Exploring and Predicting Search Task Difficulty

Jingjing Liu[1], Chang Liu[2], Michael Cole[2], Nicholas J. Belkin[2], Xiangmin Zhang[3]
1. School of Library and Information Science, University of South Carolina
2. School of Communication and Information, Rutgers University
3. School of Library and Information Science, Wayne State University
jingjing@mailbox.sc.edu, {changl, m.cole, belkin}@rutgers.edu, xiangminz@gmail.com

ABSTRACT
We report on an investigation of behavioral differences between users in difficult and easy search tasks. Behavioral factors that can be used in real-time to predict task difficulty are identified. User data was collected in a controlled lab experiment (n=38) where each participant completed four search tasks in the genomics domain. We looked at user behaviors that can be obtained by systems at three levels, distinguished by the time point when the measurements can be done. They are: 1) first-round level at the beginning of the search, 2) accumulated level during the search, and 3) whole-session level by the end of the search. Results show that a number of user behaviors at all three levels differed between easy and difficult tasks. Models predicting task difficulty at all three levels were developed and evaluated. A real-time model incorporating first-round and accumulated levels of behaviors (FA) had fairly good prediction performance (accuracy 83%; precision 88%), which is comparable with the model using the whole-session level behaviors which are not real-time (accuracy 75%; precision 92%). We also found that for efficiency purpose, using only a limited number of significant variables (FC_FA) can obtain a prediction accuracy of 75%, with a precision of 88%. Our findings can help search systems predict task difficulty and adapt search results to users.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *relevance feedback, search process*

General Terms
Performance, Measurement, Experimentation, Human Factors.

Keywords
User modeling, task difficulty, difficulty prediction, user behavior, first-round level, accumulated level, whole-session level

1. INTRODUCTION
Search engines perform adequately for simple search tasks, but are not successful for all search tasks. It is common to see that users have a hard time finding information for some "difficult" tasks. The ability to monitor the real-time search behaviors of users and observe when they are having difficulty can help the system determine if it is necessary to intervene and assist users. Such intervention could prevent users from becoming frustrated and/or switching to other systems.

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

There has been research that examines search difficulty, for example, by comparing users behavioral differences in easy and difficult tasks (e.g., [12], [6], [2]), and building task difficulty prediction models (e.g., [8], [15], [17]). Most of these approaches have not addressed the problem of building real-time prediction models. To our knowledge, the only attempt to use real-time behaviors to build difficulty prediction models was Liu, Gwizdka, Liu & Belkin [15]. They used in their models a set of within-session variables that can be captured by the system with the ongoing session before its completion. This work had limited real-time prediction success probably because of limitations on the number of within-session behaviors examined. In addition to examining the differences of user behaviors between easy and difficult tasks, the key questions for task difficulty predictions are: What user behaviors can be used by search systems to predict task difficulty in real-time? What significant user behaviors should be used in task difficulty prediction models? What is the best method to build task difficulty prediction models using observable (significant) user behaviors?

To address these questions, we conducted a laboratory user experiment and collected rich data from the client-side user system interaction logs as well as user self-assessed difficulty judgments after conducting the search tasks. This enabled analysis of the relationships between a comprehensive list of user behaviors and task difficulty and construction of task difficulty prediction models. The models developed through this analysis were then applied to prediction of task difficulty in an entirely different study of information seeking behavior, with positive results. Both studies were designed to simulate work task environments, so although the laboratory settings controlled for users, tasks, and environment, it is plausible these results can be generalized beyond the specific constraints of the two studies in which they were realized.

Our study has contributions in several aspects. First, we conducted an extremely extensive examination on the differences between user behaviors in easy and difficult tasks, at three levels, depending on when the behaviors can be captured: the first-round (early in a search), accumulated (during the search), and whole-session (after the search task) levels. Second, we constructed a number of models that can predict task difficulty using search behaviors, and found the best real-time prediction model has an overall accuracy of 83% and precision of 88%. Third, we showed that including only a limited number of significant behavioral variables in a real-time prediction model can receive comparable performance, with an overall accuracy of 75% and a precision of 88%, with using an extensive list of variables. Fourth, we found that significant behaviors for high accuracy real-time difficulty prediction can actually be captured at a quite early phase in a search session. These all shed light on information user behavior and interactive information retrieval (IR) research.

2. RELATED WORK

2.1 Search task difficulty and user behaviors

Much research attention has been attracted by search task difficulty. In their comprehensive task classification scheme, Li & Belkin [13] noted that task difficulty can only be subjective, as assessed by task doers. With a similar conceptualization of task difficulty as Li & Belkin [13], Kim [12] suggests that difficulty is the task doer's perception of task complexity, that it could be both pre- and post-task perceptions, and that task type is a variable in task difficulty. In a study to examine the effect of task difficulty on user behaviors, she used three types of tasks: factual, interpretive, and exploratory. Through a correlation examination of task difficulty and some user behaviors measured on the entire session level, it was found that in factual tasks, post-task difficulty was significantly associated with task completion time, and the numbers of queries and documents viewed; in exploratory tasks, user behaviors were significantly correlated with pre-task difficulty, but in interpretive tasks, most correlations between behaviors and task difficulty were not significant. The findings of this study pictured the relationship between task difficulty and users behaviors, but there was no prediction effort made, and it was not clear what factors can be used to build a model predicting task difficulty.

Aula, Khan, & Guan [2] also took the approach of post-task difficulty assessment, but it was determined by users' success or failure in finding the answers to their tasks, which were closed information tasks that have a single, unambiguous answer. They conducted a lab experiment with 23 participants to gain an understanding of users' behavioral changes when having difficulty, and then tested the observations using a large-scale study with 179 participants, each completing an average of 22.3 tasks in a pool of 100 tasks. Their studies found that in difficult tasks, users started by formulating more diverse queries, used advanced operators more, and spent longer time on the SERPs.

