17]. Satisfaction is a more general term that not only takes goal completion into consideration, but also effort and more subjective aspects of the search experience such as searcher's prior expectation [23]. Satisfaction has been studied extensively in a number of areas such as psychology [30] and commerce [33]. Within search, satisfaction and success can be framed in terms of search system evaluation, essential in developing better search technologies. Kelly [25] comprehensively summarizes different methods for evaluating search systems with searchers.

On a session level, Huffman and Hochster [22] found a strong correlation between session satisfaction and the relevance of the first three results for the first query, the number of events and whether the information need was navigational. Hassan et al. [17] showed that it is possible to predict session success in a model that is independent of result relevance. Jiang et al. [23] found that it is necessary and possible to predict subtle changes in session satisfaction using graded search satisfaction. Most prior studies regard search tasks or sessions as the basic modeling unit, from which holistic measures (e.g., total dwell time [44]) can be computed. Beyond tasks and sessions, interest has also grown in modeling satisfaction associated with specific searcher actions [26, 39]. These estimates can then be applied to improve rankings for future searchers [19]. Insights into satisfaction and success are used to predict satisfaction for individual queries [13, 18] and for sessions [17, 22, 23].

Frustration and Difficulty. Related to satisfaction are other key aspects of the search process such as task difficulty and searcher frustration. These have been studied using a variety of experimental methods, including log analysis [20], laboratory studies [5, 12], and crowd-sourced games [2]. Feild et al. [12] found that when given difficult information seeking tasks, half of all queries submitted by searchers resulted in self-reported frustration. Ageev et al. [2] provided crowd-workers with tasks of different levels of difficulty and found that more successful searchers issue more queries, view more pages, and browse deeper in the result pages. When searchers experience difficulty in finding information, their struggle may be apparent in search behaviors such as issuing numerous search queries, more diverse queries or visiting many results within a search session [5]. However, rather than struggling, these longer sessions can be indicative of searchers exploring and learning [10, 43]. Hassan et al. [20] have recently developed methods to distinguish between struggling and exploring in long sessions using only behavioral signals. They found that searchers struggled in 60% of long sessions. Scaria et al. [36] examined differences between successful and abandoned navigational paths using data from a Wikipedia-based human computation game. They compared successful and abandoned navigation paths, to understand the types of behavior that suggest people will abandon their navigation task. They also constructed predictive models to determine whether people will complete their task successfully and whether the next click will be a back click (suggesting a lack of progress on the current path). The terminal click has also been used in other studies of search to better understand searchers' information goals [9] or point people to resources that may be useful to other searchers [40]. In this paper, we focus on struggling tasks (that are thus likely to be unsatisfactory) to understand how some of them end up successful while others end up unsuccessful. We target traditional Web search in this study given its prevalence. Recently, others have studied struggling and success in the context of engagement with intelligent assistants [23].

These studies and those on searcher satisfaction are valuable in understanding these important concepts, and yield insights and signals that can directly improve search systems and their evaluation [19]. They provide important clues on what searchers might do next, such as switching to a different search engine [15, 41] or turning to a community question answering service [29]. Ideally, a search engine would interpret these signals of struggling and frustration to provide personalized hints to help the searcher succeed. These hints can be learned from more successful and advanced users [2, 42] or provide examples that are may work generically for some search intents, such as leading searchers to longer queries [1]. Moraveji et al. [32] showed that the presentation of optimal search tips, where they are presented for a task where they are known to have benefit, can have a lasting impact on searcher efficiency. Savenkov and Agichtein [35] showed that providing a searcher with task-specific hints improves both success and satisfaction. Conversely, generic hints decrease both success and satisfaction, indicating that it is paramount to understand what a searcher is struggling with before providing hints.

Query Reformulation and Refinement. More detailed insight on searcher behavior can be found by analyzing query reformulations. Query reformulation is the act of modifying the previous query in a session (adding, removing, or replacing search terms) with the objective of obtaining a new set of results [18]. For this, a number of related taxonomies have been proposed [3, 14, 21, 27, 38]. Huang and Efthimiadis [21] surveyed query reformulation taxonomies and provided a mapping between these and their own approach. While they provide interesting insight into searchers trying to articulate

their information needs, these approaches all focus on superficial lexical aspects of reformulation. Anick [3] examined usage and effectiveness of terminological feedback in the AltaVista search engine. No difference in session success was found between those using the feedback and those not using it, but those using it did continue to employ it effectively on an ongoing basis. A number of recent studies have shown that search tasks provide rich context for performing log-based query suggestion [11, 24, 28], underscoring the importance of studying query reformulation in search tasks.

Contributions over previous work. No previous study examines how searchers struggle and what makes them ultimately succeed. We employ large-scale log analysis and a crowd-sourced labeling methodology to provide new insight into the nature of struggling and what contributes to their success. Based on this, we propose a new taxonomy for intent-based query reformulation that goes beyond the superficial lexical analysis commonly applied in the analysis of query transitions (see [21]). Building on our characterization, we propose predictive models of key aspects of the struggling process, such as the nature of observed query reformulations and the prediction of the pivotal query. Based on insights gleaned from our data analysis, we also provide some examples of the types of support that search systems could offer to help reduce struggling.

3. CHARACTERIZING STRUGGLING

We apply large-scale search behavioral analysis to characterize aspects of struggling search tasks and to understand how some of these searches end up successful while others end up unsuccessful.

3.1 Definitions

We focus on a particular type of search task that exhibits search behavior suggestive of struggling. We assume a broad view of struggling behavior and apply the following definitions:

Struggling describes a situation whereby a searcher experiences difficulty in finding the information that they seek. Note that in this definition they may or may not eventually locate the target of their search [20].

Sessions are a sequence of search interactions demarcated based on a 30-minute user inactivity timeout [9, 42].

Tasks are defined as topically-coherent sub-sessions, i.e., sequences of search activity within sessions that share a common subject area [20]. We follow the approach of [20] and assume that two queries belong to the same task if they are less than ten minutes apart and the queries match one of the following conditions: (i) share at least one non-stop word term, or (ii) share at least one top ten search result or domain name (where popular domains such as wikipedia.org are excluded).

Struggling tasks describe topically coherent sub-sessions in which searchers cannot immediately find sought information.

Quick-back clicks describe result clicks with a dwell time of less than ten seconds [26].

Since we cannot infer that searchers experience difficulty from a single query, we consider only longer struggling tasks in our analysis. Our aim is to obtain a broad understanding of struggling search behavior in natural settings. To do this, we study search tasks where struggling is very apparent. Intuitively, when a searcher cannot locate the information they are seeking, they are much less likely to click search results and examine landing pages. We focus on tasks where a searcher does not examine any of the search results for the first two queries in detail. This includes queries that do not receive any clicks as well as queries with clicks that result in only very short dwell time on the landing page. We use a dwell time of less than 10 seconds to identify these quick-back clicks.

