Characterizing and Early Predicting User Performance for Adaptive Search Path Recommendation

Wang, Ben School of Library and Information Studies, University of Oklahoma, USA | benw@ou.edu

Liu, Jiqun School of Library and Information Studies, University of Oklahoma, USA | jiqunliu@ou.edu

ABSTRACT

User search performance is multidimensional in nature and may be better characterized by metrics that depict users' interactions with both relevant and irrelevant results. Despite previous research on one-dimensional measures, it is still unclear how to characterize different dimensions of user performance and leverage the knowledge in developing proactive recommendations. To address this gap, we propose and empirically test a framework of search performance evaluation and build early performance prediction models to simulate proactive search path recommendations. Experimental results from four datasets of diverse types (1,482 sessions and 5,140 query segments from both controlled lab and natural settings) demonstrate that: 1) Cluster patterns characterized by cost-gain-based multifaceted metrics can effectively differentiate high-performing users from other searchers, which form the empirical basis for proactive recommendations; 2) whole-session performance can be reliably predicted at early stages of sessions (e.g., first and second queries); 3) recommendations built upon the search paths of system-identified high-performing searchers can significantly improve the search performance of struggling users. Experimental results demonstrate the potential of our approach for leveraging collective wisdom from automatically identified high-performance user groups in developing and evaluating proactive in-situ search recommendations.

KEYWORDS

User performance; Web search; search path recommendation; evaluation; proactive information retrieval

INTRODUCTION

Characterizing user search performance is crucial for modeling user behavior and providing them with customized recommendations in interactive information retrieval. The search path and strategies of users with high performance can be leveraged to enhance other users' search results and experience under similar search tasks, cognitive states, and intents (Hassan Awadallah et al., 2014; Hendahewa & Shah, 2017; J. Liu & Shah, 2022). Users' search performance is shaped by several factors, such as domain knowledge (i.e., users' knowledge background and familiarity with the search task/topic), search expertise (i.e., technical skills and topic-independent search experience) (Suzuki & Yamamoto, 2021), and contextual factors (J. Liu & Shah, 2019a, 2019b). Focusing on search performance, previous research suggested that good-performing users usually have better search strategies and can effectively accumulate knowledge and cues through interactive search activities (Mao et al., 2018; Suzuki & Yamamoto, 2021). Through leveraging collective wisdom from expert users, collaborative query recommendation can help struggling users with lower performance and fewer skills and improve their search interactions (Halvey & Jose, 2012; Morris, 2008; Morris & Horvitz, 2007).

Existing research has examined user search performance using unified *behavior-based* or *self-report-based* metrics to measure performance. The user performance is related to users' interactions with the results (e.g., clicks, dwell time, and continue or stop searching) and their perceived search experience (e.g., search success and satisfaction) (M. Liu, Liu, Mao, Luo, Zhang, et al., 2018; Mao et al., 2016). Concerning user performance prediction, some existing studies employed the performance from previous searches and implicit feedback features (e.g., dwell time on results) to predict their search performance in current and subsequent query segments (Hendahewa & Shah, 2015; Piech et al., 2015), but many of them only considered a small set of online search behavior features collected at the whole session level (after the search session is concluded). These measures of single aspects are not capable of capturing the multidimensional nature of user performance, which may lead to incomplete and biased results in user performance evaluation (Hendahewa & Shah, 2017; Moraveji et al., 2011; Odijk et al., 2015). In addition. although various methods have been developed to support users by improving their search experience in both IR and HCI communities (Huurdeman & Kamps, 2015; Smith, 2017), it is still unclear how we can leverage the knowledge about evaluating and predicting user performance at different early points of search sessions, and how effective the performance measurement is.

