Struggling or Exploring? Disambiguating Long Search Sessions

Ahmed Hassan, Ryen W. White, Susan T. Dumais, and Yi-Min Wang
Microsoft Research
Redmond, WA 98052 USA
{hassanam, ryenw, sdumais, ymwang}@microsoft.com

ABSTRACT

Web searchers often exhibit directed search behaviors such as navigating to a particular Website. However, in many circumstances they exhibit different behaviors that involve issuing many queries and visiting many results. In such cases, it is not clear whether the user's rationale is to intentionally explore the results or whether they are struggling to find the information they seek. Being able to disambiguate between these types of long search sessions is important for search engines both in performing retrospective analysis to understand search success, and in developing real-time support to assist searchers. The difficulty of this challenge is amplified since many of the characteristics of exploration (e.g., multiple queries, long duration) are also observed in sessions where people are struggling. In this paper, we analyze struggling and exploring behavior in Web search using log data from a commercial search engine. We first compare and contrast search behaviors along a number dimensions, including query dynamics during the session. We then build classifiers that can accurately distinguish between exploring and struggling sessions using behavioral and topical features. Finally, we show that by considering the struggling/exploring prediction we can more accurately predict search satisfaction.

Categories and Subject Descriptors

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

Keywords

Exploratory search; Exploring vs. struggling.

1. INTRODUCTION

Web search engines are a primary mechanism by which people seek information and solve problems. When searchers experience difficulty in finding information, their struggle may be evident in their search behavior via indicators such as issuing numerous search queries or visiting many results within a search session [2]. However, these same search behaviors may also be indicative of *exploratory* search activity, whereby people actively try to learn a topic and discover new information [24][29]. Accurately distinguishing between struggling and exploring is an important issue for search engines. For example, during post-hoc analysis of search logs, the ability to distinguish between these two situations can help to more accurately identify underperforming scenarios, signaling user frustration or unhappiness. In addition, predictive models could be applied to provide appropriate system support if the search situation can be discerned, e.g., providing a revised experience for exploratory searches such as a guided tour through related topics [13][34]. Since searchers may be reluctant to describe their situation explicitly to the search engine it is be desirable to predict this automatically from observed search activity.

WSDM'14, February 24 - 28 2014, New York, NY, USA
Copyright 2014 ACM 978-1-4503-2351-2/14/02…$15.00.
http://dx.doi.org/10.1145/2556195.2556221

1:20:15 PM	**Query**	can you use h & r block software for more than one year
1:20:55 PM	**Query**	how do I file 2012 taxes on hr block
1:20:58 PM	**Click**	http://www.hrblock.com
1:33:17 PM	**Query**	can you only use h & r block one year
1:33:29 PM	**Click**	http://www.consumeraffairs.com/finance/hr_block_free.html
1:34:21 PM	**Click**	http://financialsoft.about.com/od/taxcut/gr/HR-Block-At-Home-...
1:36:23 PM	**Query**	do I have to buy new tax software every year
1:36:38 PM	**Click**	http://financialsoft.about.com/od/simpletips/f/upgrade_yearly.htm...
1:55:10 PM	**Click**	http://askville.amazon.com/buy-version-Tax-Software-year/Answer...
1:55:32 PM		**END OF SESSION**

(a) A *struggling* session

5:54:51 PM	**Query**	career development advice
5:55:03 PM	**Click**	http://www.sooperarticles.com/business-articles/career-devel...
5:55:48 PM	**Query**	employment issues articles
5:55:52 PM	**Click**	http://jobseekeradvice.com/category/employment-issues/...
6:01:02 PM	**Query**	professional career advice
6:01:05 PM	**Click**	http://ezinearticles.com/?Career-Advice-and-Professional-Ment...
6:03:09 PM	**Click**	http://askville.amazon.com/buy-version-Tax-Software-year/Answer...
6:03:35 PM	**Query**	what is a resume
6:04:21 PM	**Click**	http://en.wikipedia.org/wiki/R%C3%A9sum%C3%A9...
6:07:15 PM		**END OF SESSION**

(b) An *exploring* session

Figure 1. Examples of *struggling* and *exploring* sessions.

There is a need to develop methods to automatically identify struggle and exploration from search behavior. Research on leveraging implicit feedback [9][14] has shown that we can learn when users are dissatisfied based on features of their search activity such as short landing-page dwell times and post-click query reformulation. Previous work on behavioral analysis has shown that multiple query refinements and multiple search-result clicks are associated with users experiencing difficulty [2], frustration [7], and to events such as search engine switching [31]. However, more querying is not necessarily a negative indicator if the user is learning and consuming content on their journey [27], and searchers can benefit from the information that they are exposed to as they search [32].

To illustrate the challenge of distinguishing between struggling and exploring automatically, let us present an example of user behavior with a single search session. Figure 1 (a) shows an example of a search session where the user is seeking information on tax software. We can see that in this session the searcher issued many related queries and clicked on many results (four queries and five results in total), providing some evidence that they are exploring. However, the queries are closely related, and the inter-query time is relatively short for some queries (suggesting impatience). This provides stronger evidence that the user is in fact struggling to find relevant information pertaining to the annual purchase of specific

tax software. In Figure 1 (b), we see an example of a user exploring different aspects of a topic "career development" and issuing multiple queries and having multiple clicks. While the difference between the two types of sessions may be discernible to human judges, existing automatic methods may be unable to perform this distinction using limited information such as the number of queries in the session or session duration. To date, researchers have not developed methods to *automatically* distinguish between struggling and exploring. Such mechanisms would have utility for a range of applications, including, as we demonstrate later in the paper, enhancing search satisfaction models, e.g., [12]. We address this shortcoming with the research presented in this paper.

We use behavioral data gathered from a large commercial Web search engine and consensus judgments about the type of session from external human assessors. Through our analysis, we show clear behavioral differences between struggling and exploring, and use these insights to develop machine-learned models capable of accurately distinguishing between the two situations. Specifically, we make the following three research contributions with this work presented in this paper:

- Characterize differences in the search behavior associated with struggling and exploring.
- Build predictive models to distinguish between struggling and exploring given behavioral data.
- Integrate the prediction into model of search satisfaction and demonstrate gains in prediction accuracy by considering whether the searcher is struggling or exploring.

The remainder of this paper is structured as follows. In Section 2, we describe related work in areas such as search satisfaction and searcher frustration. Section 3 defines struggling versus exploring and describes the labeled data that we use in our analysis. In Section 4 we compare and contrast search behavior for each of search situations. Section 5 describes the predictive model and we provide the findings of our experiments in Section 6. We discuss the implications and limitations in Section 7, and conclude in Section 8.

2. RELATED WORK

There is relevant related work in four main areas: (1) search satisfaction, (2) task difficulty, (3) searcher frustration, and (4) exploratory search and multi-query analysis. We address each area in turn.