Liu, Liu, Gwizdka & Belkin [17] examined how user behaviors vary in tasks with different difficulty levels as well as of different types. They conducted a lab experiment with 48 participants, in which each worked on 6 out of a total of 12 tasks, in three different task types: single-item closed tasks, multiple-item closed tasks, and open-ended tasks. They looked at both the section-level and the task level behaviors; the former which can be tracked real-time and the later which can only be obtained after the search session. They found that in difficult tasks, users had longer task completion time, issued more queries, viewed more content pages, and had longer dwell time on content pages. These are good indicators of task difficulty, but they did not attempt to build prediction models using these behavioral indicators.

2.2 Task difficulty prediction

Although the current paper focuses on search task difficulty, as defined in the above subsection, it is helpful to note that there have been a number of studies in the IR literature that looked at query performance/difficulty prediction. Cronen-Townsend et al. [5] calculated the query clarify score based on the entropy between the language models of the query and the collection. Cramel et al. [3] calculated topic difficulty using the distances between three components of a topic: the query (the textual expression describing the information need), the Qrels (the set of documents relevant to the topic), and the entire collection of documents. Collins-Thompson & Bennett [4] used class-based statistics, instead of the entire content of top-ranked results, to compute measures of search result quality. Their empirical study findings showed that using class predictions, which reduce computing overhead, could offer comparable performance to full language models. These studies mainly attempted to predict query difficulty from the language model perspective.

The literature has also seen a continuing interest in the prediction of search task difficulty, which is more closely related to the effort of the current research. Gwizdka & Spence [6] examined how users' behaviors could indicate the difficulty of a factual information-seeking task. Task difficulty was self-assessed by users after each task. Their results indicated that higher search effort, lower navigational speed, and lower search efficiency were good predictors of task difficulty tested by regression models. One limitation of this study is that the predictive factors are not easily captured in real-time, and variables such as search efficiency cannot be obtained only after the search session is complete.

Liu, Gwizdka, Liu, & Belkin [15] examined relationships between search behaviors and tasks with different levels of difficulty and attempted to find some behavior variables to predict task difficulty. The tasks in the investigation were categorized as easy and difficult based on users' post-task judgment on tasks' difficulty levels. Users' behaviors were grouped at the whole-task-session level and the within-task-session level. Their study found that both whole-session level and within-session level user behaviors can serve as task difficulty predictors in logistic regression models. Whole-session level variables showed higher prediction accuracy, but do not enable real-time prediction. On the other hand, while within-session level factors can ensure real-time prediction, the prediction accuracy in general was mediocre, especially in some types of tasks, possibly because of the limited number of within-session factors that were considered and used in their model.

2.3 Other related user modeling in IR

In addition to the efforts made to predict task difficulty based on user behaviors, researchers in the IR community have been modeling users based on their search behaviors in other aspects. These include modeling of search success, frustration, satisfaction, and search engine switching behaviors, etc., by using different types of modeling techniques.

White & Dumais [20] examined search engine switching behaviors, and developed and evaluated predictive models of switching behavior using logistic regression. The study combined large-scale log-based analysis and survey data. Behaviors in the prediction model included the active query, the current session, and user search history. The study demonstrated the relationship between search engine switching and factors such as dissatisfaction with the quality of the results, the desire for broader topic coverage or verification of encountered information, and user preferences.

Feild, Allan, & Jones [9] extracted features from query logs and physical sensors in a controlled lab user study to build models of searcher frustration prediction using logistic regression. They found that the behavioral measures that were most useful for detecting frustration are the same as those White & Dumais [20] found most useful for detecting when a user switches search engines, including: the most recent query's length in characters, the average token length of the most recent query, the duration of the task in seconds, the number of user actions in the task, and the average number of URLs visited per task for the current user.

Ageev et al. [1] analyzed searcher success through user behaviors in a designed game consisting of 10 search tasks. They built their

prediction models using machine-learning methods. They found that more successful users issue more queries, clicks, and browse more pages for each question, issue shorter queries, and more actively use query reformulations and advanced query syntax. Also aimed at detecting users' success, but in a mobile search environment, Guo, Yuan, & Agichtein [6] used machine learning techniques to predict smart phone users' search success and satisfaction. They investigated client-side interaction signals, including the number of browsed pages, and touch screen-specific actions such as *zooming* and *sliding*. Their method resulted in nearly 80% accuracy for predicting searcher success which significantly outperformed the previous models.

Fox et al. [10] examined implicit behaviors for user satisfaction prediction using Bayesian modeling, decision trees, as well as a new usage behavior pattern analysis "gene analysis". They found an association between implicit measures of user activity and the user's explicit satisfaction ratings. The best models for individual pages include the behaviors: clickthrough, time spent on the search result page (SERP), and how a user exited a result or ended a search session. Behavioral patterns found through the gene analysis can be used to predict user satisfaction for search sessions.

Kotov et al. (2011) proposed methods for modeling and analyzing searchers' behaviors for multi-session search tasks. They attempted to identify, given a current query, what related queries were given in previous sessions, as well as to predict if the user will return to the task in the future. They adopted a machine learning methodology using different sets of behavioral features including: query-based (e.g., number of characters in a query, number of terms in a query), session-based (e.g., number of queries, number of clicks), history-based (e.g., number of sessions in the user's search history), and pair-wise features (e.g., number of overlapping terms between two queries). Evaluation showed that it is possible to effectively model and analyze cross-session search behavior.

3. USER EXPERIMENT

3.1 Participants

Our difficulty prediction models were developed using data from a study of 38 students from a U.S. research university conducting searches on selected topics from the TREC Genomics track task topics. Participants were drawn from medical- and health-related schools and departments, including biology, pharmacy, animal science, biochemistry, and so on. Their educational level ranged from undergraduates to graduate students and post-docs. Each was compensated $25 for the completion of the experiment.

3.2 Task topics

Task topics were selected from the 2004 TREC Genomics Track topic pool for our experiment. The topics were presented unchanged from the TREC Genomics Track and included the topic description, question, and context.

