Predicting Search Task Difficulty

Jaime Arguello

University of North Carolina at Chapel Hill
jarguello@unc.edu

Abstract. Search task difficulty refers to a user's assessment about the amount of effort required to complete a search task. Our goal in this work is to learn predictive models of search task difficulty. We evaluate features derived from the user's interaction with the search engine as well as features derived from the user's level of interest in the task and level of prior knowledge in the task domain. In addition to user-interaction features used in prior work, we evaluate features generated from scroll and mouse-movement events on the SERP. In some situations, we may prefer a system that can predict search task difficulty early in the search session. To this end, we evaluate features in terms of whole-session evidence and first-round evidence, which excludes all interactions starting with the second query. Our results found that the most predictive features were different for whole-session vs. first-round prediction, that mouseover features were effective for first-round prediction, and that level of interest and prior knowledge features did not improve performance.

1 Introduction

Search engine users engage in a wide variety of search tasks. A large body of prior research focused on characterizing different types of search tasks (see Li and Belkin [10]). The motivation behind this prior work is to understand how task characteristics influence search behavior and how search systems can provide customized interactions for different task types. One important search task characteristic is *search task difficulty*. Search task difficulty refers to the user's assessment about the amount of effort required to complete the search task. In this work, we learn and evaluate predictive models of *post-task* difficulty, which refers to the user's assessment *after* completing the search task. Predicting search task difficulty has important implications for IR. First, it can help system designers determine the types of search tasks that are not well-supported by the system. Second, it can help researchers discover correlations between search task difficulty and undesirable outcomes such as search engine switching. Finally, predicting search task difficulty in real-time would enable a system to intevene and assist the user in some way.

To train and evaluate our models, we first conducted a user study to collect search-interaction data and post-task difficulty judgments from searchers. In order to collect data from a large number of search sessions and users, the study was conducted using crowdsourcing. Participants were given carefully constructed search tasks and asked to use our search system to find and bookmark webpages that would be useful in completing the task. We used search tasks that we

thought would cause participants to experience varying levels of difficulty. After completing each search, participants were given a post-task questionnaire that included several questions about the level of difficulty experienced while searching. Responses to these questions were averaged into single difficulty scale and this measure was used to group search sessions into *easy* and *difficult* searches. Our goal was to learn models to predict whether a search session was considered easy or difficult using behavioral measures derived from the search session. We investigate features derived from queries, clicks, bookmarks, mouse-movement and scroll events on the SERP, dwell-times, and the session duration.

Past studies also considered the task of predicting post-task difficulty using behavioral measures from the search session [12, 13]. Our work is the most similar to Liu *et al.* [13], with three main methodological differences. First, we used a larger set of search sessions for training and testing our models (600 vs. 117 [13]). Second, we used a larger number of participants (269 vs. 38 [13]), which potentially introduced more variance in search behavior. Third, we used more search tasks (20 vs. 5 [13]). Using more search tasks allowed us to avoid training and testing on search sessions from the same task. Thus, we believe that our evaluation emphasizes a model's ability to generalize to previously unseen users and tasks.

In addition to differences in methodology, we extend prior work in two directions: (1) we investigate new sources of evidence and (2) we investigate predicting task difficulty at different stages in the search session. In addition to using similar features used in prior work, we experimented with features derived from mouse movement and scrollbar events on SERPs produced by the system. To our knowledge, this is the first study to consider mouse and scroll data for predicting search task difficulty. Additionally, we experimented with features derived from the user's level of interest in the task and level of prior knowledge (domain knowledge and search experience). Our goal was not to *infer* this information about the user. Instead, as a first step, we wanted to assess the value of this information for predicting task difficulty. Thus, our level of interest and prior knowledge features were derived from responses to a pre-task questionnaire.

In certain situations, we may want the system to predict task difficulty *before* the end of the session. This may be the case, for example, if the goal to intervene and assist the user. To this end, we divided our analysis in terms of *whole-session analysis* and *first-round analysis*. Our first-round analysis excludes all search interactions starting with the second query (if any). We evaluate different types of features based on whole-session evidence and first-round evidence.