3.2 Mining Struggling Tasks

To better understand struggling search behavior in a natural setting, we analyze millions of search sessions from the Microsoft Bing Web search engine. We select these sessions from the interaction log of Bing for the first seven days of June 2014. All logged interaction data (i.e., queries and clicked search results) are grouped based on a unique user identifier and segmented into sessions. We mine struggling tasks from all sessions using the following steps:

1. **Filter sessions:** We included only search sessions originating from the United States English language locale (en-US), and excluded internal traffic and secure traffic (https). Furthermore, we only considered sessions that started with a typed query (and not with a click on a query suggestion).
2. **Segment sessions into tasks:** Using a time-based segmentation can lead to combining multiple unrelated tasks into a single session. We therefore further refine sessions into topically coherent sub-sessions that cover a single task.
3. **Filter struggling tasks:** From these tasks, we select those with at least three queries where the first two lead to either no clicks or only quick-back clicks.
4. **Partition based on final click:** Lastly, we partition the struggling tasks based on searcher interaction for the last query in the task. Although we cannot directly infer whether the searcher successfully fulfilled their information need, search activity for terminal queries has been shown to be a reasonable proxy for success [13, 17]. We validate this predictor using crowd-sourced annotations in Section 4. More specifically, we partition the tasks into three sets:
   (a) **Unsuccessful:** If a searcher does not click on any result for the final query or when their only clicks are quick-back clicks (less than 10 seconds), we assume the searcher was unsuccessful and abandoned their search task without satisfying their need.
   (b) **Successful:** If a searcher clicks on a search results and we observe no interaction for at least 30 seconds, we assume the searcher was successful. Note that this includes tasks in which the searcher does not return at all to the result page before the session times out.
   (c) **Other:** All other tasks have clicks where searchers examine landing pages between 10 and 30 seconds. Based on previous research [13], we consider these task outcomes to be ambiguous and exclude them.

Following these steps, we obtain two sets of tasks that differ only based on the interaction with the terminal query. The combined dataset contains nearly 7.5 million struggling tasks. We now describe characteristics of these tasks to provide insight into struggling behavior, seeking to understand how some searches result in success, while others are unsuccessful. We begin with struggling tasks, and then consider queries and query reformulations.

3.3 Task Characteristics

Of the struggling tasks in our set, approximately 40% are successful per our definition. Around half of the tasks comprised three queries and successful tasks were slightly shorter than their unsuccessful counterparts. Focusing on the relationship between task duration and task outcome, Figure 2 shows the percentage of struggling tasks that were continued after a specified number of queries. We observe from the figure that the percentage of tasks that are continued increases with the number of queries already issued (i.e., the likelihood of a re-query increases with each successive query). Unsuccessful tasks are continued more frequently than successful tasks. There are many

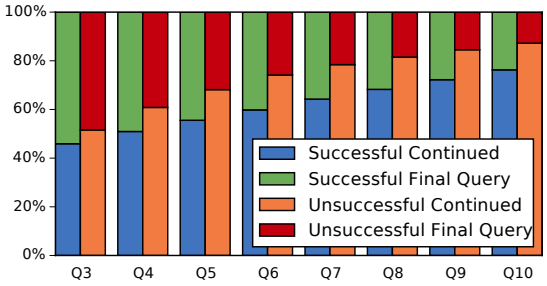


Figure 2: Percentage of successful and unsuccessful struggling tasks continued (onto another query) or completed (task terminates) after the third query (Q3) until the tenth query (Q10).

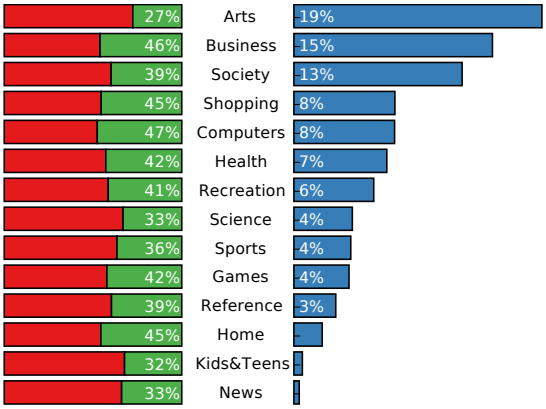


Figure 3: Prevalence of top-level ODP categories in all analyzed tasks (right) and proportion of successful and unsuccessful struggling tasks for that category (left).

important factors, such as searcher tenacity and sunk cost, that may explain more of a reluctance to abandon unsuccessful tasks.

For some topics, searchers experience more difficulty in finding what they are looking for than for others. For example, Hassan et al. [20] reported that exploratory behavior is more than twice as likely as struggling behavior when shopping for clothing compared to downloading software. To obtain greater insight on the relationship between search topics on struggling behavior, we analyze the top 10 search results returned to searchers. We classify each document in the search results using the top-level categories of the Open Directory Project (ODP). For this we use an automatic content-based classifier [6]. We assign the most frequently-occurring topic across all queries in a task as the topic of that task. Figure 3 shows the prevalence of topics in struggling tasks. The proportion of successful tasks ranges from 47% in Computers to 27% in Arts, which is also the most prevalent topic. Next, we analyze the changes within a task by considering characteristics at different stages.

3.4 Query Characteristics

Since we have tasks of different length, we can only directly compare tasks of the same length or align queries in tasks of different length. To analyze the development over the broadest set of tasks, we consider the first and last query in the search task, and collapse all intermediate queries.

Query Length: We observe that the first query in both successful and unsuccessful tasks is typically short (3.35 and 3.23 terms respectively on average). Intermediate queries are typically longer, averaging 4.29 and 4.04 terms respectively. The final queries are also longer, averaging 4.29 and 3.93 terms respectively. Increased success associated with longer queries has motivated methods to encourage searchers to issue queries with more terms [1].

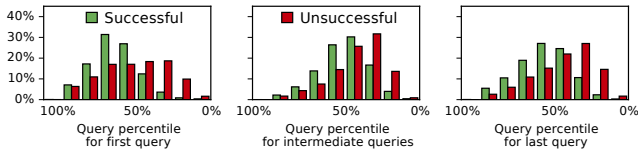


Figure 4: Query percentile. Larger percentile = more frequent.