2.1 Search Satisfaction

There is significant literature on estimating task success or failure from online search behavior. Methods that have been used include correlating search behavior, such as search-result clicks and dwell time for clicks, with either self-reported success or labels of success provided by expert judges. Fox et al. [9] used an instrumented browser to determine whether there was an association between explicit ratings of satisfaction and implicit measures of searcher interest and identified the measures that were most strongly associated with user satisfaction. They found that there was a relationship between user activity and search satisfaction ratings, and that clickthrough, dwell time, and session termination activity combined to make good predictors of satisfaction for Web pages. Fox et al. found that short dwell times and clicking numerous (four or more) search results for a query were both indicators of dissatisfaction.

Behavioral patterns have also been used to predict user satisfaction for search sessions in addition to individual queries. Huffman and Hochster [15] found a relatively strong linear correlation between session satisfaction and the relevance of the first three results returned for the first query in a search task, whether the information need was navigational, and the number of events in the session. Hassan et al. [12] developed models of user behavior to accurately estimate search success on a session level, independent of the relevance of documents retrieved by the search engine. Ageev et al. [1] propose a formalization of different types of success for informational search, and presented a scalable game-like infrastructure for crowdsourcing search behavior studies, specifically targeted towards capturing and evaluating successful search strategies on informational tasks with known intent. They show that their model can predict search success effectively on their data and on a separate set of logs comprising search engine sessions. Later in the paper, we will show that by considering evidence of struggling or exploring, we can improve the performance of search satisfaction models such as that described in previous work [12].

2.2 Task Difficulty

One explanation for why users struggle is the difficulty of the task being attempted. A study by Aula and colleagues [2] examined the behavior of searchers engaged in challenging closed informational search tasks, where the answer was difficult to find. They examined users' search behavior in two studies—a usability study and an online study—and showed that there were differences in search behavior when searchers were experiencing difficulty in finding relevant information. Specifically, searchers applied more advanced operators, spent longer on the search result page, and issued the longest search query toward the middle of the search session. We explore the role of these and related behavioral signals in the analysis described later in the paper.

Other research on the effects of task difficulty on search behavior has shown that beyond queries, there are other behavioral signals that correlate with task difficulty. For example, as task difficulty increases, the number of results viewed increases and average dwell time on landing pages also increases [22], as does the number of pages retained for later inspection (e.g., via bookmarking [20]). Gwizdka and Spence [11] showed that the number of unique web pages visited, the time spent on each page, the degree of deviation from the optimal path, and the degree of the navigation path's linearity, were good predictors of subjective task difficulty. Beyond behaviors, studies have also shown that signals of task difficulty are mediated by domain knowledge, and that should be considered when interpreting those signals [23]. Information about domain expertise is typically unavailable to search engines, but it could be estimated from signals such as topical interests and estimates of the reading difficulty of viewed Web pages [21].

2.3 Searcher Frustration

Users experiencing difficulty in finding the information they seek may experience frustration that can manifest in search behavior and other signals (including physiological indicators). While satisfaction and frustration are closely related, they are distinct. Searchers can ultimately satisfy their information need, but still be quite frustrated in the process [4]. We may therefore need to consider frustration separately in developing models of user behavior.

Although others have studied frustration from an information science perspective (e.g., [19]) they have not attempted to model frustration in a way that could be utilized directly by search engines. Feild et al. [7] developed methods to predict user frustration from behavioral signals gathered during the search process. They assigned users difficult information seeking tasks and monitored their degree of frustration via behavioral logs and physical sensors. They showed that behavioral features such as the total duration of the session and query complexity were good predictors of search frustration. These aligned with features that were shown to be useful in predicting another behavior, search engine switching, defined as the voluntary transition between different search engines. White and Dumais [31] showed that there are behavioral patterns such as

a sequence of queries with no intervening clicks that could help predict engine switching.

2.4 Exploratory Search

As described earlier, during exploratory search users are focused on learning about a topic and gathering information [24]. This may be associated with searchers seeking different opinions on a topic, exploring or discovering aspects of a topic, or trying to ascertain an overview of a topic. Research on exploratory search has focused on characterizing the exploratory search process and the types of support that are required to help people perform exploratory searches [29]. For example, if it can be determined that the user is engaged in a complex search task then support can be offered to help them within the current session using tours or trails [13][34], tailored search support [3], or even across multiple search sessions with the ability to preserve and restore search state [25].

Other research has used within-session query reformulation for evaluation purposes, as well as supporting various aspects of the search process. Kanoulas and colleagues [19] proposed measures for evaluating search systems across multiple queries. Radlinski and Joachims [26] proposed the use of query chains of connected query reformulations, independent of exploring or struggling, to improve the ranking of search results. Recent research has studied intrinsically-diverse search tasks that typically require multiple user searches on different aspects of the same information need [27]. The authors proposed an approach that could alter the rankings presented to the user, to also provide them information on aspects of the task for which the user will search in the future.

The research presented in this paper extends previous work in a number of ways. First, we focus specifically on the important problem of *automatically* identifying exploring and struggling behavior, and distinguishing between the two situations even though they appear similar given coarse analysis of search behavior such as query reformulation counts or session duration. Second, we developed a labeling methodology to capture possible explanations for observed session behavior from external assessors. Using this approach we gather a large labeled set of struggling and exploring sessions from the logs of the Microsoft Bing search engine, labeled based on the consensus judgments of those assessors. Third, we characterize the key behavioral differences between struggling and exploring that help to inform feature generation in our predictive models. Finally, we develop predictive models and use the predictions to improve the performance of established models of search satisfaction.

3. STRUGGLING VS. EXPLORING

We begin by outlining the two different search scenarios and how we define them for the purposes of this study. We then describe the process by which we obtained labels on a sample of sessions taken from search engine logs.

3.1 Definitions

We focus our effort in this paper on studying long topically-coherent sessions (i.e., cases where users are observed pursuing information related to a common subject area). The behavior in these sessions could reflect either exploring or struggling. We adopt the following definitions for each of these concepts:

Definition: *Exploring sessions* are those where users are engaged in an open-ended and multi-faceted information-seeking task to foster learning and discovery.

Definition: *Struggling sessions* are those where users are experiencing difficulty locating the required information. Note that struggling may not necessarily result in failure (i.e., failing to locate the required information).

The challenge is in distinguishing between these types of sessions. In practice, users who are exploring or struggling issue multiple queries that share a common subject area.

In exploring sessions, multiple queries are intended to address different aspects of a topic and provide users with a combination of sought information and direction for future exploration. Because of the open-ended nature of such search sessions, the information goal is likely to be satisfied with information encountered during the search. Hence, multiple search queries can provide relevant information, reflecting increased engagement in the topic of interest.

In contrast, during struggling sessions, multiple queries are likely an indication of trouble in locating the required information. In such cases, people may have a well-defined information need and are trying to locate specific information, with little or no success.