The topics were selected based on their difficulty levels, which were determined by retrieval performance using the topic titles as queries in our search system. In the experiment, we originally selected 4 topics: topics 2, 7, 45, and 49. About half-way through the study (after 19 participants completed the experiment), we found that topic 49 was too difficult for users (details below), so we replaced it with task 42. The task topics and the designed and user assessed difficulty levels are listed in Table 1.

Table 1. Task topics

TREC topic id	Topic title keywords	Designed difficulty level	User rated difficulty level	Mean of user ratings of task difficulty
2	Generating transgenic mice	Difficult	Difficult	4.53
7	DNA repair and oxidative stress	Easy	Easy	3.83
42	Genes altered by chromosome translocations	Easy	Easy	4.32
45	Mental Health Wellness-1	Difficult	Difficult	4.88
49	Glyphosate tolerance gene sequence	Easy	Difficult	5.24

Our tasks asked participants to find and save all of the documents useful for answering the topic questions. Each participant did four tasks. All of the participants completed tasks 2, 7, and 45. Nineteen participants did topic 42 and the other nineteen did task 49. The substitution of tasks is clearly a limitation of the experiment, but in this paper we do not distinguish between tasks in the analysis. The questions presented to the participants were the TREC genomics track descriptions including the need and context. An example of a task is shown in Table 2.

Table 2. User's task using Topic 45

Imagine you are gathering information for a class project on the following topic:

45 Mental Health Wellness-1

Need: What genetic loci, such as Mental Health Wellness 1 (MWH1) are implicated in mental health?

Context: Want to identify genes involved in mental disorders.

Please try to find and save all the articles on this topic. You will have up to 15 minutes to search on this task.

3.3 Search system

A search system was designed using the data set taken from the TREC Genomics collection, a ten-year, 4.5 million document subset of the MEDLINE bibliographic database [11]. To allow for reasonable retrieval efficiency, we used the documents from the 2000-2004 period (n=1.85 million). The system was implemented using Indri from the Lemur toolkit[1]. The system provides a web search interface in Internet Explorer (IE) 6.0. Figures 1 and 2 depict the search results page presented to the user and the abstract of a document (a document).

3.4 Procedure

Participants were invited individually to the site of the experiment, an on-campus information interaction lab. Each session lasted about 2.5 hours. After reading and signing a consent

[1] http://lemurproject.org

form, filling out a questionnaire about their background, and completing a self-assessment of their familiarity with selected MeSH terms that corresponded to the search topics categories, participants were given a brief demo using a training task about how to use the experiment system. Then they were given up to 15 minutes to conduct each of the four assigned tasks. They were asked to save as many documents as possible that helped to answer the task topic questions. Before and after each task topic, participants completed questionnaires about their self-rated perception of task difficulty. After each task, they rated the usefulness of the saved documents. After all 4 tasks, they completed an exit questionnaire about their final thoughts on the experiment. The interaction between the participants and the system was recorded by the logging software Morae[2].

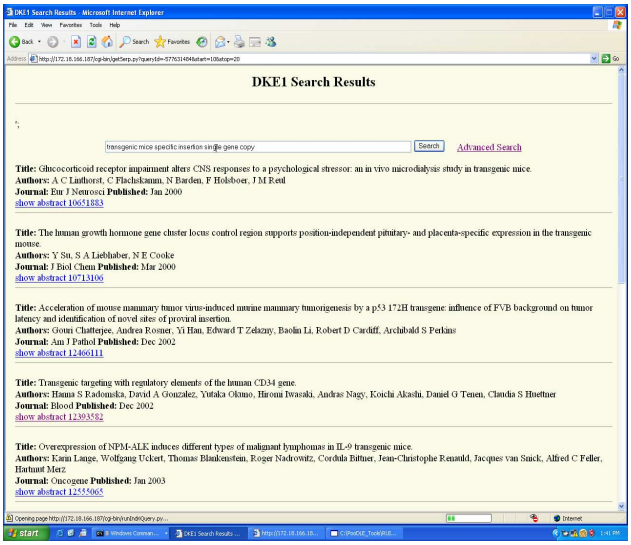

Figure 1. Search result interface

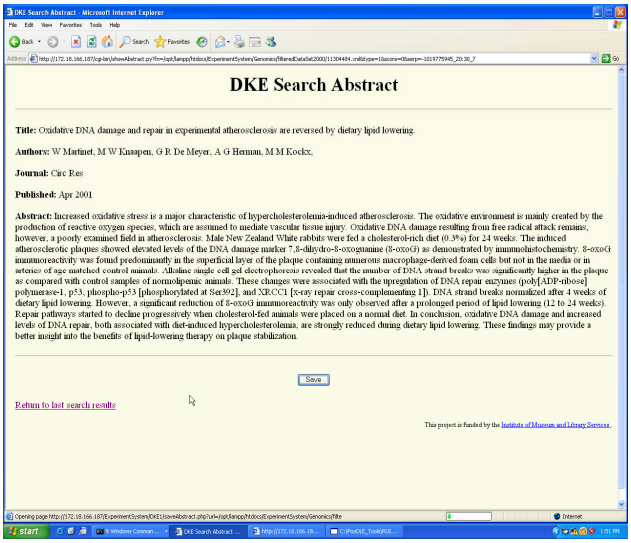

Figure 2. Document page interface

[2] http://www.techsmith.com/morae.html

4. VARIABLES

4.1 Difficulty measurements

There were three ways in our study to assess or control a task's difficulty level. The first way was by task design. As mentioned earlier, we chose task topics to include both easy and difficult topics in our study. However, task difficulty is clearly a subjective variable and the difficulty level of the same task could vary among different users. In comparison, we also elicited task difficulty in our study by asking users to self-rate how difficult they thought the tasks were both before and after working on them. The pre-task measure was what they expected, and the post-task measure was what we call "reflected". Between the two of them, the reflected measure seems a more accurate indicator of the users' actual perception of the difficulty of the task since it was made after their experience working on it. Therefore, post-task difficult rating was used in our analysis. The detailed measurement methods are described in Section 5 below.