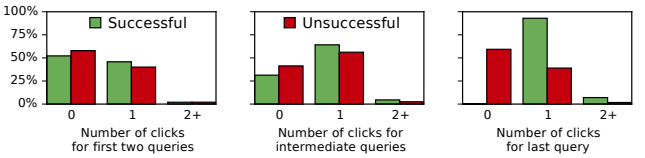


Figure 5: Number of result clicks at different queries.

Query Frequency: For all queries in our dataset we compute the frequency of that query during the previous month. Figure 4 shows the percentile of a query, based on how often the query was issued in the previous month. Queries tend to become less common as the task progresses (distributions shift to the right in each plot). The first query is different than the ones that follow. For successful outcomes, the first query is more common than the successive queries, and more common than the first query for unsuccessful outcomes. This could suggest two different causes for struggling on the first query: (i) if it is common, it may be general and ambiguous, and (ii) if it is uncommon, it might be overly specified.

3.5 Interaction Characteristics

Next, we turn to the interaction within a task. We follow a similar query grouping approach as in Section 3.4, using the position in the task. Since we select struggling tasks based on limited interaction on the first two queries, we separate those from the others. We now have three groups: first and second query, last query and all other intermediate queries. Note that we have no intermediate queries for tasks of only three queries (about half of the tasks). Figure 5 shows the change in the number of clicks in successful and unsuccessful tasks. We observe that for the first and second query, just over 40% of the tasks have a quick-back click. By our definition of struggling task, all clicks for the first two queries were quick-back clicks.

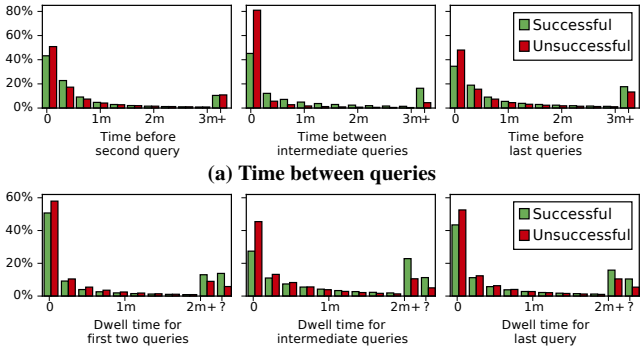
As described in Section 3, we used the clicks on the search results of the final query to partition our dataset into successful and unsuccessful tasks. We observe in Figure 5 that indeed the final query for all successful tasks has at least one click. The characteristics for the final query in the unsuccessful tasks is very similar to the first two queries (that are selected in the same way). When comparing the successful and unsuccessful tasks up to the final query, we observe that queries without clicks are more common in unsuccessful tasks.

Figure 6a shows the time between queries and Figure 6b shows the dwell time (i.e., the time spent on a landing page after a click). Considering the successful tasks, we see the time between queries increases as the task progresses, mostly due to an increase in dwell time. Interestingly, the pattern for unsuccessful struggling tasks is different. We observe less time between queries mid-task perhaps due to fewer clicks (Figure 5) and more quick-back clicks.

3.6 Query Reformulation

To better understand the change in queries within a task, we analyze how the text of a query is refined over time. We classify each query reformulation into one of six types, described in Table 1. We use the algorithm presented in [16] which builds an automatic classifier for a subset of the query reformulation types presented in [21].

Figure 7 shows the distribution over query reformulation types. The most common type of query reformulation is specialization, i.e., adding a term. This is most likely to occur directly after the first query, when it accounts for almost half of all query reformulations



(a) Time between queries

(b) Average dwell time (i.e., the time spent on a page after a click)

Figure 6: Time between queries and spent on a clicked page.

Table 1: Lexical-based query reformulation types.

Type	Description
New	Shares no terms with previous queries.
Back	Exact repeat of a previous query in the task.
Spelling	Changed the spelling, e.g. fixing a typo.
Substitution	Replaced a single term with another term.
Generalization	Removed a term from the query.
Specialization	Added a term to the query.

in successful tasks and a bit less in unsuccessful tasks. Substitutions and generalizations (removing a term) are the next most common reformulations. These query reformulation types are substantially less likely for the first query reformulation and more likely in the middle. Around 10% of the query reformulations are spelling corrections and an equal percentage entirely new queries. Spelling reformulations are most likely directly after the first query. New queries are most likely as second query and, surprisingly, as the last query. Others have observed similar patterns, e.g., Aula et al. [5] showed that searchers try many distinct queries toward the end of difficult search tasks. Lastly, revisiting a prior query is rare and is more likely in the middle of the task than at the final query.

One could argue that of these query reformulation types a specialization is most informative, since it defines an information need in greater detail. This is the most common type and substantially more common in successful tasks (39% of reformulations) than in unsuccessful tasks (30%). For these unsuccessful tasks, almost all other query reformulation types are more common. With more substitutions, spelling corrections, completely new queries and returning to previous queries, it appears that searchers in the unsuccessful tasks experience more difficulty selecting the correct query vocabulary.

Inspired by Lau and Horvitz [27], we examined the temporal dynamics of the query reformulations in addition to their nature. Figure 8 shows the likelihood of observing a particular reformulation type given the time that has elapsed since the previous query. If a new query is issued within seconds after the previous, it is most likely to be a substitution or a spelling change. In fact, a spelling change is unlikely to occur after more than about fifteen seconds.

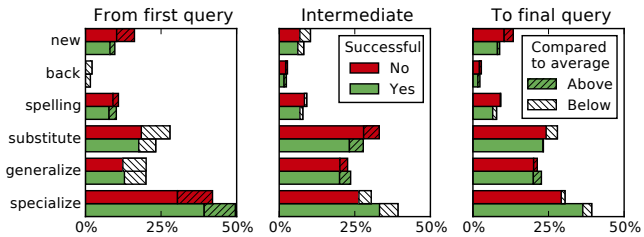


Figure 7: Distribution of query reformulation types at different stages of the search task for successful/unsuccessful tasks.

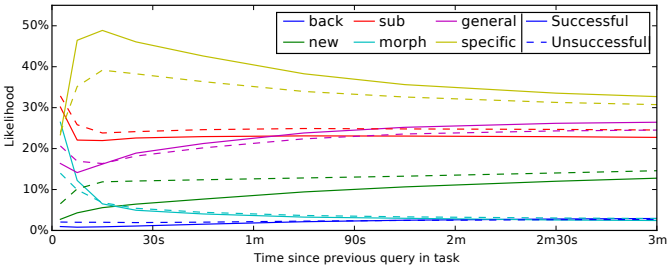


Figure 8: Likelihood of observing a query reformulation type in successful/unsuccessful tasks, given time since last query.