2 Related Work

A large body of prior work has focused on defining different task characteristics or dimensions (see Li and Belkin [10]). Two different, yet sometimes confused characteristics are task complexity and task difficulty. In this work, we make the same distinction made by Kim [9] and Li and Belkin [10]. Task complexity is an inherent property of the task, independent of the task doer, while task difficulty refers to a user's assessment about the amount of effort required to complete the task. In our study, we manipulated task complexity in order for our participants

to experience varying levels of difficulty. Prior work found that task complexity influences task difficulty [3, 18] and we observed a similar trend in our study.

Different characterizations of task complexity have been proposed [3–5, 8, 17]. Jansen *et al.* [8] (and later Wu *et al.* [18]) defined task complexity in terms of the amount of cognitive effort and learning required to complete the task. To this end, they adopted a taxonomy of learning outcomes originally developed by Anderson and Krathwohl for characterizing educational materials [1]. We used search tasks created using this *cognitive* view of task complexity.

Several studies investigated the effects of task difficulty or complexity on search behavior [2, 8, 11, 13, 14, 18]. Results show that task difficulty and complexity affect a wide range of behavioral measures. For example, difficult tasks are associated with longer completion times, more queries, more clicks, more clicks on lower ranks, more abandoned queries, more pages bookmarked, longer landing page and SERP dwell-times, and greater use of query-operators. Given the correlation between task difficulty and different behavioral measures, prior work also focused on predicting search task difficulty [13, 12]. Liu *et al.* [13] combined a large set of features in a logistic regression model and were able to predict two levels of post-task difficulty with about 80% accuracy. Our work builds upon Liu *et al.* [13] and investigates new features derived from the whole search session and from only the first round of interactions.

Prior work also focused on predicting user actions and emotions likely to be related to task complexity and difficulty. White and Dumais [16] focused on predicting search engine switching—whether a user's next action will be to switch to a different search engine—and evaluated features from the search session, from the user's history, and from interactions from other users for the same query. Most features were found to be complementary. Field *et al.* [7] focused on predicting searcher frustration. Searchers were periodically asked about their level of frustration and the goal was to predict the user's response. Feild *et al.* combined search interaction features with physical sensor features derived from a mental state camera, a pressure sensitive mouse, and a pressure sensitive chair. Interestingly, the search interaction features were more predictive than the physical sense features.

3 User Study

In order to train models to predict task difficulty, it was necessary to run a user study to collect search interaction data and difficulty judgements. Participants were given a search task and asked to use a live search engine to find and bookmark webpages that would help them accomplish the task. The user study was run using Amazon's Mechanical Turk (MTurk).[1] Using MTurk allowed us to collect data from a large number of participants. Each MTurk HIT corresponded to one search session. Our HITs were implemented as external HITs, meaning that everything besides recruitment and compensation was managed by our own server. This allowed us to control the assignment of participants to search tasks

[1] Mechanical Turk is a crowdsourcing marketplace where *requesters* can publish *human intelligence tasks* or *HITs* for *workers* to complete in exchange for compensation.

and to record all user interactions with the search system. Search results were returned by our server using the Bing Web Search API. As described below, we used the same set of 20 search tasks developed by Wu *et al.* [18]. Each of the 20 search tasks was completed by 30 unique MTurk workers for a total of 600 search sessions. Tasks were assigned to participants randomly, except for two constraints: (1) participants were not allowed to "see" the same search task more than once and (2) in order to gather data from a large number of participants, each worker was not allowed to complete more than eight tasks. Each HIT was priced at $0.50 USD. To help filter malicious workers, we restricted our HITs to workers with an acceptance rate of 95% or greater and, to help ensure English language proficiency, to workers in the US.

All user interaction data was recorded at the server-side. Clicks on search results were recorded using URL re-directs. Clicking on a search result opened the landing page in an HTML frame embedded in a webpage produced by our system. In order to record landing-page dwell-times, we used Javascript and AJAX to catch focus and blur events on this page and communicate these events to our server. Similarly, we used Javascript and AJAX to record and communicate scrolls and mouse movements on the SERP.

