

Proactive Identification of Query Failure

Jiqun Liu

School of Communication & Information, Rutgers
University, New Brunswick, NJ
jl2033@scarletmail.rutgers.edu

Chirag Shah

School of Communication & Information, Rutgers
University, New Brunswick, NJ
chirags@rutgers.edu

ABSTRACT

When a user fails to find any useful information to support the task at hand after issuing a query, the user experiences a *query failure*. Since users possess limited cognitive resources, query failures often lead to user frustration as no clear benefit is obtained from the associated search interactions. Therefore, to improve users' search experiences, we conducted a controlled-lab study with 40 participants, seeking to explore the extent to which query failures can be proactively identified *before users start examining the retrieved results*. Specifically, based on the data collected from 693 query segments generated in 80 search sessions, we used past search behaviors and current query attributes as features to build classifiers and examined the performance in capturing query failures. We report that (1) analytics algorithms utilizing past search behavioral data have significantly better performances than the baseline model in tasks of different types, and (2) The knowledge of users' search intentions can help improve the performance of the prediction model. Results pave way for developing proactive system supports for task-based search interactions.

KEYWORDS

Query Failure; Proactive Information Retrieval; Search Intention; Task-Based Information Searching

ASIS&T THESAURUS

Information Seeking; Information Retrieval; Search Behavior

INTRODUCTION

Search interaction is a major component of information-intensive tasks. Users often go through multiple rounds of search interactions in complex tasks where the needed item(s) and dimension(s) for result evaluation are not clearly defined (Liu, 2017; Liu, Mitsui, Belkin, & Shah, 2019). When a user fails to find any useful information to support the task at hand after issuing a query, the user experiences a *query failure*, which often lead to search frustration (Feild, Allan, & Jones, 2010; Smith & Kantor, 2008). Hence, to help users save their limited cognitive resources and improve their search efficiency

and experience, search systems need to *proactively* capture query failures and provide useful interventions *before* users allocate time and energy to reviewing the results presented on the associated search engine result pages (SERPs).

In the past interactive information retrieval (IIR) studies on query quality evaluation, most researchers sought to infer query performance and capture low-performing queries based on the associated search behavioral features (e.g., Guo, White, Dumais, Wang, & Anderson, 2010; Kumaran & Carvalho, 2009; Wu, Kelly, & Huang, 2012). Although this approach can enhance our understanding on the connection between query performance and user behavior, the associated models were built upon the search interactions that happened *after* the target query was issued. Hence, this inference approach leaves no room for the system to help users by saving their time and energy in the current query segment. Given this problem, our study sought to provide “preventive care” for search system users by developing predictive methods to proactively identify query failures (see Figure 1).

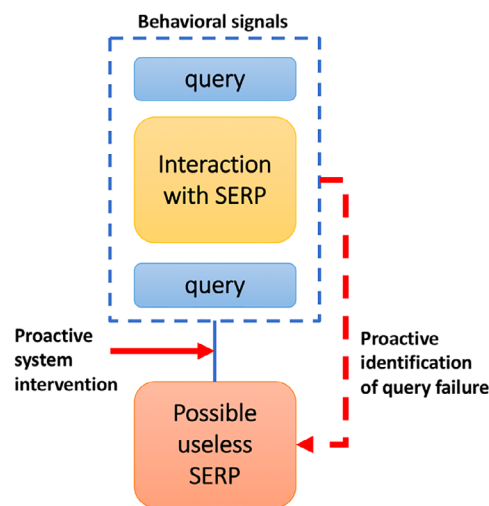


Figure 1. Proactive identification of query failure.

With respect to the operational definition, this study considers two states of query: (1) *useful query*: one or more useful documents can be identified in the associated SERP, and (2) *query failure*: no useful document is identified in the associated SERP. Based on these definitions, we adopted a machine learning approach and built classifiers to predict

query failure from the search interactions extracted from the previous query segment. Specifically, this study sought to answer the following research questions (RQs):

RQ1: How is a user's search interaction associated with the subsequent query state (i.e. useful query or query failure)?

RQ2: To what extent can query failure be predicted from past search behaviors?

RQ3: To what extent can we leverage the knowledge of users' query-level search intentions in predicting query failures?

By addressing the RQs above, this study has multiple contributions and implications:

- Our results reveal the underlying connection between past search interactions and the quality or performance of the subsequent query in task-based search process.
- The query failure prediction model is a critical step toward achieving comprehensive proactive supports and recommendations in interactive information retrieval.
- Methodologically, the study procedure, behavioral features and techniques employed in this study can be reused in examining and predicting other dimensions of search interactions (e.g., user interest and expectation).

RELATED WORKS

Corresponding to the RQ1 and RQ2, this section first reviews the findings from previous works on inferring and evaluating query quality. Then, to test the usefulness of search intention feature in different contexts (RQ3), this section discusses the existing research progresses made in characterizing tasks and search intentions in the context of information searching.

Inferring Query Performance

Among the IIR studies focusing on query analysis and evaluation, some of them sought to infer query quality and performance based on users' interactions with the SERP and content pages retrieved from the target query. For instance, He and Ounis (2006) focused on a series of query features in performance prediction and their TREC-based experiments indicated that many of the query-based predictors have significant association with query performance measured by the average precision of document retrieval. Similarly, Hauff, Hiemstra, and de Jong (2008) identified 22 pre-retrieval predictors based solely on query terms and assessed the prediction performance using three TREC test collections. Kumaran and Carvalho (2009) took a step forward by combining query features and the attributes of retrieved documents (i.e. IDF-based features) in building models for query quality prediction. Guo et al. (2010) used features of the query, retrieved search results, and user behavioral features (e.g., result browsing, dwell time on pages) to infer query performance. Their findings demonstrated that using interaction signals can significantly improve the performance of predicting query performance. Wang, Liu, Mandal, and Shah (2017) conducted a field study on query-

level search failures and stated that users' pages visiting and dwell time on pages can significantly affect their query-level search outcomes. To understand query quality evaluation from the user side, Wu et al. (2012) conducted a controlled-lab user study, aiming to figure out how users evaluate query quality based on query features as well as their own interactions with the SERPs. They found that for exploratory search tasks, users' ratings of query quality were highly correlated with the number of relevant documents and the proportion of relevant documents viewed. Similar to previous works, Wu et al. (2012) also confirmed the importance of including search interaction features in query performance prediction.