Research Objective

To fill the above research gaps, we developed multiple **user performance measures** that capture both the gain (e.g., the numbers of relevant/useful documents) and cost (e.g., the number of clicks, dwell time) aspects of users' search interactions based on both online search behavioral features and offline relevance and usefulness labels. Differing from most of the existing measures, our metric set characterizes users' in-situ interactions with multidimensional gain and cost. Then, based on the user performance groups identified through clustering from the measure scores, we built machine learning classifiers to **predict user performance at early points of search sessions** (e.g., first,

86th Annual Meeting of the Association for Information Science & Technology | Oct. 27 – 31, 2023 | London, United Kingdom.

second, or third query). Moving forward toward practical application, with the simulated query segment recommendation, we investigated the effectiveness of the performance measurement and prediction. The main contributions and implications of our study include:

- This study measures user performance with multidimensional metrics, and tested the predictability and effectiveness of the performance evaluation process under varying search datasets, which can be widely applied and reused in characterizing search performance and evaluating search paths.
- The study demonstrates the feasibility of early predicting user performance based on their search behaviors and collaboratively generating adaptive search path recommendations at early points of search sessions.
- At the practical level, this study demonstrates a viable approach to leveraging the collective wisdom of multiple users in collaborative search path recommendations and helping struggling users engage in complex search tasks going beyond ad hoc retrieval contexts.

RELATED WORK

In this section, we introduced the theoretical foundations of our proposed framework and the research gap. First, as we aimed to evaluate users' search performance, we reviewed previous studies about the measures and the importance of users' performance in search tasks. Then, we planned to predict users' search performance based on their behaviors, so we reviewed methods investigating user behavioral features. Furthermore, with user performance measures and predictors, we proposed to design a method to help struggling users in search tasks, which is the interest of proactive information retrieval. Finally, we specified the research gaps in these fields and brought up our research questions.

User Online Search Performance

Improving user search performance and experience with search system support is fundamental in information retrieval (Huurdeman & Kamps, 2015). User search performance is shaped by many factors, such as domain knowledge (M. Liu, Liu, Mao, Luo, Zhang, et al., 2018; Suzuki & Yamamoto, 2021), task complexity (P. Liu & Li, 2012; Payne, 1976; Willett et al., 2014), search skills, and search experience (Bailey, 2017; Wildemuth, 2004). Users from various domains with different levels of search experience will perform differently on specific search tasks (M. Liu, Liu, Mao, Luo, Zhang, et al., 2018; Moraveji et al., 2011; Suzuki & Yamamoto, 2021). Evaluating user search performance usually involves assessing their success in the search task, which can be done through self-reports or analyzing user responses (M. Liu, Liu, Mao, Luo, Zhang, et al., 2018; Odijk et al., 2015). However, collecting user-generated annotations can be costly, so there is a need for metrics that evaluate user performance with limited user annotations and implicit feedback. Recent research has focused on online metrics for information retrieval systems that combine offline metrics (e.g., relevance-based metrics) and online user behaviors (F. Zhang et al., 2020). These metrics can reflect user search behaviors and performance from multidimensional aspects (Hendahewa & Shah, 2017; J. Liu & Yu, 2021). According to information foraging theory, user interaction signals are associated with their search outcomes in varying ways, and this outcome-behavioral relationship is affected by the goal of information gain, tolerance of cost, or the rate of gain and cost (Azzopardi et al., 2018). However, each online metric only evaluates search results from a single aspect, making it challenging to represent and evaluate multidimensional user search performance using user-interaction signals (Chen et al., 2017).

On the other hand, researchers have studied query performance prediction (QPP) methods to estimate query performance or system effectiveness without relevance labels (Cronen-Townsend et al., 2002; Raiber & Kurland, 2014). These methods estimate performance using textual features of the query and documents in the search results (Zendel et al., 2019). However, QPP methods neglect users' experience and are not designed for improving users' performance in web search (Zamani et al., 2018; Zhou & Croft, 2007). Therefore, how to adopt user-interaction signals to represent and evaluate the multidimensional search performance remains an open question.

Online Search Behaviors

User search performance is closely associated with their online search behaviors (F. Liu, 2016). With the development of the modern search engine, researchers have collected user interaction signals in real-time and at a large scale, aiming to develop user-oriented search supports (Bennett et al., 2011, 2012). These behavioral signals normally involve query reformulation, pagination, clicks, mouse movements, dwell time, etc. Based on information-seeking theories and user models, previous research has analyzed users' online search behaviors to personalize their interactions with search engines by inferring their implicit states (Dumais et al., 2014; Fox et al., 2005; White & Morris, 2007). Various methods have been implemented to investigate users' search behaviors, including collecting user behavior data from user studies or large datasets of search logs and developing machine learning models to predict users' implicit states using behavioral data (Hendahewa & Shah, 2015; Li et al., 2021). The implicit states include task difficulty, user information intention, information need, satisfaction, etc (Arguello, 2014; J. Liu et al., 2020; F. Zhang et al., 2020). In addition, researchers also found distinguishable patterns in online search behaviors