Experimental Protocol. Upon accepting the HIT, participants were first given a set of instructions describing the goal of the HIT and the search interface (e.g., how to add/delete bookmarks and view the current set of bookmarks). After clicking a "start" button, participants were shown the search task description and were asked to carefully read the task. Following this, participants were asked to complete a pre-task questionnaire (described below). After completing the pre-task questionnaire, participants were directed to the search interface. Participants were instructed to search naturally by issuing queries and clicking on search results. Clicking a search result opened the landing page inside an HTML frame. Participants were able to bookmark a page using a button labeled "bookmark this page" located above the HTML frame. While bookmarking a page, participants were asked to provide a 2-3 sentence justification for why they bookmarked the page. Participants were not allowed to leave this field blank. At any point in the search process (either from the search interface, the landing page display, or the bookmark view page), participants were able to revisit the task description and to review the current set of bookmarks. From the bookmark view page, participants were able to delete bookmarks and to terminate the search task and proceed onto a post-task questionnaire (described below).

Pre-task Questionnaire. The pre-task questionnaire was completed immediately after reading the search task description. Participants were asked one question about their level of interest and indicated their responses on a five-point scale: How interested are you to learn more about the topic of this task? (not at all interested, slightly interested, somewhat interested, moderately interested, and very interested). Participants were asked two questions about their level of prior knowledge and indicated their responses on a four-point scale: (1) How much do you already know about the topic of this task? (nothing, a little, some,

a great deal) and (2) How many times have you searched for information about this task? (never, 1-2 times, 3-4 times, 5 times or more).

Post-task Questionnaire. The post-task questionnaire was completed after terminating the search task. Participants were asked five questions about task difficulty. The first four asked about the amount of effort expended on different search-related activities: (1) How difficult was it to *search* for information for this task? (2) How difficult was it to *understand* the information the search engine found? (3) How difficult was it to *decide* if the information the search engine found would be *useful* in completing the task? and (4) How difficult was it to determine when you had *enough information* to finish? The fifth question was designed to elicit a summative judgment about the task difficulty: (5) Overall, how difficult was the task? Responses were indicated on a five-point scale: not at all difficult, slightly difficult, somewhat difficult, moderately difficult, and very difficult. We averaged responses to these five questions to form a single difficulty measure. Participant responses indicated a strong internal consistency (Cronbach's $\alpha = .903$).

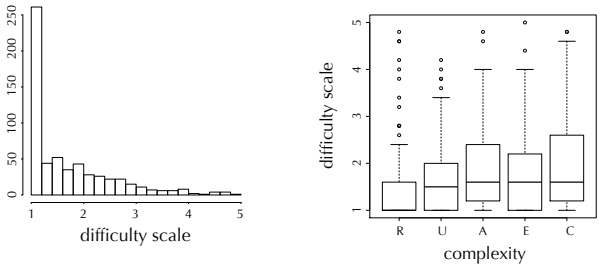
Search Tasks. We manipulated task complexity in order for participants to experience varying levels of difficulty. To accomplish this, we used the same set of 20 search tasks developed by Wu *et al.* to study the effects of task complexity on search behavior [18]. The tasks from were constructed to reflect different levels of *cognitive* complexity, which refers to the amount of learning and cognitive effort required to complete the task, and are evenly distributed across 4 topical domains (commerce, entertainment, health, and science & technology) and 5 cognitive complexity levels from Anderson and Krathwol's Taxonomy of Learning [1]:

- **Remember:** Retrieving relevant knowledge from long-term memory.
- **Understand:** Constructing meaning through summarizing and explaining.
- **Analyze:** Breaking material into constituent parts and determining how the parts relate to each other and the whole.
- **Evaluate:** Making judgments through checking and critiquing.
- **Create:** Putting elements together to form a new coherent whole.

Figure 1(a) shows the overall distribution of our difficulty scale across all 600 searches and Figure 1(b) shows the individual distributions for each cognitive complexity level. Two trends are worth noting. First, the 20 tasks were not found to be too difficult ($M = 1.749$, $SD = 0.866$). The median difficulty was 1.400. Second, more complex tasks were perceived to be more difficult, which is consistent with previous studies [3, 18]. A Kruskal-Wallis test showed a significant main effect of task complexity on difficulty ($\chi^2(4) = 36.60, p < .001$). Bonferroni-adjusted (Mann-Whitney) post-hoc tests showed significant differences between Remember (R) and all other complexity levels (U, A, E, C).