Despite of the contributions of the previous studies discussed above, they still have limitations as they either (1) only consider query attributes, or (2) seek to infer query quality from the associated search interactions (after the retrieval step), leaving no room for proactive system interventions (although in many of these cases, the authors used the term *prediction* to label their works on *inferring* query performance). In addition, the impacts of cognitive-level search intentions and task type on query failure prediction still remains ambiguous.

Given the interactive nature of Web search, different query segments are likely to be connected within the same task context (Liu, Mitsui et al., 2019). The multi-dimensional associations between adjacent query segments can create valuable opportunities for capturing query failure from past search behaviors. Hence, to provide proactive supports for users, search systems need to combine past search behavior features with current query attributes in predicting the query performance *before the retrieval step*. The proactive identification of query failure can push us closer toward the vision of proactive information retrieval (Shah, 2018).

Task and Search Intention

The existing IIR research have shown that many facets of users' search interaction vary significantly across tasks of different types (Jiang, He, & Allan, 2014; Liu, Cole et al., 2010). Thus, the predictability of query failure in information search process may also change according to multiple task features.

IIR researchers have developed a series of frameworks and typologies to define task types. For instance, Capra, Arguello, O'Brien, Li, and Choi (2018) focused on task prior determinability (i.e. the level of uncertainty about task outcomes and processes) and argued that the variations in needed items and the clarity of dimensions for result evaluation can significantly affect the overall task determinability. Kelly, Arguello, Edwards, and Wu (2015) examined the cognitive complexity of task and adopted Bloom (1956)'s taxonomy of learning domains in characterizing different levels of task complexity. Li and Belkin (2008) took a step back and looked at the bigger picture of task. Specifically, instead of merely focusing on one aspect of task, they developed a faceted approach to characterize multiple dimensions of tasks (e.g., task product, task goal, task complexity) and tested the effects of multiple facets on

users' search behavior (e.g., Liu et al., 2010). This multi-faceted scheme allows researchers to characterize tasks over multiple dimensions and conceptually supports the IIR studies focusing on how task traits affect users' cognitive states and search behaviors.

In addition to the global task context, users' search patterns are also affected by their changing cognitive spaces (Ingwersen, 1996). In some recent IIR studies, researchers have investigated the roles of query-level search intentions in local query segments. In these studies, a query segment is defined as a search session segment which starts from one issued query and ends at the next query (Mitsui, Liu, Belkin, & Shah, 2017). Ruotsalo et al. (2018) represented users' interactive intents as keywords for interaction and developed an intent extraction and visualization technique to support search interactions. Rha, Mitsui, Belkin and Shah (2016) empirically investigated the connections between search intentions and queries and found that users' search intentions in query segments are significantly associated with their query reformulation types. Hence, knowing search intentions can help researchers better understand what a user is trying to accomplish with a formulated or reformulated query. Furthermore, Liu, Mitsui et al. (2019) sought to build a multi-level framework of Web search and studied the relations between task, search intention and user behaviors. They found that (1) task features (i.e. objective task type and subjective task perception) have significant impacts on the frequencies of occurrences of different intentions, and (2) users' search evaluation is closely associated with many search intentions in query segments. Therefore, using search intentions as features in prediction models may help systems better capture query failures in task-based search interactions.

Summary

Based on the discussions above, to achieve the goal of query failure detection, it is important to combine past search interaction features with query features. In addition, researchers should also consider the roles of task types and search intentions in query failure prediction.

METHOD

To answer the proposed RQs, we conducted a controlled-lab user study and performed correlation and prediction analyses based on the collected data. The specific methods and techniques are fully explained in the following subsections.

Search Tasks and Participants

In the controlled-lab user study, we designed and assigned four work tasks, all within journalism domain, in order to control the task context: Copy editing (CPE), Story pitch (STP), Relationships (REL), and Interview preparation (INT). In addition, to control the variance in task familiarity, we recruited 40 upper-division undergraduates who have finished either one journalism writing or reporting class.

We defined four different types of search tasks based on four task facets extracted from the Li and Belkin (2008)'s faceted classification scheme, as modified by Cole, Hendahewa, Belkin, and Shah (2015): Product, Goal, Level, and Named. The values of these facets are explained in Table 1.

Each of the four task types has two different versions (i.e. different topics: coelacanths; methane clathrates and global warming). We selected these two specific topics for controlling participants' familiarity with the task topic, as our participant population was considered likely to not be familiar with either. To avoid possible biases in task types and topics, we asked each participant to do two tasks of different topics and types, in Latin Square design, pairing CPE with INT, and STP with REL. The four specific tasks for the topic of coelacanth are listed as follows. Note that the work task associated with each search task was designed as a context or cover story. Participants only need to complete the search task part in this study.

Facets	Values and definitions
Product	Intellectual: a task which produces new ideas; Factual: a task locating facts, data, or other information objects
Goal	Specific: a task with explicit and measurable goal; Amorphous: the goal of task is not explicitly defined and cannot be quantitatively measured
Level	Document level: a task for which a document as a whole is judged; Segment level: a task for which a part or parts of document are judged
Named	True: what is to be found is explicitly named; False: what is to be found is not explicitly named

Table 1. Facet values and definitions.