After a few seconds, the most likely type of reformulation is a specialization, with a peak at around fifteen seconds, where nearly half of the queries are expected to be reformulations. After this, some of the likelihood mass is taken over by generalizations and completely new queries. The likelihood per query reformulation type for successful vs. unsuccessful tasks appears similar, except for an increased likelihood of new queries for unsuccessful tasks, accompanied by a decreased likelihood of specialization. Anchoring on previous queries can harm retrieval performance [7]. The increase in new queries may represent an attempt to reset the query stream for the current task. Temporal dynamics such as these are interesting and may prove to be useful signals for the reformulation strategy prediction task described later in the paper.

3.7 Summary

In the analysis in this section, we have shown there are significant differences in how struggling searchers behave given different outcomes. These differences encompass many aspects of the search process, including queries, query reformulations, result click behavior, landing page dwell time, and the nature of the search topic. Given these intriguing differences, we employ a crowd-sourcing methodology to better understand struggling search tasks and the connection between struggling and task outcomes.

4. CROWD-SOURCED ANNOTATIONS

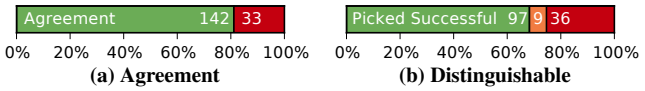
We performed a series of detailed crowd-sourced annotations to obtain a better understanding of what struggling searchers experience. This is also important in validating the assumptions made in the log-based estimation of search success from the previous section. Informed by an initial exploratory pilot with open-ended questions on a task level, we annotate tasks on a per-query basis.

To obtain a representative and interesting sample of tasks (and importantly, to also control for task effects), we group the tasks described in Section 3 based on the first query in the task. Recall that these tasks are all struggling tasks, either successful or unsuccessful. For each initial query, we count the number of successful and unsuccessful tasks. We then filter these queries to have an approximately equal number of successful and unsuccessful tasks (between 45% and 55%).

Upon inspection of the selected tasks, we noticed a small set were unsuitable for annotation. To this end, we exclude initial queries that were deemed too ambiguous, are navigational in nature, or show many off-topic follow up queries. We randomly sample from these initial queries and verify whether they meet our criteria until we manually selected 35 initial queries for deeper analysis in an exploratory pilot study.

4.1 Exploratory Pilot

For each of the 35 initial queries, we generate five task pairs, by randomly sampling a successful task and an unsuccessful task that both start with that initial query. These 175 task pairs are annotated



(a) Agreement **(b) Distinguishable**

Figure 9: (a) Agreement and (b) Distribution of task outcomes of the distinguishable (majority agreement) tasks.

for struggling and success by three judges. We recruited a total of 88 judges from Clickworker.com, a crowd-sourcing service providing access to human annotators under contract to Microsoft and others. We created an interactive judgment interface that displays the search activity (queries and clicks) associated with two paired tasks side-by-side, in random order. To validate whether the activity on the terminal query is a reasonable proxy for success, we asked judges in which of the two tasks they believe the searcher was more successful using three options (one of the two sessions was more successful or they where indistinguishable). We have agreement by majority vote (i.e., at least two of the three judges agreed) for 81% of the pairs (Figure 9a). If there is a majority, the successful task is picked for 68.4% of the pairs and the unsuccessful struggling task for 25.4% of the pairs, while in 6.3% of the cases it is agreed that the tasks are indistinguishable (Figure 9b). These findings suggest that the judgment task is tractable and that tasks we labeled as successful automatically are indeed more often deemed successful by our judges.

We first considered on the search activity connected to the task in which the judge believed the searcher was more successful in more detail. We informed the judge that we believed that the searcher started off struggling with their search, and asked them to look in detail at the task and answer a number of open-ended questions, starting with:

- Could you describe how the user in this session eventually succeeded in becoming successful? *Describe what made the user pull through. Note that this could be multiple things, for example: both fixing a typo and adding a specific location.*

We then considered the task in which the searcher was less successful and asked the judge to look in detail at that task and answer:

- Why was the user in this session struggling to find the information they were looking for? *Describe what you think made the user have trouble locating the information they were looking for. Note that this could be multiple things, for example: looking for a hard to find piece of information and not knowing the right words to describe what they are looking for.*
- What might you do differently if you were the user in this session? *Describe how you would have proceeded to locate the information they were looking for. Note that this could be multiple things, for example: using different terms (please specify) and eventually ask a friend for help.*

The answers to the questions provided diverse explanations for why searchers were struggling and what (could have) made them pull through. The answer typically described specific changes from one query to the other. Some of the explanations were topic-specific (e.g., suggesting a particular synonym for a query term) and some were quite generic (e.g., a suggestion to expand an acronym). We observed from these answers that the main reasons and remedies for struggling are not very well captured by the query reformulation types that are typically used, including the ones we described in Table 1. We adapted our main annotation efforts accordingly.

4.2 Annotations

The exploratory pilot suggested that third-party judges agree on labeling the success or failure of search tasks (especially in the positive case, where it may be more clear that searchers have met their information goals), and provided diverse clues on how and why searchers are struggling in a task. We now dive deeper into specific queries and what searchers did to remedy struggling.

Table 2: Intent-based taxonomy presented to judges for each query transition, multiple response options could be selected.

Added, removed or substituted	☐ an **action** (e.g., download, contact) ☐ an **attribute** (e.g., printable, free,), specifically (if applicable): ☐ a **location** (or destination or route) ☐ a **time** (e.g., today, 2014, recent) ☐ **demographics** (e.g., male, toddlers)
Specified	☐ a **particular instance** (e.g. added a brand name or version number)
Rephrased	☐ Corrected a spelling error or **typo** ☐ Used a **synonym** or related term
Switched	☐ to a **related** task (changed main focus) ☐ to a **new** task

In our main annotation efforts, we consider how a searcher reformulates queries within a task. Based on the open question answers of the exploratory annotation pilot, we propose a new taxonomy of query reformulation strategies (depicted in Table 2). In contrast to the lexical reformulation taxonomy in Table 1, this new taxonomy captures the intent of a query reformulation. Rather than simply observing that a term was substituted, we want to know if this is a related term to the one replaced or if it is a specification of an instance (e.g., refining [microsoft windows] to [windows 8]).