4 Predicting Task Difficulty

In this work, we cast the difficulty prediction problem as a binary classification problem. Search sessions were grouped into *easy* and *difficult* using a mean split. Search sessions with a difficulty rating equal to or lower than the mean (1.749) were considered easy and search sessions with a difficulty rating greater



(a) Difficulty (b) Difficulty vs. complexity

Fig. 1. Difficulty scale distribution.

than the mean were considered difficult. A mean-split resulted in 358 easy and 242 difficult search sessions. We trained L2-regularized logistic regression models using the LibLinear Toolkit [6]. In prior work, logistic regression performed well for the task of predicting search task difficulty [13, 12] and related tasks such as predicting search engine switching [16] and searcher frustration [7].

Features. Our models were trained to predict task difficulty as a function of a set of features. We were interested in evaluating different types of features. Thus, we organized our features into the following categories: *query features* were derived from the queries issued to the system, *click features* were derived from clicks on search results, *bookmark features* were derived from bookmarked webpages, *mouse* and *scroll features* were derived from mouse movements and scroll events on the SERP, *dwell-time features* were derived from time spent on a landing page, and *duration features* were derived from the task completion time. In addition to features derived from user-interaction data, we were interested in assessing the value of knowing the user's level of interest in the task and level of prior knowledge of the task domain (domain knowledge and search experience). We did not attempt to infer this information about the user. As a first step, we used participant responses from the pre-task questionnaire.

Before describing our features, a few concepts need clarification. A search result *click* is as instance where the participant clicked on a search result. In contrast, a result *view* is an instance where the participant actually inspected the landing page (recall that we recorded focus and blur events in our landing page display). In most browsers, users can right or mouse-wheel click a result to open the landing page in a hidden tab. In some cases, participants right or mouse-wheel clicked a result, but did not actually opened the hidden tab. A *pagination click* is one where the participant requested results beyond the top 10. A *mouseover* is an instance where the participant's mouse-pointer entered a transparent bounding-box around a search result. Finally, the *scroll position* is a number in the range [0,100] indicating the position of the scrollbar on the SERP (0 = top and 100 = bottom).

We evaluated different feature types based on whole-session and first-round evidence. The following list summarizes our features. Features associated with 'per-query' statistics were not used in our first-round analysis (which included only those interactions associated with the first query). Features included in our first-round analysis are marked with a '*'.

- **Query Features**
  - NumQueries: number of queries.
  - AvgQueryLength: average number of terms per query.
  - NumQueryTerms*: total number of query-terms.
  - UniqueQueryTerms*: total number of unique query-terms.
  - TokenTypeRatio*: NumQueryTerms / UniqueQueryTerms
  - AvgStopwords*: average percentage of stopwords per query.
  - AvgNonStopwords*: average percentage of non-stopwords per query.
  - NumAOLQueries*: total number of queries found in the AOL query-log.
  - NumQuestionQueries*: total number of queries with question words.
- **Click Features**
  - NumClicks*: total number of search results clicked.
  - AvgClicks: average number of clicks per query.
  - AvgClickRank*: average rank associated with all clicks.
  - AvgTimeToFirstClick*: average time between a query and the first click.
  - NumViews*: total number of search results viewed.
  - AvgViews: average number of views per query.
  - AvgClickRank*: average rank associated with all views.
  - NumPageClicks*: total number of pagination clicks.
  - NumAbandon*: total number of queries with no clicks
  - PercentAbandon: percentage of queries with no clicks.
- **Bookmark Features**
  - NumBook*: total number of pages bookmarked
  - AvgBook: average number of bookmarks per query.
  - AvgBookRank*: average rank associated with all bookmarks.
  - NumQueriesWithBook: total number of queries with a bookmark
  - PercentQueriesWithBook: percentage of queries with a bookmark
  - NumQueriesWithoutBook: total number of queries without a bookmark.
  - PercentQueresWithoutBook: percentage queries with without a bookmark.
  - NumClicksWithoutBook*: total number of clicks without a bookmark.
  - PercentClicksWithoutBook: percentage of clicks without a bookmark.
  - NumViewsWithoutBook*: total number of views without a bookmark.
  - PercentViewsWithoutBook: percentage of views without a bookmark.
- **Mouse Features**
  - TotalMouseovers*: total number of mouseovers in the session.
  - AvgMouseovers: average number of mouseovers per query.
  - MaxMouseover*: max mouseover rank in the session.
  - AvgMaxMouseover: average max mouseover rank per query.
- **Scroll Features**
  - TotalScrollDistance*: total scroll distance in session.
  - AvgScrollDistance: average scroll distance per query.
  - MaxScrollPosition*: max scroll position in session.
  - AvgMaxScrollPosition: average max scroll position per query.
- **Dwell-time Features**
  - TotalDwell*: total time spent on landing pages.
  - AvgDwell*: average time spent on a landing page.
- **Duration Feature**
  - Duration*: total time to task completion.
- **Interest Feature**
  - Interest*: pre-task level of interest response.
- **Prior Knowledge Features**
  - PriorKnowledge*: pre-task level of prior knowledge response.
  - PriorSearch*: pre-task level of prior search experience response.