- **CPE** (Factual Specific Segment Named True).

Work task: You are a copy editor at a newspaper and you have only 20 minutes to check the accuracy of six italicized statements in the excerpt of a piece of news story below.

Search task: Please find and save an authoritative page that either confirms or disconfirms each statement.

- **STP** (Factual Amorphous Segment Named False).

Work task: You are planning to pitch a science story to your editor and need to identify interesting facts about the coelacanth, a fish that dates from the time of dinosaurs and was thought to be extinct.

Search task: Save pages that contain the six most interesting facts about coelacanths and/or research about their preservation.

- **REL** (Intellectual Amorphous Document Named True).

Work task: You are writing an article about coelacanths and conservation efforts. You have found an interesting article about coelacanths but in order to develop your article you need to be able to explain the relationship between key facts you have learned.

Search task: In the following, there are five italicized passages, find an authoritative Web page that explains the relationship between two of the italicized facts.

- INT (Intellectual Amorphous Document Named False).

Work task: You are writing an article that profiles a scientist and their research work. You are preparing to interview Mark Erdmann, a marine biologist, about coelacanth and conservation programs.

Search task: Identify and save authoritative Web pages for the following: Identify two (living) people who likely can provide some personal stories about Dr. Erdmann and his work. Find the three most interesting facts about Dr. Erdmann’s research. Find an interesting potential impact of Dr. Erdmann’s work.

Query Failure

Compared to the traditional relevance-based measurements, usefulness can serve as a better evaluation criterion as it considers both task success and the value of system support over the entire information seeking episodes (Cole, Liu et al., 2009). Hence, in this study, we define query state based on the notion of usefulness. Specifically, query failure refers to the query segments where users fail to find any useful document to support the task. In contrast, useful query refers to the query segments where users find at least one useful document. Built upon this operational definition of query failure, we classified query segments into the two categories and built classifiers for the query failure prediction task.

Search Intentions

In this study, search intentions are defined as the things that users seek to accomplish *within a query segment*. Knowing search intentions is likely to be helpful for proactively capturing query failures. To systematically elicit data from participants about their intentions, we adopted the typology of search intention proposed by Rha et al. (2016) and asked

Category	Search intention
Keep record	Keep record of a link (KR)
Identify search information	Identifying something to start (IS); Identify something more to search (IM)
Learn	Domain knowledge (LK); Database content (LD)
Find	Find known item(s) (FK); Find specific information (FS); Find items sharing a named feature (FN); Find items without predefined criteria (FW)
Access item(s)	Access a specific item (AS); Access items with common characteristics (AC); Access a website or similar (AW)
Evaluate	Evaluate correctness of an item (EC); Evaluate usefulness of an item (EU); Pick best items from all the useful ones (EB); Evaluate specificity of an item (ES); Evaluate duplication of an item (ED) (i.e., determine whether the information in one item is the same as in others)
Obtain	Obtain specific information to highlight or copy (OS); Obtain part of an item (OP); Obtain a whole item(s) (OW)

Table 2. Search intentions in query segments.

participants to identify their intentions for *each query segment* based on the typology. This typology is relatively comprehensive and includes 20 intentions (see Table 2). The annotation process is explained in the following section.

Study Procedure

The study procedure (see Figure 2) started with a demographic questionnaire and a tutorial video on the browser interface features offered by the plugin. After finishing the demographic, participants read the task description and completed a pre-search questionnaire on topic familiarity, task experience, and estimated task difficulty. Each participant had up to 20 min to complete the assigned search task, by searching on the Web within the supplied browser. During the search process, our Firefox browser plugin automatically recorded participants’ search activities (e.g., query, timestamp, URL, snippet, page type).

After completing the search task and post-search questionnaire on task perception, participants were asked to read a guidance of intention annotation task. Then, the browser plugin replayed the entire search process, query segment by query segment, asking participants to annotate their intention(s) (i.e., identify if the intention was present or absent) for each segment based on the predefined typology. Note that within each query segment, participants were allowed to select any number of intentions from the list as they might have multiple search intentions in one local step of search. Hence, if we consider the totality of all intentions as users’ cognitive space during search, then the specific search intentions are essentially separate dimensions of the cognitive space, rather than different values of the same categorical variable. Figure 3 presents the post-search replay and intention annotation interface.

The same procedure (from pre-search questionnaire to intention annotation) was repeated for the second search task. The

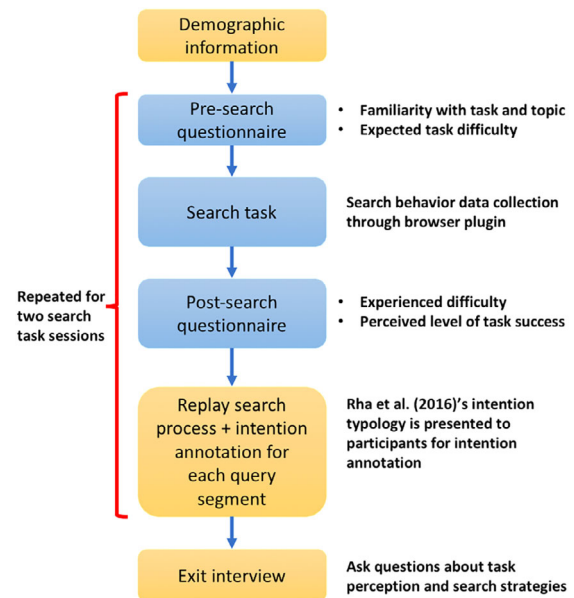


Figure 2. Study procedure.