We hypothesize that there is a point during a struggling search task where searchers switch from struggling with little progress to making progress toward task completion (i.e., the so-called *pivotal* query). This could be associated with many factors, including the receipt of new information from an external source such as a search result or snippet. Understanding the pivotal query can provide valuable insight into what enabled the searcher pull through and is an important starting point when finding means to support searchers who are struggling. We ask judges to select the point in the task where they believe the searcher switched from struggling to being successful in finding what they were looking for. Judges could select either a query or a clicked URL in a task presented as shown in Figure 1. Judges could reissue the query to Bing and inspect the clicked pages that the original (logged) searcher selected. Furthermore, the crowd-workers were asked to judge how successful they think the searcher was in the task on a four-point scale: *not at all, somewhat, mostly, completely*. For annotation, we selected two separate, but closely-related sets of tasks. The first is based on the dataset from the exploratory pilot. We exclude six initial queries from the set that showed low agreement on picking the most successful task. For each of the 29 remaining initial queries, we sampled ten new successful tasks that started with that query. In a second set, we sampled 369 successful tasks with any initial query (as a control). This results in a total of 659 tasks.

4.3 Annotation Results

Each task was judged by three human annotators via an interactive interface. This results in a total of 1,977 annotations for 659 tasks. We removed the annotations of three of the 111 judges (80 annotations in total), because their responses were unrealistically quick. On average, judges spent 37 seconds to familiarize with a task and to judge the success and pick the pivotal query. Subsequently, they spent twelve seconds on average to judge each query transitions (at least two per task, depending on the number of queries).

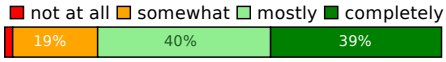


Figure 10: Judgments for the success of tasks.

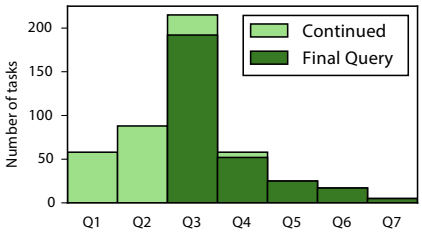


Figure 11: Pivotal query within a task, i.e. where the searcher appeared to switch from struggling to making progress.

Judges show 67% majority agreement on the task of judging the success of a searcher on a four-point scale. Krippendorff's α measures 0.33, signaling fair agreement [4] between three judges. For judging what query is the pivotal query, we observe a 71% majority agreement for choosing one out of three or more queries ($\alpha = 0.44$, signaling moderate agreement between three judges). This demonstrate that third-party labeling is feasible but also challenging given the subjectivity of the labeling task. We deem this satisfactory for our purposes and consider only annotations with majority agreement in the remainder of this section.

Figure 10 illustrates the distribution of success judgments. We selected these tasks assuming that the searcher was struggling, but somewhat successful. We observe that in these tasks searchers are deemed to be at least mostly successful (79%), with only eight tasks (1.8%) not successful at all. This suggests that almost all studied searchers make at least some progress toward completing the task, even though they struggled along the way. We asked judges to consider the tasks carefully and select the pivotal query where searchers appeared to switch from struggling to succeeding. Figure 11 shows the distribution of positions of this pivotal query across the search task. In 62% of tasks, the pivotal query represents the final query in the task.

Query transitions. Judges were asked to examine each sequential pair of queries, while added and removed terms were highlighted. We asked judges to characterize the transition between these two queries by selecting one or more applicable options from the taxonomy. Figure 12 shows the distribution of query reformulation types. The most common are adding, substituting or removing an attribute and specifying an instance. The most common subtype of attribute modified is location with 17.8%, whereas time is only 5% of the attributes and demographic information (e.g., gender, age) occurs in only nine transitions (1.2%). None of the more fine-grained attribute subtypes show qualitatively different characteristics, so we will discuss only the attribute category, without the subtypes. Turning our attention to the different stages within tasks, we observe that adding attributes or actions and specifying an instance is relatively common from the first to the second query. For the transition towards the final query in a task, substituting or removing an attribute, rephrasing with a synonym and switching task are relatively common. This suggests more emphasis on broadly specifying information needs at the outset of tasks and more emphasis on refining it by adding specific details towards the end.

Transitions and task outcomes. Figure 13 shows the relationship between query transitions and success. It is worth noting that switching to a new task occurs substantially more often in less successful tasks. Specifying an instance also occurs relatively often in the less successful tasks. Addition, substitution, and deletion actions or attributions typically occurs in more successful tasks.

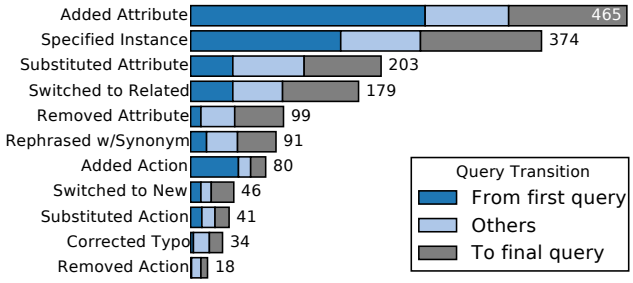


Figure 12: Distribution of query reformulation types.

Figure 13 also shows how often a query transition is deemed pivotal if it occurs. Interestingly, both switching tasks and specifying an instance are more common in less successful tasks, but are relatively frequently considered to be pivotal. Substituting an action is most often seen as pivotal, whereas substituting an attribute and correcting a typo are least frequently pivotal. The differences in the prevalence of pivotal queries as a function of the reformulation type suggests that some actions may be more effective than others and that accurately predicting the next action presents the opportunity for early corrective intervention by search systems. We will discuss this and similar implications toward the end of the paper.

4.4 Summary

Through a crowd-sourcing methodology we have shown that there are substantial differences in how searchers refine their queries in different stages in a struggling task. These differences have strong connections with task outcomes, and there are particular pivotal queries that play an important role in task completion.

5. PREDICT REFORMULATION STRATEGY

Given the differences in query reformulation strategies and to help operationalize successful reformulations in practice, we develop classifiers to (i) predict inter-query transitions during struggling searches according to our intent-based schema (Table 2), and (ii) identify pivotal queries within search tasks. This facilitates the development of anticipatory support to help searchers complete tasks. For example, if we predict a searcher wants to add an action to the query, we can provide query suggestions and auto-completions with actions. Our classifiers can also be applied retrospectively, e.g., to identify frequent transitions between queries and pivotal queries that form useful query suggestions or alterations.

Features. We include five sets of features, described in detail in Table 3. Some of the features relate to the characterizations that have been described in previous sections of the paper.