As described below, we trained and tested our models using cross-validation. For each individual train/test pair, all features were normalized to zero minimum and unit maximum using the training set min/max values.

5 Evaluation Methodology

We collected user-interaction and self-report data for 20 different search tasks and each search task was completed by 30 different study participants, for a total of $20 \times 30 = 600$ search sessions. Training and testing was done using cross-validation. In a production environment, search engine users are likely to engage in a large number of search tasks. For this reason, we felt it was important to not include search sessions for the same task in the training and test data. In other words, we wanted to test a model's ability to generalize to previously unseen tasks. To this end, we used 20-fold cross-validation. Each training set corresponded to the 570 search sessions associated with 19 tasks and each test set corresponded to the 30 search sessions associated with the held-out task. We can also view this as leave-one-task-out cross-validation. Regularized logistic regression uses parameter C to control the misclassification cost on the training data. Parameter C was tuned using a second-level of cross-validation. For each top-level train/test pair, we conducted a second level of 19-fold cross-validation on each training set and used the value of C with the greatest average performance. Parameter C was tuned across values of 2^x where $x = -4, -3, -2, -1, 0, 1, 2, 3, 4$.

Prediction performance was measured using average precision (AP). Logistic regression outputs a prediction confidence value (the probability that the search session is difficult). We used average precision to evaluate a model's ability to rank search sessions in descending order of difficulty. Average precision is proportional to the area under the precision-recall curve. Statistical significance was tested using an approximation of Fisher's randomization test [15]. We report the mean of AP values across our 20 training/test-set pairs. Thus, the randomization was applied to the 20 pairs of AP values for the two models being compared.

6 Results

Our goal was to evaluate different features using whole-session and first-round evidence. Before presenting our classification results, we present an analysis of each feature in isolation. Results in Table 1 show differences in feature values between easy and difficult searches, both in terms of evidence aggregated at the whole-session and first-round level. We used non-parametric Mann-Whitney U tests to compare feature values between easy and difficult searches.

In terms of whole-session evidence, most features had significant differences. Difficult searches had more interaction: longer search sessions, more queries, more clicks and bookmarks, lower-ranked clicks and bookmarks, more pagination clicks, more mouseovers and scrolls on the SERP, and lower-ranked mouseovers and scrolls on the SERP. Difficult searches also had more backtracking: more queries without a click, more clicks without a bookmark, and shorter dwell-times, which suggests more clicks where the participant quickly found the landing page not useful. Features characterizing the types of queries issued were not significantly different. These included the average query length, average number of stopwords, number of queries with a question word, and number of queries in the AOL query-log, which we used as a proxy for query popularity (easy tasks).

In terms of first-round evidence, fewer features had significant differences. There were no significant differences in the number of clicks, views, and book-

marks, and no significant differences for any of the features associated with the query. This indicates that such features become more informative after multiple queries. Interestingly, there were significant differences in the average rank associated with clicks, views, and bookmarks. Mouseover and scroll features were also significant.

Our level of interest and prior knowledge features were not based on interaction data, so their values are the same in both analyses. The observed trend is in the direction we expected (easier searches were associated with greater levels of interest and prior knowledge). However, the differences were not significant.

Figure 2 shows our classification results based on average precision (AP). The row labeled all corresponds to a model using all user-interaction features (excluding our level of interest and prior knowledge features). The rows labeled no.x correspond to models using all user-interaction features except for those in group x. The rows labeled inc.interest and inc.pk correspond to models using all user-interaction features plus our level of interest and level of prior knowledge features, respectively. Finally, the rows labeled only.x correspond to models using only those feature in group x.

As expected, using all user-interaction features (all), whole-session prediction was more effective than first-round prediction ($p < .05$). The whole-session model had access to a greater number of features and a greater number of whole-session features had significant differences between easy and difficult searches (Table 1).