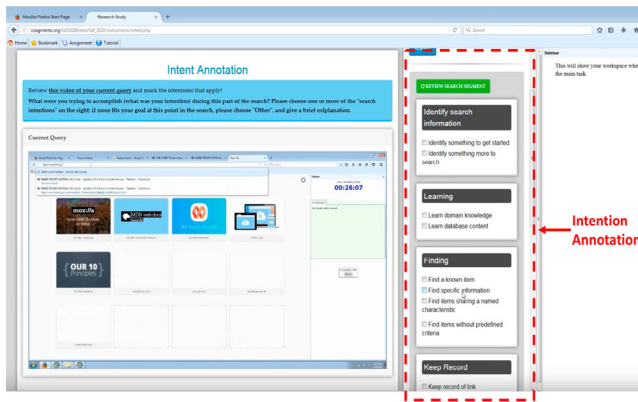


Figure 3. Intention annotation interface. When the main screen was replaying a query segment, the user can identify associated search intention(s) within the side bar on the right.

whole study session (around 2 hr in total) ended with an exit interview in which we asked questions about participants' task perceptions and search strategies.

Behavioral Measures and Data Analysis

The central research problem of this study is to proactively identify or predict query failure from past search behaviors and current query attributes. To address this problem, based on previous empirical studies on query quality prediction and evaluation, we extracted a series of behavioral features from multiple aspects of users' search interactions (see Table 3). Regarding the dwell time and open time features, dwell time on a page refers to the amount of time that a user actually spent on reading the contents on the page. In contrast, open time measures how long a page has been kept open within a query segment (Hienert, Mitsui, Mayr, Shah, & Belkin, 2018; Mitsui et al., 2017). Note that the query behavior category includes both previous query features and current query attributes. Query reformulation type includes: generalization, specialization, repeat existing query, new query, word substitution, and other reformulation types. We sought to predict query failure at the point where a query is formulated but the retrieval and result examination step is not yet executed (see Figure 1).

Types	Behavioral measurements
Query behavior	Query length, query reformulation type
Browsing	Number of clicks, number of tabs, number of pages visited, number of sources (i.e. unique Internet domains) visited
Dwell time	Mean dwell time on content page, total dwell time on content pages, total dwell time on SERP, total session elapsed time
Display/open time	Total open time on content pages, total open time on SERP

Table 3. Behavioral Measurements.

With respect to data analysis, to answer the RQ1, this study mainly used Mann-Whitney U test to examine the connection between past search behavior and current query state (binary)

as none of the search behavior data was normally distributed. The statistical test for each behavior feature was conducted between the useful query group (at least one useful document bookmarked) and query failure group. To answer RQ2, we employed multiple classifiers (e.g., Logistic regression, Naïve Bayes, SVM) to predict query failure. For RQ3, we added query-level search intentions as features into the classifier/prediction models and see to what extent we can improve the performance of query failure prediction with the knowledge of user intentions. In addition, given that users' search interactions and performances often vary across tasks of different types (Liu et al., 2010), we also ran the prediction models in separate task contexts and compared the prediction performances across different task types, aiming to enhance our understanding of the predictability of query failure. In prediction analyses, we compared the classifiers built upon a series of search features with the baseline model (ZeroR). We trained and evaluated models using 5-fold cross-validation. The results are explained in the following section.

RESULTS

To answer the proposed research questions, we collected search behavior and intention data from 693 query segments generated by 40 participants in 80 task-based sessions. Since the last query segment of each session does not have subsequent query, we did not use the search interaction data generated in these query segments. Hence, we conducted data analyses based on the search interaction and intention data extracted from 613 query segments (184 query segments in CPE tasks, 90 in STP tasks, 137 in REL tasks, and 202 in INT tasks). The results of data analyses are organized according to the RQs in the following subsections.

RQ1: Search behavior and subsequent query state

To examine the associations between search behaviors in the current query segments and subsequent query states, we ran Mann Whitney U test for most pairs of behavior measure and query state (see results shown in Table 4). To statistically test the correlations between query reformulation and subsequent query state, we adopted Pearson's Chi-square test as query reformulation type is a categorical variable.

Since this study applied nonparametric tests, Table 4 reports the medians and inter-quartile ranges (IQR) (instead of means and standard deviations) of behavior measures. In Pearson's Chi-square tests we did not find any statistically significant correlation between query reformulation type (either current or subsequent) and subsequent query state.

According to the results shown in Table 4, overall, participants who were better prepared in the current query segment were less likely to encounter query failure in the following query segment. Specifically, when a participant was more active in interacting with and extracting knowledge from current query segments (e.g., spent longer time on content pages and SERP, visited more unique Internet domains), the

Behavior measure	Useful	Failure	Mann-Whitney test
Current query length	4(3)	4(3)	$z=.479$, $p=.632$
Next query length*	5(3)	4(3)	$z=2.16$, $p=.029$
Number of clicks	3(4)	3(4)	$z=.716$, $p=.474$
Number of tabs	3(3)	3(2)	$z=-1.84$, $p=.064$
Number of pages	5(4)	5(3)	$z=-1.48$, $p=.140$
Number of sources*	4(2)	3(3)	$z=-2.23$, $p=.026$
Mean dwell time content*	10.2(15.9)	8.4(13.5)	$z=2.15$, $p=.031$
Total dwell time content	48.5(83.2)	39.5(81.7)	$z=.949$, $p=.343$
Total dwell time SERP*	10.5(15.5)	7.2(9.3)	$z=2.49$, $p=.013$
Total session elapsed time*	6.6(7.6)	8.3(7.7)	$z=-2.43$, $p=.015$
Total open time content	155(287)	157(341)	$z=-.135$, $p=.892$
Total open time SERP**	39.2(33.6)	25.8(18.5)	$z=2.72$, $p=.007$

Table 4. Correlations between current search behavior and next query state: Median (IQR) and Mann-Whitney U test. The unit of total session elapsed time is minute. *: $p<.05$, **: $p<.01$. The statistically significant correlations are boldfaced.

participant was more likely to enter into the useful query state in the subsequent query segment. This may be because when participants engaged in deeper interactions with current SERPs and results, they had a better chance to obtain more useful information and knowledge from the retrieved pages and then to use them in facilitating query formulation and result examination in the following query segment. In addition, we found that as the search session proceeded, participants were more likely to experience query failure, especially in the middle and later stages (see Figure 4). This may be because participants gradually raised the threshold for page usefulness judgment as they obtained more and more useful information along the way. Another possible reason is that in later search stages participants encountered more repeated documents which were judged as useful documents earlier and had difficulty in finding new useful pages to support the task at hand.