4.2 Behavioral variables

We considered a comprehensive list of behavioral variables that were extracted from the logged user-system interaction data. These behaviors can be divided into three levels according to the time point when the variable can be observed by the system. The first two levels are within-session variables, which can be captured by the system during the search session, and the third level includes whole-session variables, which cannot be captured by the system until the completion of the search session.

The first level is called the "first-round level". It can be acquired early in a search session. These variables can, in principle, be obtained by the system in the initial sequence of the query-SERP-document process, with the end point being when the next query is issued. The following lists the variables and their definitions:

- *First query length*: the length of the first query in words
- *First dwell time on first SERP (seconds)*: the duration between the time when the user first viewed the first SERP to the time when he/she first left this SERP
- *First dwell time on first viewed document (seconds)*: the duration between the time when the user first viewed the first document to the time when he/she first left this document
- *Rank of first opened document*: the rank of the first document that the user opened in the whole search task session
- *Rank of first saved document*: the rank of the first document that the user saved in the whole search task session
- *Number of viewed documents at first query*: the total number of documents that the user clicked on and viewed following the first query issued
- *Number of saved documents at first query*: the total number of documents that the user saved following the first query issued
- *First query interval (seconds)*: the total duration after the first query is issued and before the second query (if any) is issued

The second level is called "accumulated level", which contains behavioral measures that can be calculated real-time *during* the search sessions, before the user has finished the search task. These measures include:

- *Mean dwell time of all documents (seconds)*: the average dwell time of all the visited documents in the session
- *Mean dwell time of unique documents (seconds)*: the average of the total dwell time of all unique documents visited in the session
- *Average first dwell time on documents (seconds)*: the average of durations between when a user opened a document and when the user first left the document
- *Mean dwell time of all SERPs (seconds)*: the average dwell time of all the SERPs in the session
- *Mean dwell time of unique SERPs (seconds)*: the average of the total dwell time of all unique SERPs in the session
- *Average first dwell time on SERPs (seconds)*: the average of durations between when a user opened a SERP and when the user first left the SERP
- *Average rank of viewed docs*: the average rank of all viewed documents
- *Average rank of saved docs*: the average rank of all saved documents
- *Number of documents per query*: average number of viewed documents per query
- *Number of unique documents per query*: average number of viewed unique documents per query
- *Number of saved documents per query*: average number of saved documents per query
- *Number of SERPs per query*: average number of SERPs visited per query
- *Number of unique SERPs per query*: average number of unique SERPs visited per query
- *Average query length*: the average length of all queries
- *Difference between first and average query length*
- *Average query interval (seconds)*: the average of all query intervals, from the time point after one query is issued and before the subsequent query is issued

The third level is called "whole-session level", which contains behavioral measures that can only be captured *after* the user has completed the entire search session. These measures include:

- *Task completion time (seconds)*: time users spent on each task
- *Numbers of all documents*: the number of all documents that the user viewed
- *Numbers of unique documents*: the number of unique documents that the user viewed
- *Number of saved documents in each session*
- *Number of SERPs*
- *Number of unique SERPs*
- *Number of queries*
- *Number of queries not leading to saved pages*: number of queries that were not followed by page saving before the next query was entered.
- *Number of queries leading to saved pages*: number of queries that were followed by page saving before the next query was entered.
- *Ratio of queries not leading to saved pages*: the ratio of the number of queries not leading to saving pages to the number of all queries in a task
- *Ratio of queries leading to saved pages*: the ratio of the number of queries leading to saving pages to the number of all queries in a task
- *Total time spent on documents (seconds)*: the sum of time users spent on all viewed documents
- *Total time spent on SERPs (seconds)*: the sum of time users spent on all SERPs
- *Ratio of document time to all*: the ratio of total time spent on documents to task completion time
- *Ratio of SERP time to all*: the ratio of total time spent on SERPs to task completion time
- *Total number of query terms (tokens)*: the sum of all tokens in all queries issued in the session
- *Total number of unique query terms*: the sum of unique tokens in all queries issued in the session
- *Total number of functional words*: the sum of stop words or non-meaningful words in all queries issued in the session
- *Total number of meaningful words*: the sum of non-stop words or functional words in all queries issued in the session
- *Total number of meaningful words from topic description*: the sum of meaningful words that appeared in topic descriptions
- *Total number of meaningful words not from topic description*: the sum of meaningful words that did not appear in topic descriptions

5. USER BEHAVIORS AND DIFFICULTY

As above mentioned, users' post-task self-assessed task difficulty scores were taken to measure the task difficulty level based on a 7-point scale, 1 - "not difficult", 4 - "somewhat difficult", and 7 - "extremely difficult". For analysis we collapsed the scores in the same way as has been done for document usefulness measurements (e.g., [19]). The mapping was based on both the distribution of difficulty scores and the meaning of the different rating points: *Difficult* for user ratings of 5-7; *Neutral* for a rated task difficulty of 4; *Easy* for task difficulty ratings of 1-3. Altogether, there were 117 user task sessions that were analyzed, with 36 easy tasks and 81 difficult tasks; the rest 35 were neutral tasks, which was not included in the analysis. The observed relationships between task difficulty level and three levels of user behaviors are reported below. In our analysis, we focused on task sessions that were rated as *Difficult* or *Easy* to compare the behavioral differences between them.

An exploration of the distribution of each behavioral variable in the difficult and easy tasks shows that the majority of them are not normal. Therefore, the non-parametric Mann-Whitney U test (a non-parametric statistical test assessing whether one of two samples of independent observations tends to have larger values than the other) was used to compare these variables between the two groups, as is reported in sections 5.1 to 5.3 below.

5.1 First-round level behaviors

The search behavioral measures on the first-round level between *Easy* and *Difficult* tasks were compared and Table 3 lists the results of the independent Mann-Whitney U Test.