In terms of whole-session evidence, in most cases, omitting a single feature type did not result in a significant drop in performance (see rows labeled no.x). Dwell time features were the only exception. Omitting dwell-time features resulted in an 8.49% drop in AP. This result suggests that our dwell-time features conveyed information not conveyed by the other features. That said, no single feature type on its own (including our dwell-time features) approached the performance of the model using all features (see rows labeled only.x). All models using a single feature type performed significantly worse. Taken together, these results suggest that given whole-session evidence, the best approach is to combine a wide range of features (including dwell-time features) in a single model.

In terms of first-round evidence, we see slightly different trends. As in the whole-session analysis, in most cases, omitting a single feature type did not result in a significant drop in performance (see rows labeled no.x). The largest drop in AP (7.77%) came from omitting bookmark features and not dwell-time features. Combined with the analysis in Table 1, this suggets that clicking and not bookmarking a page in the first round of results is highly predictive of task difficulty. In fact, using bookmark features alone (only.book) resulted in a 2.81% improvement over the model using all features.

Between mouse movement and scroll features, mouse movement features were more predictive of task difficulty. In terms of whole-session evidence, a model using only mouse features (only.mouse) performed only 5.41% worse than one using all features. In terms of first-round evidence, a model using only mouse features performed only 7.83% worse.

Table 1. Feature Analysis. Mean (STD) feature values for easy and difficult searches. A ▲(▼) denotes a significant increase(decrease) in the measure in difficult vs. easy searches ($p < .05$)

	Whole-Session Analysis		First-Round Analysis	
	easy	difficult	easy	difficult
Query Features				
NumQueries	1.810 (1.462)	2.373 (1.641)▲	-	-
AvgQueryLength	5.398 (2.980)	5.779 (3.702)	-	-
NumQueryTerms	9.073 (8.251)	12.448 (10.333)▲	5.415 (3.346)	5.772 (3.889)
UniqueQueryTerms	6.504 (3.666)	8.091 (5.039)▲	5.246 (2.999)	5.622 (3.549)
TokenTypeRatio	1.315 (0.628)	1.471 (0.590)▲	1.019 (0.068)	1.014 (0.044)
AvgStopwords	0.201 (0.212)	0.204 (0.196)	0.203 (0.225)	0.217 (0.225)
AvgNonStopwords	0.799 (0.212)	0.796 (0.196)	0.797 (0.225)	0.783 (0.225)
NumAOLQueries	0.286 (0.705)	0.295 (0.731)	0.165 (0.387)	0.112 (0.329)
NumQuestionQueries	0.286 (0.573)	0.336 (0.625)	0.216 (0.451)	0.224 (0.418)
Click Features				
NumClicks	3.263 (2.481)	4.618 (3.292)▲	2.289 (2.022)	2.527 (2.446)
AvgClicks	2.161 (1.739)	2.425 (2.033)	-	-
AvgClickRank	2.704 (1.737)	3.701 (3.517)▲	3.152 (2.645)	4.089 (3.819)▲
AvgTimeToFirstClick	8.613 (8.278)	8.351 (7.062)	48.425 (134.743)	63.253 (155.576)
NumViews	2.815 (2.055)	3.793 (2.623)▲	1.983 (1.644)	2.087 (1.980)
AvgViews	1.901 (1.507)	2.040 (1.703)	-	-
AvgViewRank	2.697 (1.795)	3.713 (3.499)▲	3.217 (2.756)	4.555 (3.988)▲
NumPageClicks	0.092 (0.450)	0.282 (0.937)▲	0.059 (0.381)	0.133 (0.724)▲
NumAbandon	0.294 (0.779)	0.378 (0.755)▲	0.132 (0.392)	0.149 (0.357)
PercentAbandon	0.078 (0.178)	0.106 (0.196)▲	-	-
Bookmark Features				
NumBook	2.336 (1.559)	2.722 (1.509)▲	1.681 (1.374)	1.531 (1.372)
AvgBook	1.620 (1.258)	1.548 (1.238)	-	-
AvgBookRank	2.713 (1.865)	3.900 (3.793)▲	3.425 (2.919)	4.971 (4.220)▲
NumQueriesWithBook	1.359 (0.790)	1.651 (0.905)▲	-	-
PercentQueriesWithBook	0.875 (0.229)	0.814 (0.257)▼	-	-
NumQueriesWithoutBook	0.451 (1.020)	0.722 (1.205)▲	-	-
PercentQueresWithoutBook	0.125 (0.229)	0.186 (0.257)▲	-	-
NumClicksWithoutBook	0.927 (1.521)	1.896 (2.821)▲	0.608 (1.237)	0.996 (1.721)▲
PercentClicksWithoutBook	0.184 (0.242)	0.275 (0.279)▲	-	-
NumViewsWithoutBook	0.479 (0.996)	1.071 (2.103)▲	0.303 (0.698)	0.556 (1.214)▲
PercentViewsWithoutBook	0.105 (0.193)	0.176 (0.253)▲	-	-
Mouse Features				
TotalMouseovers	23.039 (32.056)	42.602 (52.086)▲	15.602 (26.080)	22.494 (38.002)▲
AvgMouseovers	12.307 (13.160)	16.185 (15.026)▲	-	-
MaxMouseover	5.734 (5.229)	8.664 (7.845)▲	4.815 (4.889)	6.212 (6.268)▲
AvgMaxMouseovers	4.486 (3.346)	5.943 (4.432)▲	-	-
Scroll Features				
TotalScrollDistance	105.532 (161.087)	182.154 (244.690) ▲	64.636 (114.955)	91.699 (147.264)▲
AvgScrollDistance	55.118 (83.464)	64.382 (74.730)▲	-	-
MaxScrollPosition	39.067 (44.027)	53.610 (45.904)▲	29.528 (40.854)	39.013 (44.387)▲
AvgMaxScrollPosition	28.626 (36.012)	34.635 (35.586)▲	-	-
Dwell-Time Features				
TotalDwell	100.577 (112.695)	91.984 (105.488)	74.736 (95.838)	67.214 (105.407)▼
AvgDwell	42.998 (50.161)	29.351 (26.185)▼	42.009 (59.925)	31.458 (47.590)▼
Duration	193.596 (145.959)	223.964 (151.590)▲	140.766 (123.834)	130.235 (128.194)▼
Interest	2.838 (1.257)	2.635 (1.114)	2.838 (1.257)	2.635 (1.114)
Prior Knowledge Features				
PriorKnowledge	1.919 (0.937)	1.834 (0.845)	1.919 (0.937)	1.834 (0.845)
PriorSearch	1.437 (0.786)	1.378 (0.703)	1.437 (0.786)	1.378 (0.703)