Based on the identified connections between current search behavior and subsequent query state, in the following subsection we employed search behavior and intention features to build classifiers for predicting query failure *before* the retrieval and SERP examination steps.

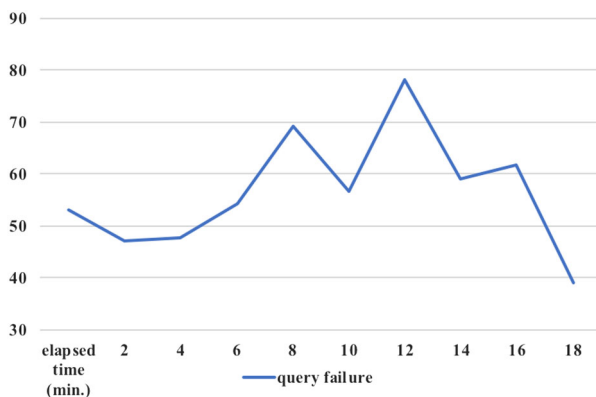


Figure 4. The frequency (%) of query failure during search.

RQ2 & RQ3: Predicting query failure

RQ2 and RQ3 speak to the problem of proactively identifying query failure in search process. Particularly, as a preparation for answering RQ3, Figure 5 reports the proportions of query segments where the intention was identified by users as present (which means the users had this intention in the associated query segments). The results regarding intention distribution indicate that for most search intentions, the proportions of the query segments where the intention was present is clearly smaller than the opposite situation (i.e. where the intention was absent). This result suggests that no search intention is ubiquitous in any of the task-based search sessions. Hence, different combinations of search intentions may correspond to largely different cognitive states, which may significantly affect users' search strategies and performances in the associated query segments.

To better compare the performances of baseline model (ZeroR), search behavior models and behavior-intention models, we report and explain the results of all prediction analyses in the same subsection. Figures 6–8 only show the performances of best classifiers in prediction for brevity. The full result tables presenting the performances of all classifiers are provided in the Tales A1–A6.

Accuracy is a way to measure overall performance improvement in a classification problem. According to the results illustrated in Figure 6, we made gains in accuracy by building prediction models based on search behavioral features (3% above the ZeroR baseline). Also, adding search intention features into the classifiers resulted in additional 5% increase in overall accuracy. Since the main focus of this study was proactively capturing query failures in search process, we also analyzed the precision of predicting query failures. The results indicate that the behavior model clearly outperforms the baseline (11% increase), and that adding intention features can obtain another 2.8% improvement.

Task type usually plays a central role in affecting users' task-based search interactions. Given that users often have different search intention patterns, search strategies, and performances in tasks of different types (Aula, Khan, & Guan,

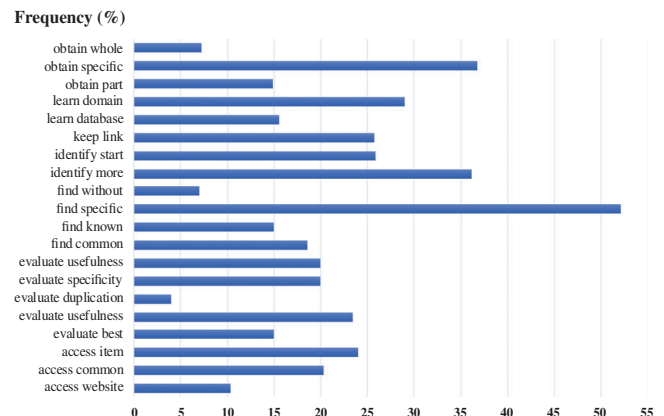


Figure 5. Intention selected in query segments (%).

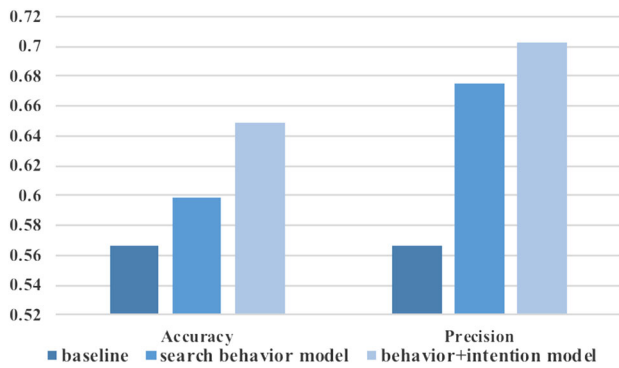


Figure 6. Prediction performance: overall accuracy and precision of predicting query failure.

2010; Liu et al., 2010; Liu, Mitsui et al., 2019), we also ran our prediction models in the four task contexts respectively.