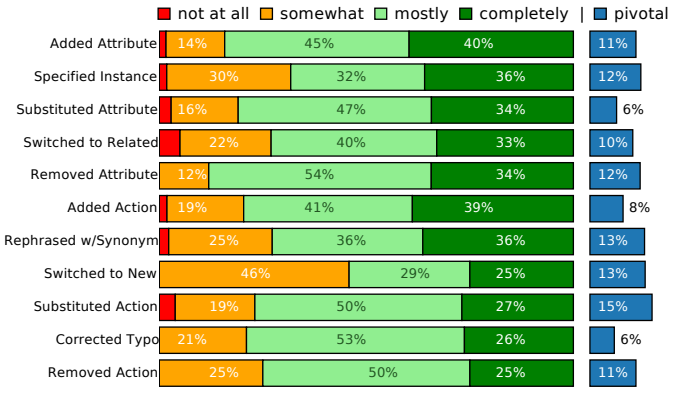


Figure 13: Task-level success on a four-point scale per reformulation type and % of reformulations considered pivotal.

Table 3: Features to predict reformulation strategy. The features marked with * are included for five similarity functions.

Name	Description
Query Features	
NumQueries	Number of queries in task
QueryCharLength	Query length in number of characters
QueryTermLength	Query length in number of terms
Interaction Features	
TimeElapsed	Time elapsed since query issued
ClickCount	Number of clicks observed since query
ClickDwelltime	Dwell time on clicked pages
QueryDwellTime	Dwell time on result page
Query Transition Features	
NewCharLength	New query length in number of characters
NewTermLength	New query length in number of terms
Levenshtein	Levenshtein edit distance in characters
normLevenshtein	Levenshtein edit distance as proportion
commonCharLeft	Number of characters in common from left
commonCharRight	Number of characters in common from right
diffPOS:<type>	Difference in the number of terms that are identified in WordNet as belonging to a specific part of speech type: noun, verb, adjective, adverb
Query Similarity Features	
ExactMatch*	Number of terms that match exactly
AddTerms*	Number of terms not in previous query
DelTerms*	Number of terms not in new query
SubsTerms*	Number of terms substituted
QuerySim*	Proportion of terms exactly match
commonTermLeft*	Number of terms in common from left
commonTermRight*	Number of terms in common from right
Query Reformulation Features	
LexicalType	Lexical query reformulation type (Table 1): new, back, morph, sub, general, specific
RuleBasedType	Lexical reformulation type using rule-based classifier of Huang and Efthimiadis [21]

Query features are used to represent the query before a reformulation. Interaction features describe how a searcher interacted with the search engine results page (SERP). We saw in Figure 8 that the type of reformulation is dependent on the time elapsed since the previous query. After observing a new query, we can compute three new sets of features. First, the query transition features describe characteristics of the new query and low-level lexical measures of how the query has changed. Query similarity features describe the terms in a query have changed. For these features, similarity is measured using five different similarity functions, described in Table 4. Terms can match exactly, approximately, on their root form or semantically. Lastly, we include features that analyze the lexical reformulation of queries. For this, we used the procedure as described in Section 3.6 and a recent rule-based approach [21].

Experimental Setup. We frame this as a multi-class classification problem; one for each of the 11 query reformulation types. We use the 659 labeled tasks with a total of 1802 query transitions (Section 4). We perform ten-fold cross-validation over the tasks and use a RandomForest classifier, since it is robust, efficient and easily parallelizable. We experimented with other classifiers (including logistic regression and SVM), and none yielded better performance. We therefore only report the results of the RandomForest classifier.

Table 4: Similarity matching functions used to compare terms for query similarity features.

Name	Description
Exact	All characters match exactly
Approximate	Levenshtein edit distance is less than two
Lemma	Terms match on their lexical root or lemma form
Semantic	WordNet Wu and Palmer measure greater than 0.5 (measures relatedness using the depth of two synsets and the least common subsumer in Wordnet)
Any	Any of the above functions match

Table 5: Groups of features available at different stages.

Feature group	First Query	First+Interaction	Second Query
Query	✓	✓	✓
Interaction		✓	✓
Transition including Similarity & Reformulation			✓

We evaluate our approach at different stages between two queries:

First query: We only observed the first query and try to predict the next reformulation strategy. At this stage, search systems can tailor the query suggestions on the SERP.

First+Interaction: We observed the first query and interactions (clicks, dwell time). The searcher is about to type in a new query. At this stage, systems can tailor auto-completions for the next query.

Second query: We observed both the first and second query and infer the reformulation strategy that the searcher applied. At this stage, search systems could re-rank results (or blend the results with those from other queries), and suggestions for the next query.

Concretely, the different stages mean that different groups of features become available (see Table 5 for the feature groups available at each stage). Finally, after the first query transition we add history features that represent the reformulation type and previous transition strategy. We report accuracy, area under the ROC curve (AUC) and F1 for our classifiers. F1 is computed as the harmonic mean of precision and recall per class. As a baseline we use the marginal (i.e., always predict the dominant class).

5.1 Prediction Results

Table 6 shows the results of our prediction experiments. The results are grouped by the transition and stage within a struggling search task. A task starts with observing the first query (Q1) and our prediction experiments end with observing the third query (Q3). We observe from the baseline results in Table 6 (lines 1 and 5) that this multi-class prediction task is a difficult problem. Apart from the first transition (where adding attributes is overrepresented, see Figure 12), the baseline approach of always predicting the dominant class is only about 25% correct, and obviously not very informative to support a struggling searcher.

For the first query reformulation (Q1 to Q2, lines 1–4 in Table 6), our classifiers already improve the reformulation strategy prediction before the second query. While just observing the first query does not provide significant improvements on all metrics (line 2), observing clicks and dwell times increases the F1 score significantly from 20% to 34% (line 3). If a struggling searcher issues a second query, we can infer the applied strategy with a 43% accuracy and significantly better on all metrics (line 4). Turning to the second query reformulation (Q2 to Q3, lines 5–8 in Table 6), the baseline performance (line 5) is substantially lower as the dominant class is less prevalent. Directly after observing the second query and using the task history, we can predict the reformulation strategy with 46% accuracy (line 6). Observing the interactions with the second

Table 6: Results for predicting query reformulation strategy. Significant differences, tested using a two-tailed Fisher randomization test against row 1 for 2–4 and row 5 for 6–8, are indicated with △ ($p < 0.05$) and ▲ ($p < 0.01$).