Consistent with the analysis in Table 1, our level of interest and prior knowledge features were not highly effective for predicting search task difficulty. Including each feature type resulted in only a slight difference in performance (inc.interest and inc.pk) and neither feature set of its own approached the performance of a model using all features (only.interest and only.pk).

Table 2. Feature Ablation Analyses. A ▼ denotes a significant drop in performance compared to all ($p < .05$).

	Whole-Session Analysis	First-Round Analysis
all	0.618	0.563
no.query	0.616 (-0.39%)	0.576 (2.28%)
no.clicks	0.617 (-0.22%)	0.551 (-2.09%)
no.book	0.616 (-0.43%)	0.519 (-7.77%)▼
no.mouse	0.616 (-0.31%)	0.568 (0.85%)
no.scroll	0.625 (1.12%)	0.562 (-0.13%)
no.dwell	0.566 (-8.49%)▼	0.558 (-0.83%)
no.duration	0.622 (0.61%)	0.561 (-0.28%)
inc.interest	0.612 (-1.08%)	0.568 (0.90%)
inc.pk	0.613 (-0.83%)	0.554 (-1.62%)
only.query	0.547 (-11.47%)▼	0.516 (-8.28%)
only.clicks	0.576 (-6.81%)▼	0.528 (-6.23%)▼
only.book	0.582 (-5.83%)▼	0.579 (2.81%)
only.mouse	0.585 (-5.41%)▼	0.519 (-7.83%)
only.scroll	0.483 (-21.88%)▼	0.490 (-12.95%)▼
only.dwell	0.526 (-14.95%)▼	0.495 (-12.00%)▼
only.duration	0.501 (-18.98%)▼	0.513 (-8.78%)
only.interest	0.467 (-24.50%)▼	0.467 (-17.06%)▼
only.pk	0.479 (-22.45%)▼	0.479 (-14.81%)▼