According to the results in Figure 7, the search behavior model had better performance in the two factual tasks (CPE and STP), compared to the two intellectual tasks (REL and INT). One possible reason behind this is that the needed information items and dimensions for result evaluations were more well-defined in factual tasks than in intellectual tasks. As a result, the connection between search behavioral patterns and query states could be more predictable in factual tasks, which might make it easier to differentiate useful queries from query failures based on search behavioral features. Overall, the behavior model achieved 2%–6% increase in accuracy for the first three tasks. There is no clear improvement in accuracy for the INT task. In terms of the behavior-intention models, it is worth noting that incorporating search intention features into the classifiers led to a significant increase (11.4%) in accuracy for the INT task where search behavior model failed to achieve any major improvement. This result suggests that there might be a big gap between users’ cognitive space and explicit search strategies, especially in intellectual amorphous task. In other words, in the INT tasks, search intention features still contained plenty of useful, implicit, predicative signals which could not be captured by explicit search behaviors. Perhaps more fine-grained behavioral measures or neuro-physiological signals can help capture the hidden variances and disconnections in the cognition-behavior gap.

Compared to improving the overall accuracy of query state prediction, it is more important to accurately capture potential query failures where timely system supports and interventions are badly needed. To what extent can we proactively identify query failures based on search behavior and intention features in tasks of different types? to answer this question, we looked at the precision of query failure prediction in different task contexts. According to the results in Figure 8, search behavior models achieved significant improvement (ranging from 9% to 14%) compared to the baseline model. This result proves that the classifiers built upon past search behavioral measures are

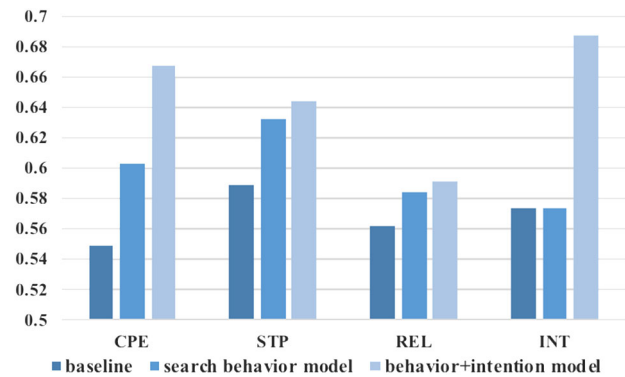


Figure 7. Results for accuracy in tasks of different types.

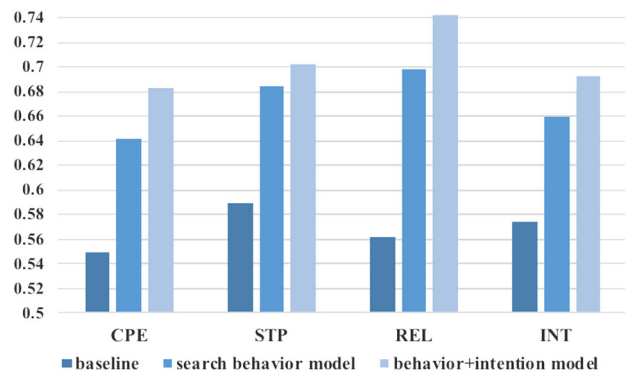


Figure 8. Precision of predicting query failure.

useful in identifying subsequent query failures and capturing the opportunities for providing proactive system interventions. Furthermore, assume we know users’ intentions for the subsequent query segments. Built upon this assumption, we found that adding search intention features can help further improve the precision score in all task types, indicating that the knowledge of users’ intentions in local steps can facilitate the identification of potential query failures. The search intention transitions and cognitive activities are often more active in the tasks where the goal is not clearly defined (i.e. goal-amorphous task) as users need to do more open-ended explorations when trying to define search goals and evaluate documents (Liu, Mitsui et al., 2019). As a result, compared to the goal-specific task (CPE task), search intention features might offer more useful information about users’ cognitive space in goal-amorphous tasks and thereby helped achieve higher precision scores in the STP, REL, and INT tasks.

DISCUSSION

Page browsing and result examination is cognitively costly for users in search interactions. Proactively capturing queries that are very likely to be useless for the task at hand can help save a user’s limited cognitive resource (e.g., working memory, visual processing system) and thereby improve his or her search efficiency and experience. To explore the possibility of offering proactive system interventions for search system

users in local search steps (i.e. query segments), this study examined the extent to which we can predict query failure from search behaviors and intentions before the retrieval and result examination steps. With respect to the research questions, we have following answers.

In terms of the association between past search behaviors and subsequent query state (RQ1), we found that participants were less likely to experience query failure in the following search step if they were more active in interacting with the current search results (e.g., spent longer time on SERP and clicked content pages, visited more unique Internet domains). This result echoes Wang et al. (2017)'s field study finding that when users spent more efforts on carefully examining search results, they usually encounter less search obstacles and problems. In addition, our finding empirically confirms the underlying connections between adjacent local steps of Web search identified in previous IIR studies (e.g., Liu, Mitsui et al., 2019; Teevan, Alvarado, Ackerman, & Karger, 2004) and highlights the importance of leveraging the existing signals from search interactions to proactively identify future opportunities for task-based system support and intervention. With respect to user persuasion and education, the answer to RQ1 also supports Liu, Wang, Mandal, and Shah (2019)'s argument that encouraging deeper, fully engaged search interaction (through explicit intervention or intelligent nudging) is critical for a search system to improve users' task performances and search experiences.

Besides, we also found that participants tended to encounter query failures more frequently in the middle and late search stages compared to the early stages (see Figure 4, the two peak points in query failure frequency at 8th and 12th minutes). This could result from the effect of *threshold priming*. Specifically, when users successfully gathered many highly relevant pages in early stages, they might gradually (and unconsciously) raise their thresholds for usefulness judgment (Scholer, Kelly, Wu, Lee, & Webber, 2013). As a result, it might become more difficult for them to identify useful pages that meet their requirements in the later stages. Another possible reason is that users might use repeated keywords and queries in later search stages and encountered repeated search results, which also led to the difficulty in identifying new useful documents. This finding indicates that the usefulness judgment process in Web searching is highly dynamic and subjective in nature, and that different dimensions of usefulness judgments and their associations with users' evolving mental models still need to be examined in detail.