Table 3. Comparison of behaviors on the first-round level

Behavioral measures on the first-round level	Task Difficulty level		Mann-Whitney U Test (p)
	Easy	Difficult	
First query length	3.81	3.42	0.15
First dwell time on first SERP (seconds)	**18.30**	**27.88**	**0.01**
First dwell time on first viewed document (seconds)	20.21	22.88	0.38
Rank of first opened document	2.72	2.83	0.16
Rank of first saved document	2.94	3.09	0.54
Number of viewed documents at first query	**4.00**	**1.47**	**0.01**
Number of saved documents at first query	**3.50**	**1.00**	**0.00**
First query interval (seconds)	159.33	84.39	0.17

Table 4. Comparison of behaviors on the accumulated level

Behavioral measures on the accumulated level	Task Difficulty level		Mann-Whitney U Test (p)
	Easy	Difficult	
Mean dwell time of all documents (seconds)	16.75	16.51	0.68
Mean dwell time of unique documents (seconds)	19.14	20.32	0.34
Average first dwell time on documents (seconds)	17.75	17.64	0.62
Mean dwell time of all SERPs (seconds)	**16.2**	**19.05**	**0.01**
Mean dwell time of unique SERPs (seconds)	48.99	42.62	0.17
Average first dwell time on SERPs (seconds)	**17.22**	**20.64**	**0.02**
Average rank of viewed documents	9.29	7.42	0.07
Average rank of saved documents	**9.37**	**7.44**	**0.03**
Number of documents per query	**3.86**	**2.31**	**0.00**
Number of unique documents per query	**3.45**	**1.99**	**0.00**
Number of saved documents per query	**3.19**	**1.48**	**0.00**
Number of SERPs per query	**5.76**	**3.96**	**0.00**
Number of unique SERPs per query	**1.96**	**1.79**	**0.05**
Average query length	4.42	4.05	0.17
Difference between first and average query length	-0.61	-0.63	0.89
Average query interval (seconds)	**164.98**	**109.8**	**0.00**

From Table 3, three of the behavioral measures showed significant differences between difficult and easy tasks. Users spent significantly more time on the first SERP in difficult tasks (27.88 seconds) as compared to easy tasks (18.3 seconds). They also viewed significantly fewer documents and saved significantly fewer documents during the first query in difficult tasks (1.47 and 1) than in easy tasks (4 and 3.5). In contrast, first query length, dwell time on first clicked documents, the rank of the first clicked or saved document, and first query interval were not significantly different between the difficult and easy tasks.

5.2 Accumulated level behaviors

Table 4 presents the results of search behavior differences at the accumulated level of analysis between easy and difficult tasks. The analysis shows that some measures are significantly different between easy and difficult tasks. Users spent significantly more time on SERPs they visited in difficult tasks (19.05 seconds) than in easy tasks (16.2 seconds). Users also spent significantly more time on each unique SERP they visited in difficult tasks (20.64 seconds) as compared to easy tasks (17.22 seconds). No significant difference was found for dwell time on clicked documents in difficult and easy tasks. The average rank of saved documents in difficult tasks (7.44) was significantly higher compared to easy tasks (9.37). We also found that users visited fewer documents and fewer unique documents per query in difficult tasks (2.31 and 1.99) compared to easy tasks (3.86 and 3.45). Users saved fewer documents per query in difficult tasks (1.48) vs. easy tasks (3.19). With respect to the search result page, users visited far fewer SERPs per query in difficult tasks (3.96) than in easy tasks (5.79), and they also visited fewer unique SERPs in difficult tasks (1.79) than in easy tasks (1.96). Users also issued queries more frequently in difficult tasks (109.8 seconds) than in easy tasks (164.98 seconds).

5.3 Whole-session level behaviors

Table 5 reports the means and the p-value of the Mann-Whitney U Test on the whole-session level behaviors. Seven out of twenty-one behavioral measures had significant differences between difficult and easy tasks. Specifically, users saved far fewer documents in each session for easy tasks (8.92 than difficult tasks (5.19). Users visited many more unique SERPs in difficult tasks (9.38) than in easy tasks (6.97). Considering user query behavior, we found users issued many more queries in difficult tasks (6.54) than in easy tasks (3.83). More queries failed to result in saved pages for difficult tasks (4.28) vs. easy tasks (1.67). For difficult tasks users had a higher percent of queries that did not lead to saved pages (54%) than in easy tasks (29%). Finally, users spent a lower percent of all task completion time on documents in difficult tasks (20%) than in easy tasks (26%).

Even though the mean of task completion time in difficult tasks was longer than in easy tasks, the difference was not significant. Other behavioral measures related to features of query terms were examined, including the number of query terms, number of unique terms, and number of meaningful terms. These measures did not show significant differences between difficult and easy tasks.

6. DIFFICULTY PREDICTION

6.1 Eight models

The previous section showed users' behavioral differences between easy and difficult tasks. In this section, we report modeling of task difficulty. The goal of the prediction task was to predict if a task session was difficult given observable user behaviors. We used logistic regression to build binary classification prediction models for task difficulty, i.e., a task session being easy or difficult, where the labels followed the classification method that we used in Section 5, i.e., ratings 1-3 as *Easy* and 5-7 as *Difficult*. Four models were constructed:

Table 5. Comparison of behaviors on the whole-session level

Whole-session variables	Task Difficulty level		Mann-Whitney U Test (p)
	Easy	Difficult	
Task completion time (seconds)	623.69	725.64	0.07
Numbers of all documents	10.92	9.21	0.14
Numbers of unique documents	9.78	7.70	0.05
Number of saved documents	**8.92**	**5.19**	**0.00**
Number of SERPs	17.86	18.73	0.52
Number of unique SERPs	**6.97**	**9.38**	**0.03**
Number of queries	**3.83**	**6.54**	**0.00**
Number of queries not leading to saving pages	**1.67**	**4.28**	**0.00**
Number of queries leading to saving pages	2.17	2.26	0.86
Ratio of queries not leading to saving pages	**0.29**	**0.54**	**0.00**
Ratio of queries leading to saving pages	**0.71**	**0.46**	**0.00**
Total time spent on documents (seconds)	159.15	144.22	0.72
Total time spent on SERPs (seconds)	276.49	338.14	0.06
Ratio of document reading time to all	**0.26**	**0.20**	**0.02**
Ratio of SERP time to all	0.43	0.46	0.19
Total number of query terms (tokens)	8.67	8.91	0.85
Total number of unique query terms	5.75	5.8	0.98
Total number of functional words	0.56	0.59	0.83
Total number of meaningful words	5.19	5.21	0.95
Total number of meaningful words from topic description	3.89	4.19	0.43
Total number of meaningful words not from topic description	1.31	1.02	0.13

