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.

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.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

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

Copyright is held by the owner/author(s). Publication rights licensed to ACM.
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 intensity-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 a subject of study from different perspectives using a range of 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 [3, 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. Aghevli 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 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.
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, we can combine the results of multiple 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 a 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
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 within a task by considering characteristics at different stages.

For 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.

### 3.4 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 task search, 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].
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 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 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...
for struggling and success by three judges. We recruited a total of 88 judges from CrowdFlower (a crowd-sourcing service) with access to human annotators under contract to Microsoft. 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 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.

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 to 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 judge the success and pick the pivotal query. Subsequently, they spent twelve seconds to judge each query transition (at least two per task, depending on the number of queries).
Judges show 67% majority agreement on the task of judging the success of a searcher on a four-point scale. Krippendorff’s $\alpha$ 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 (18%) 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.

**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 of 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 characteristics that have been described in previous sections of the paper.

![Figure 12: Distribution of query reformulation types.](image-url)
We therefore only report the results of the RandomForest classifier. Logistic regression and SVM), and none yielded better performance.

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

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

Query features are used to represent the query before a reformulation. Interaction features describe how a searcher interacted with the search engine results pages (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, or 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.

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

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 the 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).

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

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

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).

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>F1</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Baseline</td>
<td>0.6245</td>
<td>0.5091</td>
</tr>
<tr>
<td>2. Query + Interaction</td>
<td>0.6352▲</td>
<td>0.5817▲</td>
</tr>
</tbody>
</table>

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 behavioral 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.

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 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