Intuitively, both exploring and struggling sessions are topically coherent but they can have either multiple facets (exploring) or a single facet (struggling). In addition, both types of sessions are, by definition, long. A single-query session cannot be an exploring session because a single query cannot cover multiple facets of the search task. A single-query session cannot be a struggling session either because there is insufficient evidence that the user is experiencing difficulty in finding the information that they seek. Note, however, that a single-query session can be unsuccessful (e.g., a user who just gives up after a single query). We will elaborate on the difference between struggling and failure in Section 3.3.

3.2 Mining Search Sessions from Log Data

Our data consists of a sample of hundreds of thousands of search sessions from the logs of the Bing search engine. We analyzed a total of four weeks of interaction logs from March 2013. Log entries include a unique user identifier, and a timestamp for all queries and clicked Web pages. Intranet and secure (https) URL visits were excluded at the source. Any personally identifiable information was removed from the logs prior to analysis. In order to remove variability caused by geographic and linguistic variation in search behavior, we only include entries generated in the English speaking United States locale. Every session began with a query and could contain further queries or Web page visits. A session ended if the user was idle for more than 30 minutes. Similar criteria have been used in previous work to demarcate search sessions, e.g., [6][31].

To mine sessions that could potentially be related to exploring or struggling, we extract long sessions that exhibit topical coherence. To find sessions that meet these criteria, we do the following:

1. **Filter navigational queries:** Before we identify sessions. We start by collapsing all duplicate queries and removing the top 250 frequent navigational queries (e.g., facebook, amazon, etc.). Due to their navigational nature, we declare that these queries cannot be part of exploring or struggling sessions.
2. **Segment sessions into topically-coherent sub-sessions:** Since search sessions are segmented using a time threshold, it is likely that single sessions may contain multiple tasks [18] with unrelated information needs. To ensure that sessions are topically coherent, we predict two consecutive queries as belonging to the same topically-coherent session if they are no longer than 10 minutes apart and if one of the following conditions apply: (i) the two queries share at least one non-stop word terms, (ii) the two queries share at least one top ten search results, or (iii) the two queries share at least one domain name in their top ten results. Frequent domains that are not topically coherent (e.g., Wikipedia) were excluded.

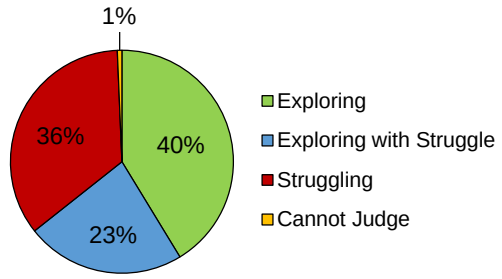

Figure 2. Session type distribution as labeled by judges.

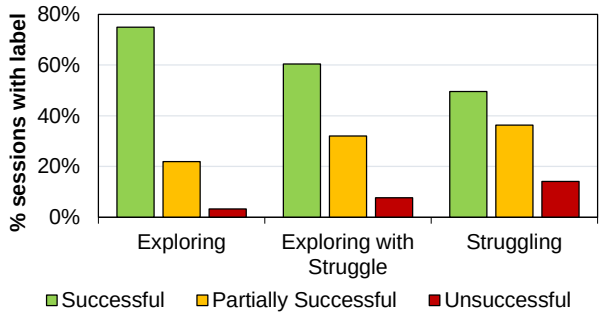

Figure 3. Session success distribution as labeled by judges.

3. **Filter short sessions:** After removing navigational queries and segmenting the logs into topically coherent sessions, we exclude sessions with less than three unique queries since these are unlikely to be exploring or struggling sessions.

We applied these criteria to identify long topically-coherent sessions because there are sessions in which users are likely to show exploring or struggling behavior. There were many thousands of such sessions in our data. We sampled 3000 of them and instructed external human judges to examine each session, try to understand the user's experience, and identify the reason for the observed behavior. We now describe the process by which the labels were collected from external judges.

3.3 Labeling Exploring and Struggling

Judges were recruited from the crowdsourcing service Clickworker.com, which provided access to crowd workers under contract. Judges resided in the United States and were fluent in English. Judges were shown sessions such as those illustrated in Figure 1. The interface was similar to Figure 1, and showed all queries, result clicks, and timestamps of all actions in the search session. They were instructed to examine the queries, the results pages (by clicking the "Query" text), the clicked pages, and to label the sessions as: *exploring*, *exploring with struggle*, or *struggling*. They were shown the definitions of exploring and struggling sessions using the same text as in the definitions provided in Section 3.1. Additionally, the following definition for *exploring with struggle sessions* was provided to judges:

Definition: *Exploring with struggle sessions* are exploring sessions where the user had experienced some difficulty in locating information about one or more of the facets being explored.

Judges were also instructed to label sessions with completely unrelated queries or sessions in a foreign language as "*cannot judge*". Figure 2 shows the distribution of labels collected from the judges. Judges excluded only 1% of the sessions as having unrelated queries or queries in a foreign language. Around 40% of the sessions were labeled as *exploring*, 23% as *exploring with struggle*, and the remaining 36% as *struggling* sessions. In total, the data set consisted of 3000 sessions with 17,117 queries, 13,168 distinct queries, and 13,780 result clicks.

We also asked the judges to assess the success of each session using the following labels:

- **Successful:** Sessions where searchers were able to locate the required information.
- **Partially Successful:** Sessions were searchers failed to locate some of the required information.
- **Unsuccessful:** Sessions where searchers failed to locate the required information.

Note that struggling is a characterization of the search process while success is a characterization of its outcome. Hence it is possible for a user who has difficulty locating the required information (struggling) to end up locating it (success). It is also possible for a user to fail in locating the required information without struggling (e.g., submit a single unsuccessful query then give up).

The distribution of success labels across session types is shown in Figure 3. The figure shows that most of the exploring sessions are successful (more than 75%) or partially successful (more than 20%). This agrees with our definition of exploring sessions, which are open-ended and multi-faceted in nature. Hence, failing to locate the required information is likely to prevent exploring early on in the session and in the cases when exploring does happen, the session is typically either successful or partially successful. Struggling sessions have a different success profile, with fewer sessions being successful (less than 50%) and more than 15% being unsuccessful.

The Cohen's kappa (κ) of inter-rater agreement is 0.59 for the exploring vs. struggling label, and 0.62 for the successful vs. unsuccessful label, signifying good agreement according to [8]. In both cases, we considered binary labels with *exploring* and *exploring with struggle* belonging to one class and *struggling* in the other. For the success label, *successful* was treated as one class and *partially successful* and *unsuccessful* were treated as another class. We use the same binary labels in the prediction task described later.

4. CHARACTERIZING EXPLORING AND STRUGGLING BEHAVIOR

In this section, we examine several characteristics of exploring and struggling sessions focusing on queries, result clicks, and topical dimensions.

4.1 Query Characteristics

At the outset of our analysis, we examine a number of different aspects of the session queries: (1) the number of unique queries, (2) similarity between queries, and (3) the nature of query transitions.