To answer RQ2, we adopted several classification algorithms and built prediction models based on a series of search behavioral measures. Overall, our result proves that behavior-based classifiers always outperform their corresponding baseline models in search tasks of different types. Particularly, compared to the baselines, the prediction models built on past search behaviors achieved 9% to 14% improvements in the precision of query failure prediction. This result demonstrates that search behavioral signals are quite useful in proactive identifications of

query failures, especially in complex, goal-amorphous tasks where search processes are not well supported and users often need to take a zigzag path toward the ultimate task goal (Capra et al., 2018; Liu, Mitsui et al., 2019). Through building a behavior-based model for predicting query failure, our study goes beyond the traditional inference approach and speaks to the issue of developing proactive supports for real-time search interactions. To further improve the performance of behavioral model, future studies can investigate more diverse, fine-grained behavioral measures that can help capture some other implicit aspects of search tactics in local steps.

For RQ3, we added search intentions as features into the prediction models and found that the behavior-intention models can further improve the overall accuracy and the precision of query failure prediction to varying degrees in tasks of different types. This finding supports Ingwersen (1996)'s argument that from a cognitive perspective, users' search interactions are affected by multiple dimensions of cognitive space. Also, it demonstrates that a deeper understanding of cognitive changes can help researcher better study search interactions and predict search outcomes. In addition, the result speaks to the practical value of predicting local-step search intentions using behavioral signals collected from search sessions (cf. Mitsui et al., 2017).

As always, there are several limitations to this study as well as needs for future research efforts. First, the size of dataset and task types are limited. Many of the potential task facet combinations defined by Li and Belkin (2008)'s framework still remained untouched in this study. To further test the predictive power and generalizability of the classifiers, we will run them on multiple, larger scale datasets generated in different task contexts. Second, search intentions as features still require participants' annotations. The findings from recent works on characterizing and inferring search intentions (e.g., Liu, Mitsui et al., 2019; Mitsui et al., 2017) can be of help for partially addressing this issue. Third, the behavioral features extracted for model building are by no means complete. Therefore, defining and extracting more fine-grained behavioral features from multiple sources (e.g., search logs, eye tracking devices) will be the next step of our future research on predicting query failures in Web search.

CONCLUSION

Few IIR research have tried to leverage the signals from past search interactions in proactively capturing future query failures. Given this research gap, we built query failure prediction models based on search behavioral features and tested them in predicting the query state (useful query or query failure) in the subsequent query segments. Furthermore, to examine the usefulness of search intention features as signals from cognitive level, we integrated intention features with behavioral features in behavior-intention models and applied the models in our research context. The results demonstrate that search behavior models always outperform the baseline models in all tasks types, and that using the knowledge of

users' search intention(s) can help further improve the overall accuracy as well as the precision of query failure prediction.

The ultimate goal of this research line is to provide proactive system supports and interventions in real-time search interactions before users encounter search problems (in this case, query failure). To this end, future research needs to (1) take advantages of the findings from other relevant research and further examine different dimensions of search problem detection, such as user frustration prediction (Feild et al., 2010) and expectation disconfirmation prediction (Liu & Shah, 2019), and (2) explore the possible directions for designing proactive system recommendations and interfaces.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation (NSF) grant IIS-1717488. It also uses datasets generated as a part of the project supported by the NSF grant IIS-1423239.

REFERENCES

- Aula, A., Khan, R. M., & Guan, Z. (2010). How does search behavior change as search becomes more difficult? In *Proceedings of the Conference on Human Factors in Computing Systems* (pp. 35–44). ACM.
- Bloom, B. S. (1956). *Taxonomy of educational objectives Cognitive domain* (Vol. 1, pp. 20–24). New York, NY: McKay.
- Capra, R., Arguello, J., O'Brien, H., Li, Y., & Choi, B. (2018). The effects of manipulating task determinability on search behaviors and outcomes. In *the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 445–454). ACM.
- Cole, M., Liu, J., Belkin, N., Bierig, R., Gwizdka, J., Liu, C., ... Zhang, X. (2009). Usefulness as the criterion for evaluation of interactive information retrieval. In *Proceedings of HCIR* (pp. 1–4).
- Cole, M. J., Hendaheba, C., Belkin, N. J., & Shah, C. (2015). User activity patterns during information search. *ACM Transactions on Information Systems (TOIS)*, 33(1), 1.
- Feild, H. A., Allan, J., & Jones, R. (2010). Predicting searcher frustration. In *Proceedings of the 33rd SIGIR Conference on Research and Development in Information Retrieval* (pp. 34–41). ACM.
- Guo, Q., White, R. W., Dumais, S. T., Wang, J., & Anderson, B. (2010). Predicting query performance using query, result, and user interaction features. In *Proceedings of Adaptivity, Personalization and Fusion of Heterogeneous Information* (pp. 198–201). ACM.
- Hauff, C., Hiemstra, D., & de Jong, F. (2008). A survey of pre-retrieval query performance predictors. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (pp. 1419–1420). ACM.
- He, B., & Ounis, I. (2006). Query performance prediction. *Information Systems*, 31(7), 585–594.
- Hienert, D., Mitsui, M., Mayr, P., Shah, C., & Belkin, N. J. (2018). The role of the task topic in web search of different task types. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (pp. 72–81). ACM.
- Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation*, 52(1), 3–50.
- Jiang, J., He, D., & Allan, J. (2014). Searching, browsing, and clicking in a search session: Changes in user behavior by task and over time. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 607–616). ACM.
- Kelly, D., Arguello, J., Edwards, A., & Wu, W. C. (2015). Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework. In *Proceedings of the 2015 Conference on The Theory of Information Retrieval* (pp. 101–110). ACM.
- Kumaran, G., & Carvalho, V. R. (2009). Reducing long queries using query quality predictors. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 564–571). ACM.
- Li, Y., & Belkin, N. J. (2008). A faceted approach to conceptualizing tasks in information seeking. *Information Processing and Management*, 44(6), 1822–1837.
- Liu, J. (2017). Toward a unified model of human information behavior: An equilibrium perspective. *Journal of Documentation*, 73(4), 666–688.
- Liu, J., Cole, M. J., Liu, C., Bierig, R., Gwizdka, J., Belkin, N. J., ... Zhang, X. (2010). Search behaviors in different task types. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries* (pp. 69–78). ACM.
- Liu, J., Mitsui, M., Belkin, N. J., & Shah, C. (2019). Task, information seeking intentions, and user behavior: Toward a multi-level understanding of web search. In *Proceedings of International ACM SIGIR Conference on Human Information Interaction and Retrieval* (pp. 123–132). ACM.
- Liu, J., & Shah, C. (2019). Investigating the impacts of expectation disconfirmation on web search. In *Proceedings of Conference on Human Information Interaction and Retrieval* (pp. 319–323). ACM.
- Liu, J., Wang, Y., Mandal, S., & Shah, C. (2019). Exploring the immediate and short-term effects of peer advice and cognitive authority on Web search behavior. *Information Processing and Management*, 56(3), 1010–1025.
- Mitsui, M., Liu, J., Belkin, N. J., & Shah, C. (2017). Predicting information seeking intentions from search behaviors. In *Proceedings of the 40th ACM SIGIR Conference on*