- First-round level variables (FR)
- Accumulated level variables (AC)
- First-round and accumulated variables (FA)
- Whole-session level variables (WS)

Three of these models use variables at the three levels, and another one was the combination of first-run and accumulated levels. Among the four, 3 could be real-time (FR, AC, and FA), and the non-real time WS model was used as a comparison to the others.

These models used all variables that were listed in the corresponding levels, so we call them plain models. Using all of the variables could be costly in computing. Therefore, we attempted to select a smaller set of significant variables that can ensure the prediction performance not much worse than using all variables. The selection was conducted using the Forward Conditional (FC) option in logistic regression, which includes only the statistically significant variables in the model. We randomly selected 10 samples from the dataset, each being 80% of all data, and ran Forward Conditional logistic regression to obtain significant variables in each model that appeared at least once in the 10 runs. This gave us another set of 4 models, named FC models. The variables are listed below:

- First-round level variables (FC_FR): first dwell time on first SERP, number of saved documents at first query, first query interval, number of viewed documents at first query.
- Accumulated level variables (FC_AC): average first dwell time on SERP, pages per query, average query interval, number of saved documents per query, average rank of saved documents.
- First-round and accumulated variables (FC_FA): variables in FC_FR and FC_AC
- Whole-session level variables (FC_WS): number of saved documents, number of query leading to saved documents, ratio of query not leading to saved documents to all, number of query, number of meaningful terms not in topic, number of meaningful terms from topic, number of content pages, ratio of content page time to all.

6.2 Model evaluation

We tested the models used an 80/20 cross validation method. We randomly selected 80% of the data as the training set and the remaining 20% as the test set. For each of the 8 models, we built the models in the training set, and then tested the models for precision and recall values in the test set. Figures 3 & 4 show their performance.

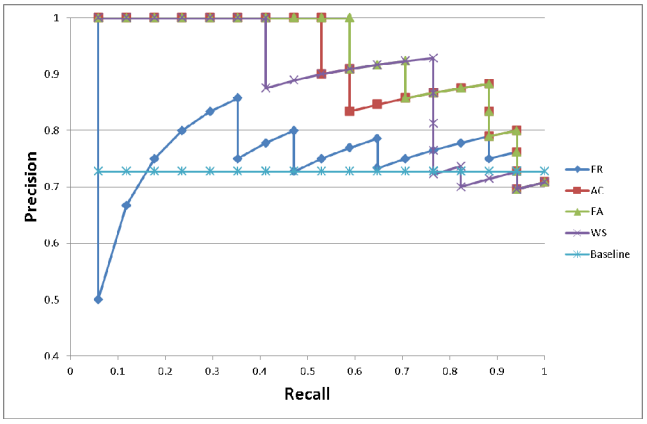

Figure 3. Plain models recall-precision graph

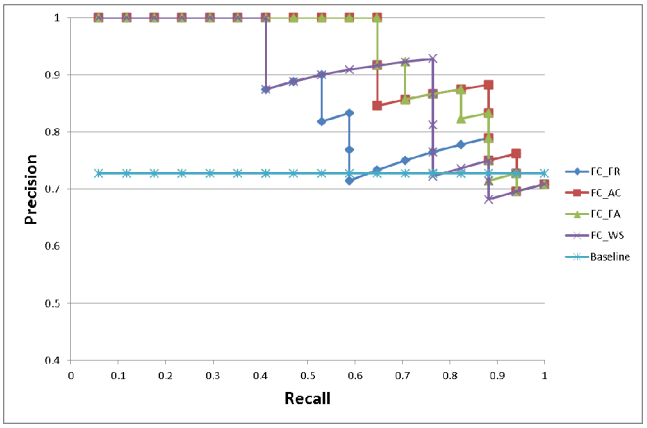

Figure 4. FC models recall-precision graph

Table 6. Model overall accuracy, precision, and F(0.5) (Those in bold are the best model value in the examined aspects)

Models	overall accuracy	precision	F(0.5)
BL	0.71	0.71	0.75
FR	0.75	0.79	0.81
AC	0.79	0.88	0.86
FA	**0.83**	0.88	**0.88**
WS	0.75	**0.92**	0.87
FC_FR	0.75	0.79	0.81
FC_AC	0.75	0.87	0.84
FC_FA	**0.79**	0.88	0.86
FC_WS	**0.79**	**0.93**	**0.89**

Table 7. Variables in the plain FA model (B and p values) (Those in bold are statistically significant variables)

	B (weight)	p
(Intercept)	-2.36	0.30
meanDwellTimeAllContentPage	0.23	0.32
meanDwellTimeUniqueContentPage	0.17	0.39
averageFirstDwellTimeContentPage	-0.45	0.22
meanDwellTimeAllSERP	-0.14	0.33
meanDwellTimeUniqueSERP	**0.11**	**0.04**
averageFirstDwellTimeSERP	0.11	0.18
averageRankViewedDocs	0.19	0.31
averageSavedDocRank	-0.11	0.43
pagesPerQuery	1.11	0.55
noUniqueContentPagesPerQuery	2.33	0.20
noSavedDocPerQuery	-0.95	0.26
noSerpPerQuery	-2.67	0.19
noUniqueSerpsPerQuery	4.23	0.06
averageQueryLength	0.23	0.48
Diff_First_AverageQueryLength	0.40	0.22
averageQueryInterval	**-0.04**	**0.04**
firstDwellTimeFirstSerp	0.01	0.79
firstDwellTimeFirstDoc	0.04	0.19
rankOfFirstViewedDocs	-0.20	0.24
rankOfFirstSavedDocs	0.14	0.48
noViewedDocFirstQuery	0.63	0.43
noSavedDocFirstQuery	-0.87	0.24
firstQueryInterval	0.00	0.68