Number of unique queries: As described earlier, both exploring and struggling sessions are long by definition. One interesting question is whether exploring or struggling leads to longer sessions. If this was the case, then session length could be employed to discriminate between exploring and struggling. To answer this question, we compute the distribution over the number of unique queries for both types of session. We used unique queries to avoid counting the same query multiple times when the user refreshes the search page or hits the back button. The average number of unique queries per session was 4.50 and 4.36 for the exploring and struggling sessions respectively. The difference between the numbers of unique queries in the two search situations is not statistically significant at the 0.05 level according to a two-tailed *t*-test, suggesting that this is unlikely to be a distinguishing factor.

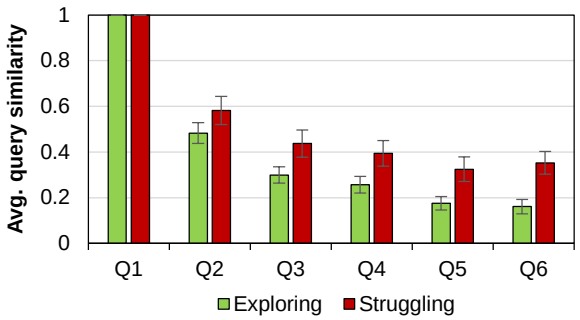

Figure 4. Average similarity between all queries to the first query in exploring and struggling sessions (±SEM).

Query similarity: Another interesting characteristic of exploring and struggling sessions may be the diversity of queries within the session. We might expect exploring sessions to contain less overlap as people revise their queries to explore alternatives. To examine this, we measure the similarity between every query in the session and the first query in the session. Our objective here is to assess how queries evolve as the user moves further into the session in both cases. If the user is struggling we might expect much of the initial query to still be present in future queries, but with terms being added or removed as the session proceeds.

To measure the similarity between pairs of queries in the session, we begin by performing standard text normalization where we lowercase the query text, replace all runs of whitespace characters with a single space, remove leading or trailing spaces, and remove stop words. Thus every query is represented as a bag of non-stop word terms. The similarity between any two queries Q_i and Q_j is computed as follows:

$$\frac{|Q_i \cap Q_j|}{|Q_i| + |Q_j| - |Q_i \cap Q_j|}$$

where $|Q_i|$ is the number of terms in query Q_i, and $|Q_i \cap Q_j|$ is the number of matched terms in Q_i and Q_j.

To calculate the number of matches in Q_i and Q_j ($Q_i \cap Q_j$), we consider two terms matched if any of the following criteria are met:

1. **Exact Match:** The two terms match exactly.
2. **Approximate Match:** To capture spelling variants and misspelling, we allow two terms to match if the Levenshtein edit distance between them is less than two.
3. **Lemma Match:** Lemmatization is the process of reducing an inflected spelling to its lexical root or lemma form. We match two terms if the lemmas of their tokens match.
4. **Semantic Match:** To capture semantic variants, we match two terms if their similarity according to the WordNet Wu and Palmer measure (wup) is greater than 0.5. The Wu and Palmer measure [33] calculates relatedness by considering the depths of the two synsets in the WordNet taxonomies, along with the depth of the Least Common Subsumer (LCS). The measure is computed as follows:

$$wup(t_i, t_j) = \frac{2 * depth(LCS)}{depth(t_i) + depth(t_j)}$$

where the depth of any synset in WordNet is the length of the path connecting it to the root node plus one.

Figure 4 show the average similarity between all queries to the first query in every session for exploring and struggling sessions ± the standard error of the mean (SEM). Query similarity was calculated as described earlier where exact, approximate, lemma, and semantic matching is taken into consideration. A similarity value of one indicates a perfect match while a similarity value of zero indicates no matched terms in the two queries. We notice from the figure that as users move through an exploring session the queries get less similar to the first query. This is different from struggling sessions where queries remain quite similar to the initial query. All differences reported in Figure 4 are statically significant at the 0.05 level according to a two-tailed t-test. This aligns with previous work [2], which also found that when people are engaged in difficult search tasks that query reformulations closely resemble their initial search.

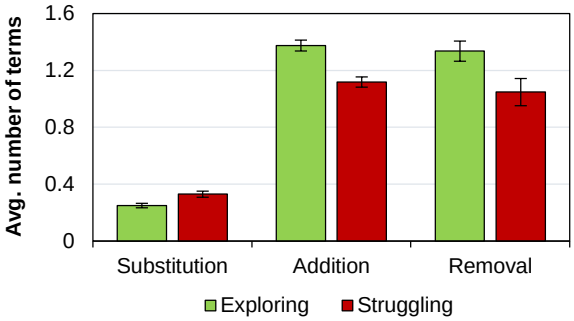

Figure 5. Average number of term addition, removal and substitution in exploring and struggling sessions (±SEM).

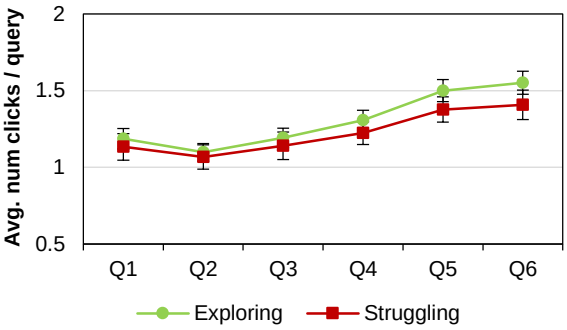

Figure 6. Average number of clicks per query for exploring and struggling sessions (±SEM).

Transition between queries: In addition to number of unique queries and similarity between queries, we also consider the strategies employed by the user when they transition from one query to another. There are multiple strategies for moving between queries. For simplicity, we consider the following three approaches, once again focusing on the relationship with the first query:

1. **Term Addition:** ≥ 1 words are added to first query.
2. **Term Removal:** ≥ 1 words removed from first query.
3. **Term Substitution:** This occurs when ≥ 1 words are substituted with a lexically or semantically matching terms. We treat one term is a substitution of another term if the two terms can be matched. Term matching is done by either exact, approximate, lemma, or semantic matching as described earlier.

Figure 5 compares the average number of term additions, removals and substitutions for exploring and struggling sessions. It is worth mentioning that our estimate of substitutions is clearly an underestimate because of the limited coverage of WordNet-based measures of synonyms. We notice from the figure that term addition and

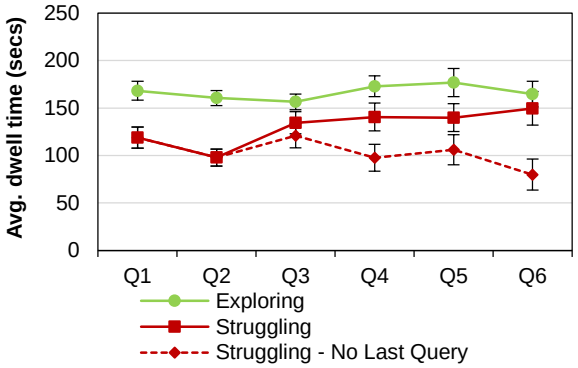

Figure 7. Average dwell time per query for exploring and struggling search sessions (±SEM).

removal are more popular reformulation strategies in exploring sessions (difference is statically significant at the 0.05 level according to a two-tailed *t*-test), while term substitution occurs somewhat more frequently in struggling sessions although this difference is not statistically significant at the 0.05 level. This aligns with our findings from the query similarity experiments which showed that struggling users are cycling as they attempt to conceive the correct query to locate a certain piece of information. Exploring users are more likely to remove old concepts and introduce new concepts as they progress in their search.