Transition	Stage	Accuracy	F1	AUC
1. Q1 to Q2	Baseline	0.3736	0.2032	0.5000
2. Q1 to Q2	Q1	0.3679	0.3046▲	0.5322△
3. Q1 to Q2	Q1+Interaction	0.3698	0.3371▲	0.5589▲
4. Q1 to Q2	Q2	0.4302△	0.3916▲	0.6077▲
5. Q2 to Q3	Baseline	0.2095	0.0971	0.4908
6. Q2 to Q3	Q2	0.4644▲	0.4595▲	0.6799▲
7. Q2 to Q3	Q2+Interaction	0.4862▲	0.4826▲	0.6896▲
8. Q2 to Q3	Q3	0.5474▲	0.5321▲	0.7306▲

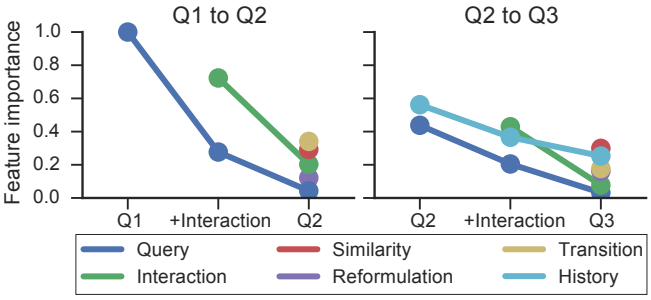


Figure 14: Feature group importance for Q1 to Q2 transition (left) and Q2 to Q3 transition (right).

query does not substantially improve prediction performance (line 7). If we observe the third query, the reformulation strategy can be inferred with 55% accuracy (line 8). All classifiers for the second query transition (lines 6–8) perform substantially better than those for the first transition (lines 2–4). This suggests that knowing what strategy a searcher has used previously helps in predicting the next reformulation strategy that they will use.

Feature Analysis. We measure the feature importance as gini importance [8] averaged over all trees of the ensemble. Figure 14 visualizes the total importance of the six groups of features. For predicting the first transition (Q1 to Q2) directly after the first query, only the query features are available. The interaction features that become available for the first query are substantially more important. If we observe the second query, the similarity and transition features are most important. The query features no longer contribute much to the classifier, as do the reformulation features. For the transition from the second to the third query (Q2 to Q3), the pattern is similar, but the history features contribute more.

Identifying the Pivotal Query. We were also interested in how accurately we can identify the pivotal query. We use a similar set-up as above, using subsets of the described features to identify the pivotal query in the 659 labeled tasks. Table 7 shows the results of this approach. The baseline always predicts the terminal query as the pivotal query (Figure 11). Using only the interaction and query features from Table 3 we significantly outperform this baseline in terms of F1 and AUC. Adding more features does not increase performance on any metric. Although this is promising, our results suggest that identifying the pivotal query is difficult.

6. DISCUSSION AND CONCLUSIONS

Search engines aim to provide their users with the information that they seek with minimal effort. If a searcher is struggling to locate sought information, this can lead to inefficiencies and frustration. Better understanding these struggling sessions is important for designing search systems that help people find information more

Table 7: Results for retrospectively identifying the pivotal query. Significant differences, tested using a two-tailed Fisher randomization test against row 1 are indicated with △ ($p < 0.05$) and ▲ ($p < 0.01$).

	Accuracy	F1	AUC
1. Baseline	0.6245	0.5091	0.7038
2. Query + Interaction	0.6352	0.5817▲	0.7296△

easily. Through log analysis on millions of search tasks, we have characterized aspects of how searchers struggle and (in some cases) ultimately succeed. We found that struggling searchers issue fewer queries in successful tasks than in unsuccessful ones. In addition, queries are shorter, fewer results are clicked and the query reformulations indicate that searchers have more trouble choosing the correct vocabulary.

We have shown quite significant behavioral differences given task success and failure. This informed the development of a crowd-sourced labeling methodology to better understand the nature of struggling searches. We proposed and applied that method to better understand the struggling process and where it became clear the search would succeed. This pivotal query is often the last query and not all strategies are as likely to be pivotal. We developed classifiers to accurately predict key aspects of inter-query transitions for struggling searches, with a view to helping searchers struggle less.

Our research has limitations that we should acknowledge. First, we focused on a very specific and very apparent type of struggling, indicated by limited activity for the first two queries within a task. More forms of struggling exist and might exhibit different search behaviors. Our determinations of search success for the log analysis were based on inferences made regarding observed search activity, especially satisfied clicks based on dwell time. Although these have been used in previous work [13], and were validated by third-party judges as part of a dedicated judgment effort, there are still a number of factors that can influence landing page dwell time [26]. Finally, the crowd-sourced annotations were based on judgments from third-party judges and not the searchers themselves. While this methodology has been used successfully in previous work [20], methods to collect judgments in-situ can also be valuable [13].

Previous work has shown that it is possible to detect struggling automatically from behavior [20, 36]. Our focus has been on better understanding struggling during search and predicting query reformulation strategy. Ideally, a search engine would interpret the behavioural signals that indicate struggling and frustration to provide personalized help to searchers to help them attain task success. The types of support possible include:

Direct application of reformulation strategies: Demonstrating the capability to accurately predict the strategy associated with the next query reformulation (rather than syntactic transformations, as has traditionally been studied) allows us to provide situation-specific search support at a higher (more strategic) level than specific query formulations. For example, if we predict that a searcher is likely to perform an action such as adding an attribute, the system can focus on recommending queries with attributes in query suggestions or query auto-completions depending on when they are applied (and augmented with additional information about popularity and/or success if available from historic data). Sets of (query → pivotal query) pairs can also be mined from log data. Such queries may also be leveraged internally within search engines (e.g., in blending scenarios, where the results from multiple queries are combined [34]) to help generate better quality search result lists, or present them as suggestions to searchers.

Hints and tips on reformulation strategies: As we demonstrated, struggling searchers, especially those destined to be unsuccessful, are highly likely to re-query. Learning the relationship between task success and the nature of the anticipated query reformulation allows search systems to generate human-readable hints about which types of reformulations to leverage (e.g., "add an action" leads to the highest proportion of pivotal queries, per Figure 13), and propose them in real-time as people are searching. Mining these reformulations retrospectively from log data also allows search systems to identify the most successful query transitions in the aggregate—rather than focusing on proxies for success, such as query or resource popularity [24, 42]. These reformulations can be learned from all searchers, or perhaps even more interestingly, from advanced searchers [2, 42] or domain experts [43].

Overall, accurate inferences and predictions about the nature of query reformulations can help searchers and search engines reduce struggling. Although our findings are promising, the nature and volume of our human-labeled data limited the types of the prediction tasks that we attempted. There are others, such as predicting search success given different query reformulation strategies that are interesting avenues for future work. Additional opportunities include working directly with searchers to better understand struggling in-situ, improving our classifiers, and experimenting with the integration of struggling support in search systems.

Acknowledgements. The first author performed this research during an internship at Microsoft Research and is supported by the Dutch national program COMMIT.

REFERENCES

[1] E. Agapie, G. Golovchinsky, and P. Qvarfordt. Leading people to longer queries. In *CHI'13*, pages 3019–3022, 2013.