We also looked at the overall accuracy, precision, and F (B=0.5) scores of each model, as shown in Table 6. As can be seen, the baseline model (BL) predicts all tasks as difficult, and its accuracy was 71% (i.e., the number of difficult task sessions in the test set, 17, divided by all user task sessions, 24). Figures 3 and 4 show that all models were better than or comparable to the BS. For the plain models, AC and FA had comparable performance with WS. For the FC models, FC_FA also had comparable performance with FC_WS.

6.3 Real-time prediction models

Although WS had good prediction performance, the behavioral variables in these models can only be obtained at the end of the search session. In the plain models, FA appeared to be quite good which can also be used in real-time to predict task difficulty. The behaviors included in the model with their p and B values are listed in Table 7.

For the FC models, again, the FC_FA showed quite good performance. The behaviors included in the model with their p and B values are listed in Table 8.

Table 8. Variables in the FC_FA model (B & p values)

	B (weight)	p
(Intercept)	1.04	0.29
firstDwellTimeFirstSerp	0.00	0.87
noSavedDocFirstQuery	-0.20	0.68
firstQueryInterval	0.00	0.76
noViewedDocFirstQuery	0.08	0.88
averageFirstDwellTimeSERP	0.06	0.22
pagesPerQuery	1.08	0.11
averageQueryInterval	-0.02	0.14
noSavedDocPerQuery	-1.21	0.06
averageSavedDocRank	0.02	0.69

7. DISCUSSION

7.1 First-round vs. accumulated level behaviors

We detected a number of variables that showed significant differences between easy and difficult tasks, as reported in the results section. Here we focus on the comparison between variables at the first-round and accumulated levels, especially those related ones.

On the first-round level, three variables showed significant differences between easy and difficult tasks:

- first dwell time at first SERP,
- number of viewed documents at first query, and
- number of saved documents at first query.

It is reasonable that if users spent more time reading SERPs before their first click on a content page, and viewed and saved fewer documents before re-issuing new queries, they would be more likely to be finding the task to be difficult.

On the accumulated level, the variables related to the three significant ones in the first-round level also showed significant differences between easy and difficult tasks. They were:

- average first dwell time on SERPs,
- number of documents per query,
- number of unique documents per query, and
- number of saved documents per query.

Meanwhile, there were two accumulated level variables appearing to be significant factors, whose related first-round variables, however, did not show significant differences. They were:

- average rank of saved documents (rank of first saved document at first-round level), and
- average query interval (first query interval at first-round level).

This indicates that the two first-round variables had a tendency to show significant differences between easy and difficult tasks, although in the beginning of the search, they did not appear to be.

Further, there were some variables that did not show significant differences between easy and tasks in both the first-round and the accumulated level. They were:

- average query length (first query length at the first-round level),
- differences between first and average query length (first query length at the first-round level),
- average first dwell time on documents (first dwell time on first viewed document at the first-round level), and
- average rank of viewed documents (rank of first opened document at the first-round level).

This indicates that these variables are not significant factors that would show differences in easy and difficult tasks.

At the accumulated level, the following variables were significant in predicting difficulty:

- mean dwell time of all SERPs
- number of SERPs per query
- number of unique SERPs per query

The following lists the accumulated level variables that did not show significant differences between easy and difficult tasks:

- mean dwell time of all documents
- mean dwell time of unique documents
- mean dwell time of unique SERPs

7.2 Whole-session level significant behaviors

Unlike the above variables, most of which are rarely examined for the relationship with task difficulty, many whole-session level variables have been examined in previous studies. Many of the significant behaviors showing differences in easy and difficult tasks in the current study had consistent patterns with findings in previous studies, including number of queries (also in [2][12][15][17]), number of unique SERPs, number of queries not leading to saved documents, ratio of queries not leading to saved documents, and ratio of queries leading to saved documents as in [15]. In addition, two other variables, number of saved documents, and ratio of total time on document reading to all also showed differences in easy and difficult tasks.

7.3 Difficulty prediction models

Several points are worth discussing with regard to the models that were built. First, the models including more variables tended to outperform those having fewer variables. Results showed that the models incorporating variables of both the first-round and the accumulated levels outperform those using only single level behaviors, whether it is a plain or an FC model. This observation tells us that the more user behaviors are measured and used in the model, the more accurate the prediction is. This is similar to the findings in previous studies that made predictions based on implicit user behaviors, for example, Fox et al. [10], which found that using a combination of implicit measures could better predict user satisfaction than by just using the base rate of satisfaction with sessions.

Second, our results indicated that it is possible that using a limited number of significant variables in the model can obtain comparable prediction performance as using an extensive list of variables. This is practically helpful in system design, considering that it may not be easy sometimes to monitor an extensive amount of user behaviors, or incorporating too many user behaviors in the model may decrease the system's efficiency. Specifically, our results showed that the FC models (using a small number of variables) received comparable precision and F scores as plain models (using a long list of variables). The FC_FR model using only 4 variables captured early in the search process provided a difficulty prediction with an overall accuracy of 75% and a precision of 79%. Using more within-session (both first-round and accumulated levels) variables in the model could reach a prediction accuracy of 79% and a precision of up to 88%. As White & Dumais [20] noted, the goal of their study was "not to optimize the model but rather to determine the predictive value of the query/session/user feature classes for the switch prediction challenge" (p. 93). Our models did well in determining the predictive value at three levels of behavioral measures, clearly showing the correct pattern of task difficulty levels.