4.2 Click Characteristics

In addition to exploring properties of the queries that people issue, we also examine the characteristics of the search results they click. Clicks provide additional information about search intentions and the information that is encountered on the landing page can shape future search interactions.

Number of clicks: We suspected that exploring sessions might have more clicks than struggling sessions since struggling users are experiencing difficulty locating information. In our dataset, we have access to all clicks performed by the user during the search session. We excluded all non-result clicks (e.g., clicks on advertisements), as well as clicks that lead to another search result page (e.g., related search clicks, search vertical clicks, etc.). We then computed the average number of clicks for different query positions in the session for both exploring and struggling sessions. Computing the number of clicks at different points in the session allows us to observe changes in intentions and interests as the session proceeds. The result of this analysis is shown in Figure 6.

The figure shows that queries in exploring sessions contain more clicks compared to struggling sessions. The difference is small initially, and not statically significant up to Q3, and becomes larger later in the session, and statistically significant at the 0.05 level using a two-tailed t-test. Having multiple clicks in exploring sessions is expected since users are exploring multiple facets and hence are more likely to click on multiple results to locate information about these facets. Interestingly, the figures shows that struggling users tend to click on many results too, yet they still cannot locate the required information. The reason why they click may be related to the difference between the perceived relevance and the actual relevance of results. As shown in previous work [16][17], people tend to click on highly ranked results even if they are not relevant (position bias) and on results with good captions (caption bias) [5][35].

Dwell time: Another interesting question related to click characteristics is the difference in dwell time on clicked results in exploratory and struggling sessions. Dwell time reflects the time spent by

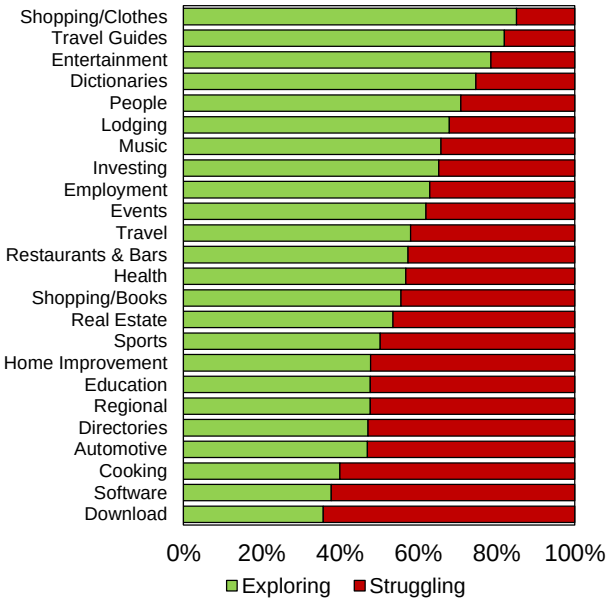

Figure 8. Likelihood for exploring and struggling for different search topics.

the user examining the clicked documents. The amount of time that people spend on pages can be an important indicator of whether they are satisfied with the content they encounter [9]. Dwell time can be estimated from click logs by computing the time between the click and the next seen click or query on the search engine.

We calculated the dwell time for every click in our dataset then we calculated the dwell time per query averaging the dwell time of all clicks corresponding to a single query. The average dwell time per query for both types of sessions is shown in Figure 7. The figure shows that clicks in struggling sessions increases during the session. However, it is much less than the dwell time for exploring sessions. The gap between the two types of sessions gets smaller as users progress through the sessions; especially in longer sessions. All difference are statically significant at the 0.05 level using a two-tailed t-test except for Q3 and Q6.

To understand the effect of the dwell time for clicks in the last query for the session, we remove the clicks of the last query before computing the average. This is important because many struggling sessions end up being successful and success typically occurs at the last query. Hence, the click dwell time of the last query in struggling sessions is likely different from other queries. There was no noticeable effect for removing the clicks of the last query in exploratory sessions. On the other hand, the average dwell time drops as users progress through struggling sessions if we ignore the last query. This shows that dwell time of clicks for the last query accounts for a large proportion of the total dwell time. This agrees with our earlier finding that many struggling sessions end up being successful or partly successful. Recall that in struggling sessions, users are typically trying to locate some information and, if successful, they typically find it at the last query and stop after finding it. Long dwell time has been shown to correlate with success in identifying the required information [9]. This explains the increase in dwell time for the last query in struggling sessions.

4.3 Topical Characteristics

Topical information has been used extensively to model search tasks and to capture users' intent. We wanted to understand the relationship between search situation and query topic. This may help

us to understand whether there are particular topics, where searchers are experiencing difficulty, and identify those where exploring may be most important. To do so, we assigned a topic to each search session and calculated the likelihood of a session being exploring or struggling given its topic (proportion of exploring/struggling sessions for each topic). The topic of every session is the most common topic assigned to the visited documents in this session. To assign topics to documents, we used the Open Directory Project (ODP, dmoz.org) ODP category labels. ODP is an open Web directory maintained by a community of volunteer editors. It uses a hierarchical scheme for organizing URLs into categories and subcategories. Many previous studies of Web search behavior have used ODP to assign topics to URLs, e.g., [28][32].

Given the large number of URLs in our set we needed to label them automatically. We performed automatic classification of URLs into ODP categories similar to [32]. URLs in the directory were directly classified according to the corresponding categories. Missing URLs were incrementally pruned one level at a time until a match was found or a miss declared. We used the top two levels of the ODP hierarchy as topic labels.

Figure 8 shows the likelihood of exploring vs. struggling for the most frequent topics in the dataset. Topics are given abbreviated names (e.g., Home/Cooking → Cooking, Reference/Dictionaries → Dictionaries) due to space considerations. The figures show exploration is much more likely in particular, topics compared to others. For example, some popular exploration topics are Shopping, Travel, Entertainment, People (e.g., Celebrities), etc. In contrast, searches in Local, Technical (e.g., Computer Software), Downloads, etc. topics are less likely to be exploratory. The strong association between exploring and the Dictionaries topic may be attributable to users researching the meaning of related concepts.

4.4 Summary

In this section, we have characterized some key aspects of struggling and exploring. We have shown that although there are no significant differences in basic behaviors such as number of queries (an important demonstration of the need for this research), there are behaviors (primarily queries rather than clicks) and topical differences that do differ between the two motivations. In the next section, we describe a model that leverages these features and others to distinguish between struggling and exploring search sessions.