7 Discussion and Conclusion

We evaluated different types of features for predicting search task difficulty at different points in the session: after the whole session and after the first round of interactions. Our results suggest that following trends. First, whole-session prediction was more effective than first-round prediction. While this may not be surprising, it is an important result because a major motivation for predicting search task difficulty is to develop search assistance interventions that would have to trigger *before* the end of the session. Second, for both whole-session and first-round prediction, the best approach is to combine a wide range of features. In our results, there were no cases where a single feature type significantly outperformed the model with all features. Third, the most predictive features were different in both analyses. Dwell-time features were the most predictive for whole-session prediction and bookmark features were the most predictive for first-round prediction. With respect to bookmarks, it is worth noting that existing search engines do not typically track bookmarks. This suggests the importance of capturing explicit relevance judgements or predicting relevance judgements implicitly for the purpose of first-round prediction. Fourth, mouse-movement features were more predictive than scroll features. For first-round prediction, a model using only mouse-movement features approached the performance of the model with

all features. Finally, including level of interest and prior knowledge features did not improve prediction performance.

In terms of future work, several open questions remain. Our experiment was conducted in a semi-controlled environment with simulated search tasks. Future work should consider predicting search task difficulty in a more naturalistic setting with user-initiated search tasks. In a real-world setting, the distribution of easy vs. difficult tasks may be highly skewed and user interaction signals are likely to be noisier. Additionally, overall our tasks were not found to be too difficult. It remains to be seen whether level of interest and prior knowledge features are predictive for highly difficult search tasks.

References

1. L. W. Anderson and D. R. Krathwohl. *A taxonomy for learning, teaching, and assessing: A revision of Blooms taxonomy of educational objectives*. 2001.
2. A. Aula, R. M. Khan, and Z. Guan. How does search behavior change as search becomes more difficult? In *CHI*, pages 35–44, 2010.
3. D. J. Bell and I. Ruthven. Searchers' assessments of task complexity for web searching. In *ECIR*, pages 57–71, 2004.
4. K. Byström and K. Järvelin. Task complexity affects information seeking and use. *Inf. Process. Manage.*, 31(2):191–213, 1995.
5. D. J. Campbell. Task complexity: A review and analysis. *The Academy of Management Review*, 13(1):40–52, 1988.
6. R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *JMLR*, 9:1871–1874, June 2008.
7. H. A. Feild, J. Allan, and R. Jones. Predicting searcher frustration. In *SIGIR*, pages 34–41, 2010.
8. B. J. Jansen, D. Booth, and B. Smith. Using the taxonomy of cognitive learning to model online searching. *Inf. Process. Manage.*, 45(6):643–663, Nov. 2009.
9. J. Kim. Task difficulty as a predictor and indicator of web searching interaction. In *CHI*, pages 959–964, 2006.
10. Y. Li and N. J. Belkin. A faceted approach to conceptualizing tasks in information seeking. *Inf. Process. Manage.*, 44(6):1822–1837, 2008.
11. J. Liu, M. J. Cole, C. Liu, R. Bierig, J. Gwizdka, N. J. Belkin, J. Zhang, and X. Zhang. Search behaviors in different task types. In *JCDL*, pages 69–78, 2010.
12. J. Liu, J. Gwizdka, C. Liu, and N. J. Belkin. Predicting task difficulty for different task types. In *ASIS&T*, pages 16:1–16:10, 2010.
13. J. Liu, C. Liu, M. Cole, N. J. Belkin, and X. Zhang. Exploring and predicting search task difficulty. In *CIKM*, pages 1313–1322, 2012.
14. J. Liu, C. Liu, J. Gwizdka, and N. J. Belkin. Can search systems detect users' task difficulty?: some behavioral signals. In *SIGIR*, pages 845–846, 2010.
15. M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *CIKM*, pages 623–632, 2007.
16. R. W. White and S. T. Dumais. Characterizing and predicting search engine switching behavior. In *CIKM*, pages 87–96, 2009.
17. R. E. Wood. Task complexity: Definition of the construct. *Organizational Behavior and Human Decision Processes*, 37(1):60–82, 1986.
18. W.-C. Wu, D. Kelly, A. Edwards, and J. Arguello. Grannies, tanning beds, tattoos and nascar: evaluation of search tasks with varying levels of cognitive complexity. In *IIIX*, pages 254–257, 2012.