[2] M. Ageev, Q. Guo, D. Lagun, and E. Agichtein. Find it if you can: A game for modeling different types of web search success using interaction data. In *SIGIR'11*, pages 345–354, 2011.

[3] P. Anick. Using terminological feedback for web search refinement: a log-based study. In *SIGIR'03*, pages 88–95. ACM, 2003.

[4] R. Artstein and M. Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

[5] A. Aula, R. M. Khan, and Z. Guan. How does search behavior change as search becomes more difficult? In *CHI'10*, pages 35–44, 2010.

[6] P. Bennett, K. Svore, and S. Dumais. Classification-enhanced ranking. In *WWW'10*, pages 111–120, 2010.

[7] D. C. Blair. Searching biases in large interactive document retrieval systems. *JASIS*, 31(4):271–277, 1980.

[8] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and regression trees*. Wadsworth, 1984.

[9] D. Downey, S. Dumais, D. Liebling, and E. Horvitz. Understanding the relationship between searchers' queries and information goals. In *CIKM'08*, pages 449–458, 2008.

[10] C. Eickhoff, J. Teevan, R. White, and S. Dumais. Lessons from the journey. In *WSDM'14*, pages 223–232, 2014.

[11] H. Feild and J. Allan. Task-aware query recommendation. In *SIGIR'13*, pages 83–92, 2013.

[12] H. A. Feild, J. Allan, and R. Jones. Predicting searcher frustration. In *SIGIR'10*, pages 34–41, 2010.

[13] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *TOIS*, 23(2): 147–168, 2005.

[14] J. Guo, G. Xu, H. Li, and X. Cheng. A unified and discriminative model for query refinement. In *SIGIR'08*, pages 379–386, 2008.

[15] Q. Guo, R. W. White, Y. Zhang, B. Anderson, and S. T. Dumais. Why searchers switch: understanding and predicting engine switching rationales. In *SIGIR'11*, pages 335–344, 2011.

[16] A. Hassan. Identifying Web search query reformulation using concept based matching. In *EMNLP'13*, pages 1000–1010, 2013.

[17] A. Hassan, R. Jones, and K. L. Klinkner. Beyond DCG: User behavior as a predictor of a successful search. In *WSDM'10*, pages 221–230, 2010.

[18] A. Hassan, X. Shi, N. Craswell, and B. Ramsey. Beyond clicks: Query reformulation as a predictor of search satisfaction. In *CIKM'13*, pages 2019–2028, 2013.

[19] A. Hassan, R. W. White, and Y.-M. Wang. Toward self-correcting search engines: Using underperforming queries to improve search. In *SIGIR'13*, pages 263–272, 2013.

[20] A. Hassan, R. W. White, S. T. Dumais, and Y.-M. Wang. Struggling or exploring? Disambiguating long search sessions. In *WSDM'14*, pages 53–62, 2014.

[21] J. Huang and E. N. Efthimiadis. Analyzing and evaluating query reformulation strategies in web search logs. In *CIKM'09*, pages 77–86, 2009.

[22] S. B. Huffman and M. Hochster. How well does result relevance predict session satisfaction? In *SIGIR'07*, pages 567–574, 2007.

[23] J. Jiang, A. H. Awadallah, X. Shi, and R. W. White. Understanding and predicting graded search satisfaction. In *WSDM'15*, pages 57–66, 2015.

[24] R. Jones, B. Rey, O. Madani, and W. Greiner. Generating query substitutions. In *WWW'06*, pages 387–396, 2006.

[25] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *FnTIR*, 3(1–2):1–224, 2009.

[26] Y. Kim, A. Hassan, R. White, and I. Zitouni. Modeling dwell time to predict click-level satisfaction. In *WSDM'14*, pages 193–202, 2014.

[27] T. Lau and E. Horvitz. *Patterns of Search: Analyzing and Modeling Web Query Refinement*. 1999.

[28] Z. Liao, Y. Song, L.-w. He, and Y. Huang. Evaluating the effectiveness of search task trails. In *WWW'12*, pages 489–498, 2012.

[29] Q. Liu, E. Agichtein, G. Dror, Y. Maarek, and I. Szpektor. When web search fails, searchers become askers: understanding the transition. In *SIGIR'12*, pages 801–801, 2012.

[30] S. Lopez and C. Snyder. *The Oxford Handbook of Positive Psychology*. Oxford University Press, 2011.

[31] G. Marchionini. Exploratory search: from finding to understanding. *CACM*, 49(4):41–46, 2006.

[32] N. Moraveji, D. Russell, J. Bien, and D. Mease. Measuring improvement in user search performance resulting from optimal search tips. In *SIGIR'11*, pages 355–364, 2011.

[33] R. Oliver. *Satisfaction: A Behavioral Perspective on the Consumer*. ME Sharpe, 2011.

[34] K. Raman, P. N. Bennett, and K. Collins-Thompson. Toward whole-session relevance: Exploring intrinsic diversity in web search. In *SIGIR'13*, pages 463–472, 2013.

[35] D. Savenkov and E. Agichtein. To hint or not: exploring the effectiveness of search hints for complex informational tasks. In *SIGIR'14*, pages 1115–1118, 2014.

[36] A. Scaria, R. Philip, R. West, and J. Leskovec. The last click: Why users give up information network navigation. In *WWW'14*, pages 213–222, 2014.

[37] M. Shokouhi, R. W. White, P. Bennett, and F. Radlinski. Fighting search engine amnesia: Reranking repeated results. In *SIGIR'13*, pages 273–282, 2013.

[38] J. Teevan. The re: search engine: simultaneous support for finding and re-finding. In *UIST'07*, pages 23–32, 2007.

[39] H. Wang, Y. Song, M. Chang, X. He, A. Hassan, and R. White. Modeling action-level satisfaction for search task satisfaction prediction. In *SIGIR'13*, pages 123–132, 2013.

[40] R. White, M. Bilenko, and S. Cucerzan. Studying the use of popular destinations to enhance web search interaction. In *SIGIR'07*, pages 159–166, 2007.

[41] R. W. White and S. T. Dumais. Characterizing and predicting search engine switching behavior. In *CIKM'09*, pages 87–96, 2009.

[42] R. W. White and D. Morris. Investigating the querying and browsing behavior of advanced search engine users. In *SIGIR'07*, pages 255–262, 2007.

[43] R. W. White, S. T. Dumais, and J. Teevan. Characterizing the influence of domain expertise on web search behavior. In *WSDM'09*, pages 132–141. ACM Press, 2009.

[44] Y. Xu and D. Mease. Evaluating web search using task completion time. In *SIGIR'09*, pages 676–677, 2009.