5. PREDICTING SESSION TYPE

We begin by formally defining our prediction task and then introducing the features used for prediction.

5.1 Problem Definition

As stated earlier, we study long sessions on a common subject area and hypothesize that searchers are engaged in either exploring sessions or struggling sessions. Given a session with more than two queries on a common subject area, as described in Section 3.1, our objective is to predict whether the current search session is exploring or struggling. For every session, we have all queries submitted by the user and all clicks, with a timestamp associated with every query or click action.

5.2 Features

Our predictive model used the following five groups of features:

Query Features: These are the features that describe general characteristics of the queries in the session such as the number of queries, query length and how the queries were issued (manual vs. clicked). If the source of a query is a click on a related search, this may suggest that the user is more likely to be exploring.

Table 1. Features used to distinguish between exploring and struggling sessions. Features marked with an "*" had three versions corresponding to the minimum, maximum and average values per query/click.

Name	Description
Query Features	
NumQueries	Number of queries issued in session
CharQueryLen*	Query length in number of characters
WordQueryLen*	Query length in number of words
TimebetQueries*	Time between queries
PercManualQueries	Percentage of manually typed queries
PercSuggQueries	Percentage of clicked queries (e.g., query suggestions)
Query Transition Features	
AvgQuerySim*	Similarity between queries (see Section 4.1)
ExactMatch*	Number of terms that exactly match the previous query
AddTerms*	Number of added terms
DelTerms*	Number of removed terms
SubsTerms*	Number of substituted terms (see Section 4.1)
NumQGeneralizations	Number of queries where 1+ terms are removed from the previous query
NumQSpecifications	Number of queries where 1+ terms are added to the previous query
Click Features	
NumClicks	Total number of clicks in session
ClicksPerQuery	Average number of clicks per query
AbandonedQuerirs	Percentage of queries with no clicks
TotalDwellTime	Total dwell time in session
DwellTimePerClick*	Dwell time per click
DwellTimePerQuery*	Dwell time per query
TimeFirstClick*	Time to first click
UniqUrls	Percentage and number of unique clicked URLs
UniqDomains	Percentage and number of unique clicked domains
Search History Features	
QueryFreq*	Number of query impressions
QueryCTR*	Query clickthrough rate
QuerySuccessCTR*	Query success (dwell time > 30 sec) clickthrough rate
QueryQBCTR*	Query quickback (dwell time < 15 sec) clickthrough rate
QueryClickEntropy*	Entropy of click distribution
Topic Features	
Topic	Binary variable for every visited URL topic (see Section 4.3)
TopicRichness	Total number of unique topics per session
TopicEntropy	Topic distribution entropy

Query Transition Features: The second group of features characterizes the way in which queries evolve as users progress in the session. These features included the number of added terms, substituted terms, and removed terms as users transitioned between queries. It also had features for the number of query generalizations and specifications in the session. See Section 4.1 for details on how we measure query similarity, added terms, substituted terms, etc.

Click Features: Click features were meant to describe the click behavior of the user during the session. They included features representing the number of clicks, their dwell time ,and features to capture whether the user was clicking on the same documents or documents from the same domain multiple times, perhaps indicative of difficulty locating a particular resource.

Table 2. Performance of predicting exploring vs. struggling sessions. * indicates statistical significance at p < 0.05 using paired t-tests compared to the *First Query Text* baseline. A majority baseline has a 64% accuracy.

	Accuracy	Exploring F1	Struggling F1	AUC
First Query Text	73.64	76.45	70.21	78.88
After 1st Query	74.22	76.63	71.28	79.90
After 2nd Query	75.72*	77.85*	73.20*	80.90*
After 3rd Query	80.41*	82.02*	78.51*	**85.34***
End of Session	**81.67***	**83.68***	**79.17***	84.84*

Table 3. Performance of predicting successful sessions. * indicates statistical significance at p ≤ 0.05 using paired t -tests compared to the baseline.

	Accuracy	F1	AUC
Baseline	70.75	72.97	76.80
+ Explore/Struggle (Predicted)	74.14*	75.82*	82.47*
+ Explore/Struggle (Truth)	**76.82***	**77.51***	**84.43***

Search History Features: The next group of features summarizes the behavior of other users who have submitted this query (e.g., clickthrough rate, frequency, and click entropy). The rationale behind including these features was that prior aggregate behavior may reveal something about the nature of search tasks of this type or engine performance on these queries (e.g., low clickthrough rate or low fractions of successful clicks).

Topic Features: The last group of features describes the topical characteristics of the session. Every visited URL in the session was mapped to an ODP category as described in Section 4.3. Given these topic labels, we compute a binary label for the presence and absence of each of the topics present in the session, topic richness, and topic entropy. As we see in Figure 8, there are differences in the nature of the search situation for different topics.

A summary of the features used for prediction is shown in Table 1. We now describe the experiments performed and the results of comparisons against a number of different baselines.

6. EXPERIMENTS AND RESULTS

In this section we describe the experimental setup and the results for the prediction experiments. We also include a detailed analysis of the performance of the features that were used in the prediction experiment, as well as an application of our predictive model for improving the performance of search satisfaction models.

6.1 Experimental Setup

We performed several experiments to study the performance of our approach. We conducted experiments using the data described in section 3 which had approximately 3000 sessions and 13100 queries. We train classifiers to distinguish between *exploring* and *struggling* sessions. We treat both *exploring* and *exploring with struggle* sessions as one class and *struggling* sessions as another class. Combining the two exploring classes has two advantages: (1) since we are adding a noisier behavioral signal to the exploring class, we can be more conservative in our estimates of the prediction performance, and (2) we can cover 99% of the labeled search sessions, with the remaining 1% as "Cannot Judge" (see Figure 2).

The model that we evaluate uses the full session behavior to predict whether the session comprises struggling or exploring, referred to as "End of Session". We also experiment with several other baselines that have access to no user behavior at all (i.e., by only considering the text of the first query, its topic, etc.) or have access to limited user behavior (e.g., behavior of the user on the first query, first two queries, etc.). This allows us to understand the effect of different features on the performance and to assess the extent to which user behavior is required to make accurate predictions. This can be useful for applications where we are interested in distinguishing between the two search situations earlier in the session. We compare our approach to four different baselines:

1. **First Query Text:** The first baselines has access to the text of the first query in the session only with no access to any behavior information forcing the classifier to use only the topical and the search history features.
2. **After 1st Query:** The second baseline has access to the behavior associated with the first query plus the topical and search history information.
3. **After 2nd Query:** The third baseline resembles the previous baseline but has access to behavior from the first two queries.
4. **After 3rd Query:** This baseline has access to all behavior information for the first three queries. It also has access to the topical and search history features.