Further, our results also show that some significant within-session variables that can play important roles in task difficulty predictions models do not necessarily need to be obtained later in a session; instead, they can be captured quite early in the search process. Some variables in the FC_FA model, i.e., number of saved documents at first query, first dwell time at first SERP, and first query interval, are all first-round level variables. This indicated that users' behaviors in the early search phase can predict difficulty of the task fairly well. For instance, if the user saves few documents at the first query, spends a short time at first SERP before clicking on any result, and issues the second query quickly, it is very likely that this user would judge this task as difficult, and what happens later in the search would play less roles instead.

Another note is that the significant variables included in the prediction models were not necessarily the same as those showing significant differences between easy and difficult tasks (as reported in Section 5). The reason could be that some variables correlated with others, given these others had stronger power, may be shadowed in the logistic regression prediction model. For example, as mentioned above, in the plain FA model, there were only two variables showing statistical significance (with a p value less than 0.05), but there were much more variables having statistical significance in the behavioral difference comparisons (results as reported in Section 5.2 and Table 4).

Although lab experiments are limited by the controlled nature in user, task, and environment, and the size of the data, we think our modeling method and the significant behaviors detected to be included in the models make good contributions to predicting task difficulty in real system application. Future studies will test these models in the real search environment settings.

8. CONCLUSIONS

Through a controlled lab experiment, we examined the differences in users search behaviors between easy and difficult tasks. We grouped user behaviors into three levels according to the time point when the behavior measures can be obtained: 1) first-round level at the beginning of the search, 2) accumulated level in the search session, and 3) whole-session level by the end of the search.

Results show that a number of user behaviors at all three levels differed between easy and difficult tasks. Some of the first-round behaviors had the same patterns as their related ones in the accumulated level, such as the first dwell time at first SERP (average first dwell time of all SERPs), number of viewed documents at first query (average number of viewed documents per query), and number of saved documents at first query (average number of saved documents per query). These are significant behavioral variables that show differences continuously along the search session. On the whole-session level, significant behaviors showing differences in the easy and difficult tasks included number of saved documents, number of unique SERPs, number of queries, number of saved documents, number of queries not leading to saved documents, ratio of queries not leading to saved documents out of all, and ratio of total time on document reading to all.

With regard to the difficulty prediction, a real-time plain model incorporating both first-round and the accumulated levels of behaviors (FA) received fairly good prediction performance (accuracy 79%; precision 88%), which was comparable with the model using the whole-session level behaviors which were not real-time (accuracy 75%; precision 92%). We also found that for efficiency purpose, using only a limited number of significant variables (FC_FA) can obtain a prediction accuracy of 79%, with a precision of 88%. Despite that this is a controlled lab experiment, our findings can help search systems predict task difficulty and adapt search results to users.

9. ACKNOWLEDGMENTS

This research was sponsored by IMLS grant LG#06-07-0105-05.

10. REFERENCES

[1] Ageev, M., Guo, Q., Lagun, D., & Agichtein, E. (2011). Find it if you can: A game for modeling different types of web search success using interaction data. *SIGIR '11*, 345-354.

[2] Aula, A., Khan, R. & Guan, Z. (2010). How does search behavior change as search becomes more difficult? Proceedings of *CHI '10*, 35-44.

[3] Carmel, D., Yom-Tov, E., Darlow, A., & Pelleg, D. (2006). What makes a query difficult? *SIGIR '06.*

[4] Collins-Thompson, K., & Bennett, P. N. (2010). Predicting query performance via classification. *ECIR '10.*

[5] Cronen-Townsend, S., Zhou, Y., & Croft, W. B. (2002). Predicting query performance. *SIGIR '02.*

[6] Guo, Q., Yuan, S., & Agichtein, E. (2011). Detecting success in mobile search from interaction. *SIGIR '11*, 1229-1230.

[7] Gwizdka, J., Spence, I. (2006). What can searching behavior tell us about the difficulty of information tasks? A study of Web navigation. *ASIST '06*.

[8] Gwizdka, J. (2008). Revisiting search task difficulty: Behavioral and individual difference measures. *ASIST '08*.

[9] Feild, H., Allan, J., & Jones, R. (2010). Predicting searcher frustration. *SIGIR '10*, 34-41.

[10] Fox, S., Karnawat, K., Mydland, M., Dumais, S., & White, T. (2005). Evaluating implicit measures to improve Web search. *ACM TOIS, 23*(2), 147-168.

[11] Hersh, W. & Voorhees, E. (2009). TREC genomics special issue overview. *Information Retrieval, 12*(1), 1–15.

[12] Kim, J. (2006). Task difficulty as a predictor and indicator of web searching interaction. *CHI '06*, 959-964.

[13] Kotov, A., Bennett, P. N., White, R. W., Sumais, S. T., & Teevan, J. (2011). Modeling and analysis of cross-session search tasks. *SIGIR '11*, 5-14.

[14] Li, Y. & Belkin, N. J. (2008). A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management, 44*, 1822-1837.

[15] Liu, J., Cole, M., Liu, C., Bierig, R., Gwizdka, J., Belkin, N.J., Zhang, J., & Zhang, X. (2010). Search behaviors in different task types. *JCDL '10*.

[16] Liu, J., Gwizdka, J., Liu, C., & Belkin, N. J. (2010). Predicting task difficulty for different task types. *ASIST '10*.

[17] Liu, J., Liu, C., Gwizdka, J., & Belkin, N. (2010). Can search systems detect users' task difficulty? Some behavioral signals. *SIGIR '10*.

[18] Roberts, P. M., Cohen, A. M., and Hersh, W. R. (2009). Tasks, topics and relevance judging for the TREC genomics track: Five years of experience evaluating biomedical text information retrieval systems. *Information Retrieval, 12*(1), 81-97.

[19] White, R., & Kelly, D. (2006). A study of the effects of personalization and task information on implicit feedback performance. *CIKM '06*, 297-306.

[20] White, R. W. and Dumais, S. T. (2009). Characterizing and predicting search engine switching behavior. *CIKM '09*, 87-96.