Research and Development in Information Retrieval (pp. 1121–1124). ACM.

Rha, E. Y., Mitsui, M., Belkin, N. J., & Shah, C. (2016). Exploring the relationships between search intentions and query reformulations. *Proceedings of the Association for Information Science and Technology*, 53(1), 1–9.

Ruotsalo, T., Peltonen, J., Eugster, M. J., Głowacka, D., Floréen, P., Myllymäki, P., ... Kaski, S. (2018). Interactive intent modeling for exploratory search. *ACM Transactions on Information Systems (TOIS)*, 36(4), 44.

Scholer, F., Kelly, D., Wu, W. C., Lee, H. S., & Webber, W. (2013). The effect of threshold priming and need for cognition on relevance calibration and assessment. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 623–632). ACM.

Shah, C. (2018). Information fostering-being proactive with information seeking and retrieval: Perspective paper. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (pp. 62–71). ACM.

Smith, C. L., & Kantor, P. B. (2008). User adaptation: Good results from poor systems. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 147–154). ACM.

Teevan, J., Alvarado, C., Ackerman, M. S., & Karger, D. R. (2004). The perfect search engine is not enough: A study of orienteering behavior in directed search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 415–422). ACM.

Wang, Y., Liu, J., Mandal, S., & Shah, C. (2017). Search successes and failures in query segments and search tasks: A field study. *Proceedings of the Association for Information Science and Technology*, 54(1), 436–445.

Wu, W. C., Kelly, D., & Huang, K. (2012). User evaluation of query quality. In *Proceedings of the 35th ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 215–224). ACM.

Classifier	Accuracy	Precision
Logistic regression	.599	.621
Naïve bayes	.566	.675
AdaBoost	.579	.579
Support vector machine	.586	.595
Random forest	.566	.602
ZeroR (baseline)	.566	.566

Table A1. overall performance: search behavior model. The best performers are boldfaced.

Classifier	Accuracy	Precision
Logistic regression	.625	.650
Naïve bayes	.648	.703
AdaBoost	.643	.654
Support vector machine	.649	.652
Random forest	.631	.652
ZeroR (baseline)	.566	.566

Table A2. overall performance: behavior-intention model. The best performers are boldfaced.

Classifier	CPE	STP	REL	INT
Logistic regression	.576	.589	.547	.480
Naïve bayes	.571	.556	.581	.490
AdaBoost	.603	.611	.518	.574
Support vector machine	.560	.611	.569	.545
Random forest	.594	.633	.584	.525
ZeroR (baseline)	.549	.589	.562	.574

Table A3. Accuracy in tasks of different types: behavior model. The best performers are boldfaced.

Classifier	CPE	STP	REL	INT
Logistic regression	.598	.654	.609	.545
Naïve bayes	.582	.659	.698	.659
AdaBoost	.619	.673	.566	.576
Support vector machine	.583	.645	.618	.575
Random forest	.642	.685	.616	.579
ZeroR (baseline)	.549	.589	.562	.574

Table A4. Precision of query failure prediction: behavior model. The best performers are boldfaced.

Classifier	CPE	STP	REL	INT
Logistic regression	.641	.578	.591	.584
Naïve bayes	.598	.622	.672	.545
AdaBoost	.658	.644	.555	.688
Support vector machine	.663	.589	.584	.589
Random forest	.668	.578	.577	.624
ZeroR (baseline)	.549	.589	.562	.574

Table A5. Accuracy in tasks of different types: behavior-intention model. The best performers are boldfaced.

Classifier	CPE	STP	REL	INT
Logistic regression	.670	.647	.624	.636
Naïve bayes	.606	.702	.742	.682
AdaBoost	.683	.672	.603	.693
Support vector machine	.679	.648	.625	.630
Random forest	.682	.642	.607	.667
ZeroR (baseline)	.549	.589	.562	.574

Table A6. Precision of query failure prediction: behavior-intention model. The best performers are boldfaced.