We used 10-fold cross validation for all experiments and Multiple Additive Regression Trees (MART) for classification [10]. The advantages of MART include model interpretability (e.g., a ranked list of features is generated), facility for rapid training and testing, and robustness against noisy labels and missing values. We experimented with other classifiers (specifically logistic regression and SVM), and they yielded similar or worse performance than MART, hence we only report the results of MART here.

6.2 Results

We now present the results of the prediction experiment. Table 2 presents the results of the *exploring* vs. *struggling* experiment. We report the performance of the five classifiers described in Section 6.1 (our predictive model plus four baselines). For every classifier we report accuracy, F1 measure of the positive class (exploring), F1 measure of the negative class (struggling), and area under the curve (AUC). We also compare the performance against the majority baseline that always predicts the label of the dominant class.

Using no user behavior information and relying on the query text only yields better results than simply using marginal class distributions (64% accuracy). This shows the predictive power of the topic and search history features. This can be explained by the analysis in Section 4.3 where we showed that exploring is more likely to occur in certain topics (e.g., Shopping, Entertainment, etc.). Search history features are also important for identifying struggling sessions since low clickthrough rate and low success clickthrough rate (success clicks are clicks with dwell time longer than 30 seconds) are likely to be correlated with queries that lead to struggle.

As more behavioral features are included, we start noticing improvement in performance. After including the behavior associated with the first query, we notice a small improvement, albeit not statically significant at the 0.05 level according to a paired t-test. Given the behavior on only one query, we only have click signals (e.g., number of clicks, click dwell time). Query transition features are still missing though. Apparently click features, especially when limited to the first query only) do not help much in terms of performance. As the classifier obtains access to more behavioral features, the performance improves. We get the best predictive performance when the classifier has access to the entire session.

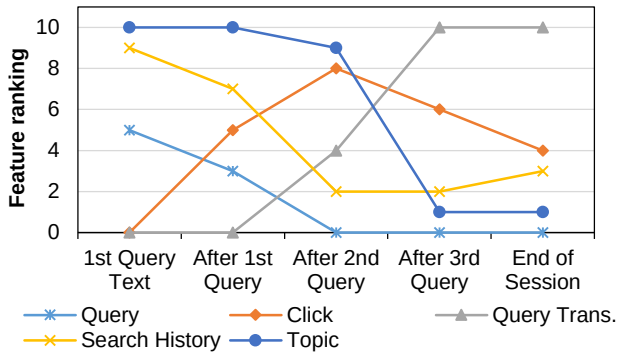

Figure 9. Feature group importance for all classifiers (0 = unused, 1-10 = relative rank (higher is better)).

6.3 Feature Analysis

We now turn our attention to the features that contribute most to the prediction task. In order to assess the importance of the proposed features, we compute the information gain of all features. We do that by computing the error reduction for each feature at each node split using the squared loss function. We then aggregate the error reduction at all splits for every feature and use that as a measure of feature gain. We do not report feature gains at the individual feature level for the five classifiers here for space considerations. Instead, we report feature importance at the feature group level. We assigned a score from 0 to 10 to each feature. The feature is assigned a 10 if it is ranked first according to feature gain, 9 if it is ranked second, and so on. Features that do not appear in the top 10 list are assigned 0. Every feature group is assigned a rank. The rank of any feature group is the rank of its best performing feature according to feature gain. We also tried to assign every group the average of the ranks of all its features and obtained similar results.

In Figure 9, we show the feature importance for every feature group for every classifier. We notice that query features are moderately important at the beginning when no behavior is available, they quickly lose their influence though as more behavior is made available to the classifier. Click features are not available at the beginning when only query text is available, but once the classifier gets access to user behavior, the importance of the click features quickly increases. It drops again though as more user behavior is available, yet remains moderately important. Topic features and search history features behave similarly. When limited behavior is available, they dominate all other features and are constantly ranked at the first positions. As more behavior becomes available, their importance drops quickly and they remain at the bottom of the list of the top features. Query transition features behave in exactly the opposite way to topic and search history features. At the beginning, they are not available to the classifier. However, as the classifier obtains access to more user behavior, they quickly climb up the list and eventually dominate all other features.

6.4 Application: Search Success Prediction

Finally, we study the effect of identifying exploring vs. struggling sessions on success prediction. Traditionally, multiple related queries in a session have been regarded as a sign of a searcher experiencing difficulty in locating required information. This is correct for struggling sessions, but for exploring sessions multiple related queries are not only not a sign of failure, but also a sign on increased engagement which can be regarded as a sign of search success.

To assess the effect of identifying session type on success prediction, we train a success prediction classifier following the work of Hassan and colleagues [12][14]. In that work, they represent every session as sequence of actions (i.e., queries and clicks). They show that the bigrams extracted from these sequences (e.g., Q-Click, Q-END, etc.) are very strong predictors of success. We use these features to train a Multiple Additive Regression Trees (MART) classifier using 10-fold cross validation. The performance of this classifier ("Baseline") is reported in Table 3. We compare this baseline to two other classifiers. The first uses the same features as the baseline in addition to another feature representing the session type (Exploratory or struggling) as predicted by the classifier described earlier. The second is just like the first but uses the truth session type label (from the human judges) as an oracle label instead of the predicted one. Results show significant improvement in performance due to adding the session type feature. We also observe that the gains get us a good way to the performance of an oracle model that was truth aware (56-74% of the maximum possible gain with our labeled data). The session type feature had the highest feature level information gain for the success prediction task.

7. DISCUSSION AND IMPLICATIONS

Search engines aim to understand whether they are meeting searchers' needs, and if and when they need to improve their engine or intervene to help improve the search process. Struggling and exploration behaviors may appear similar, but the situations have radically different experiences for the searchers involved. Searchers who are struggling may be dissatisfied and frustrated, whereas searchers who are exploring may be satisfied and content. Distinguishing between these two scenarios across a broad range of users and information needs is important for search engines in understanding search success and in providing search support. In this paper, we showed that we can build a classifier to accomplish this effectively, using features of recorded search interactions including query and topic dynamics. We also demonstrate an important application of our classifier: improving the accuracy of models of search satisfaction by considering struggling and exploring.

Although our paper shows that our method has strong potential, there are at least two limitations that we should acknowledge. The first is the nature of the labels that we applied to the data. The labels were assigned by third-party judges based on their consensus opinion of the nature of the sessions. While using the consensus may improve the reliability of the label, the fact that the judges were not those searching may lead to unforeseen issues with the labels, or for the system to simply learn what was important to the judges in making their predictions, rather than the true situation that the searcher faced. Ways to address this shortcoming to include soliciting judgments from searchers in-situ at the time of the search. This can be intrusive; however, previous work has shown that there are ways to accomplish such data capture practically, e.g., [9]. Another limitation is the way in which the struggling/exploring signal was integrated into the satisfaction model used as an example application. In our implementation, the additional signal was added as a feature – however, a potentially more powerful way would be to build separate models for success prediction for different session types that uses both general behavior patterns as well as session type specific information.

The application scenario that we have focused on in this paper has been the retrospective prediction of the rationale behind observed search behavior given the full session. We also studied how we can do such predictions at different points in the session to understand how the feature contributions evolve. While this form of retrospective prediction can have value for satisfaction analysis such as that presented in this paper and other applications, there is also potential benefit from applying this classifier in real time as the session pro-

ceeds. A real time classifier could estimate when the user is struggling or exploring and intervene to improve their experience by adapting it to suit the current need. For example, if they appear to be struggling, alternative queries could be suggested that have been shown to help searchers with similar needs. Additionally, hard-to-find resources could be presented if there is evidence that they have been sought and found in prior similar sessions [29]. If they are exploring, additional results or a new interface could be presented to help them discover and synthesize information more effectively.

8. CONCLUSIONS

There are two main reasons that searchers issue many queries during a session: (a) they are experiencing difficulty in finding required information, or (b) they are engaged in an exploratory search. Although being able to distinguish between these situations is important for improving search accuracy and experience, the behaviors associated with the situations are similar in terms of the number of queries and session duration. A more sophisticated analysis of search behavior is required to distinguish these search situations. In this paper, we have presented such methods and associated analysis. We have shown that there are differences in behavioral attributes such as query transitions and result clicks, as well as topic dynamics that can be useful in distinguishing struggling from exploring. We have also developed classifiers that have shown we can perform this prediction accurately. Future work involves the expansion of our research in this area to consider the real-time prediction of the search situation, as well as other applications of the predictor such as selecting query suggestions to bypass common areas of difficulty and get people to answers faster.

REFERENCES

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

[2] Aula, A., Khan, R., and Guan, Z. (2010). How does search behavior change as search becomes more difficult? *Proc. SIGCHI*, 35−44.

[3] Bron, M., Gorp, J., Vishneuski, A., Nack, F., Leeuw, S., and De Rijke, M. (2012). A subjunctive exploratory search interface to support media studies researchers. *Proc. SIGIR*, 425−434.

[4] Ceaparu, I., Lazar, J., Bessiere, K., Robinson, J., and Shneiderman, B. (2004). Determining causes and severity of end-user frustration. *Intl. J. of HCI*, 17(3): 333-356.

[5] Clarke, C.L.A., Agichtein, E., Dumais, S. and White, R.W. (2007). The influence of caption features on clickthrough patterns in web search. *Proc. SIGIR*, 135–142.

[6] Downey, D., Dumais, S., Liebling, D., and Horvitz, E. (2008). Understanding the relationship between searchers' queries and information goals. *Proc. CIKM*, 449–458.

[7] Feild, H., Allan, J., and Jones, R. (2010). Predicting searcher frustration. *Proc. SIGIR*, 34−41.

[8] Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions* (2nd edition). New York: John Wiley.

[9] Fox, S., Karnawat, K., Mydland, M., Dumais, S.T., and White, T. (2005). Evaluating implicit measures to improve the search experience. *ACM TOIS*, 23(2): 147−168.

[10] Friedman, J.H., Hastie, T., and Tibshirani, R. (1998). *Additive Logistic Regression: A Statistical View of Boosting.* Technical Report, Stanford University.

[11] Gwizdka, J. and Spence, I. (2006). What can searching behavior tell us about the difficulty of information tasks? A study of web navigation. *Proc. ASIST*, vol. 43: 1−22.

[12] Hassan, A., Jones, R., and Klinkner, K.L. (2010). Beyond DCG: User behavior as a predictor of a successful search. *Proc. WSDM*, 221–230.

[13] Hassan, A. and White, R.W. (2012). Task tours: helping users tackle complex search tasks. *Proc. CIKM*, 1885–1889.

[14] Hassan, A. (2012). A semi-supervised approach to modeling web search satisfaction. *Proc. SIGIR*, 275–284.

[15] Huffman, S. and Hochster, M. (2007). How well does result relevance predict session satisfaction? *Proc. SIGIR*, 567−574.

[16] Joachims, T. (2002). Evaluating search engines using clickthrough data. *Proc. SIGKDD*, 133–142.

[17] Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. *Proc. SIGIR*, 154–161.

[18] Jones, R. and Klinkner, K.L. (2008). Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs. *Proc. CIKM*, 699-708.

[19] Kanoulas, E., Carterette, B., Clough, P.D., and Sanderson, M. (2011). Evaluating multi-query sessions. *Proc. SIGIR*, 1053−1062.

[20] Kim, J. (2006). Task difficulty as a predictor and indicator of web searching interaction. *Proc. SIGCHI*, 959−964.

[21] Kim, J.Y., Collins-Thompson, K., Bennett, P.N., and Dumais, S.T. (2012). Characterizing web content, user interests, and search behavior by reading level and topic. *Proc. WSDM*, 213−222.

[22] Liu, J., Liu, C., Gwizdka, J., and Belkin, N.J. (2010). Can search systems detect users' task difficulty? Some behavioral signals. *Proc. SIGIR*, 845−846.

[23] Liu, C., Liu, J., Cole, M., Belkin, N.J., and Zhang, X. (2012). Task difficulty and domain knowledge effects on information search behaviors. *Proc. ASIST*, 49(1): 1−10.

[24] Marchionini, G. (2006). Exploratory search: From finding to understanding. *CACM*, 49(4): 41−46.

[25] Morris, D., Morris, M.R., and Venolia, G. (2008). SearchBar: A search-centric web history for task resumption and information refinding. *Proc. SIGCHI*, 1207−1216.

[26] Radlinski, F. and Joachims, T. (2005). Query chains: Learning to rank from implicit feedback. *Proc. SIGKDD*, 239–248.

[27] Raman, K., Bennett, P.N., and Collins-Thompson, K. (2013). Toward whole session relevance: Exploring intrinsic diversity in web search. *Proc. SIGIR*, 463–472.

[28] Shen, X., Dumais, S., and Horvitz, E. (2005). Analysis of topic dynamics in web search. *Proc. WWW*, 1102–1103.

[29] White, R.W. and Chandrasekar, R. (2010). Exploring the use of labels to shortcut search trails. *Proc. SIGIR*, 811–812.

[30] White, R.W. and Roth, R.A. (2009). *Exploratory Search: Beyond the Query-Response Paradigm*. Morgan Claypool.

[31] White, R.W. and Dumais, S.T. (2009). Characterizing and predicting search engine switching behavior. *Proc. CIKM*, 87−96.

[32] White, R.W. and Huang, J. (2010). Assessing the scenic route: Measuring the value of search trails in web logs. *Proc. SIGIR*, 587–594.

[33] Wu, Z. and Palmer, M. (1994). Verb semantics and lexical selection. *Proc. ACL*, 133–138.

[34] Yuan, X. and White, R.W. (2012). Building the trail best traveled: Effects of domain knowledge on web search trailblazing. *Proc. SIGCHI*, 1795–1804.

[35] Yue, Y., Patel, R., and Roehrig, H. (2010). Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. *Proc. WWW*, 1011–1018.