between users with high search performance and those with low performance (Lazonder et al., 2000; Mao et al., 2018; Suzuki & Yamamoto, 2021; Yu et al., 2018). Furthermore, users exhibit different behavioral patterns in different states of the search task. Based on these patterns, researchers have proposed to develop adaptive models to predict user states at different moments during sessions (Arguello, 2014; J. Liu & Yu, 2021). Practically, predicting users' implicit states such as domain knowledge could enable search systems to support users proactively in achieving a better search experience (X. Zhang et al., 2015).

Proactive Information Retrieval

Proactive information retrieval aims to improve users' search experience and performance from the traditional search system with a reactive nature (Bhatia et al., 2016). Users with different levels of search experience, domain knowledge, or information literacy can all get support from proactive search systems (Huurdeman & Kamps, 2015; Savenkov & Agichtein, 2014; Smith, 2017). Query recommendation is a popular proactive information retrieval method that has demonstrated usefulness for users engaging in various types of tasks (Hanauer et al., 2017; Hendahewa & Shah, 2015, 2017; Moayedikia et al., 2019; Zahera et al., 2013). High-quality, collaboratively recommended queries created by expert users can provide guidance to users and improve their search performance (Harvey et al., 2015; Morris, 2008; Morris & Horvitz, 2007; Savenkov & Agichtein, 2014; Smith, 2017). Another aspect of assisting users is helping the search system adapt to users' states according to their behaviors (J. Liu & Yu, 2021; Rha et al., 2016). The adaptive search system can improve search results during the dynamic interaction between users and the search engine (Luo et al., 2014). When users are struggling with search tasks, these proactive methods are supposed to assist users in accomplishing tasks with positive interventions. Therefore, by accurately estimating users' search performance, queries from high-performance users can be collected to help users with low performance of the search task in the proactive information system.

RESEARCH QUESTIONS

Although there is much research about adaptive search systems supporting users, most of them focus on a single facet of user behaviors or the system performance based on traditional relevance-based offline metrics. Few studies investigate user performance directly to develop adaptive search systems because of lacking appropriate evaluation metrics for multidimensional user performance. This research aimed to evaluate user performance with the combination of various online metrics in terms of gain and cost. Then, we investigated the feasibility of early prediction and intervention to support users after evaluating their performance. Therefore, we aimed to address these research questions:

RQ1: From what aspects can we measure user search performance using online-offline combined search measures?

RQ2: How early can we predict user search performance in an interactive search session?

RQ3: To what extent can we improve a user's search performance with search paths extracted from the search sessions of other users engaging in the same search task?

DATA AND METHODS

This section introduces the methodology we employed to create multidimensional user performance measures and evaluate early-prediction-based search path recommendations. First, we introduced the datasets used in this study and what features and performance measures we can extract from these datasets. Then, we described our framework of search performance evaluation, how to investigate the predictability and effectiveness of the performance evaluation by early prediction with behavioral features, and query segment recommendation simulation.

Datasets

As this research aims to support users with queries extracted from high-performance users' search sessions, some features we need include information on users' behaviors and assessments of the results to evaluate user performance. Therefore, we used the KDD19 dataset (M. Liu et al., 2019), SIGIR16 dataset (Mao et al., 2016), SearchSuccess dataset (M. Liu, Liu, Mao, Luo, Zhang, et al., 2018), and the TREC session track 14 dataset (https://trec.nist.gov/data/session2014.html) in this research. We listed the basic information about these datasets in Table 1. These datasets contain topic/task-based search logs where users could have multiple queries in one session, and their search behaviors were collected, including query reformulation, clicks, paginations, and timestamps. The KDD19 dataset also collected additional user behaviors such as mouse movements, scroll, and hover activities.

Dataset	KDD19	SIGIR16	SearchSuccess	TREC14
# Queries	1538	937	651	2014
# Sessions	450	225	166	641
# Tasks	9	9	6	60
Annotation	Relevance, Usefulness	Relevance, Usefulness	Relevance, Usefulness	Relevance

Table 1. Characteristics of Datasets.

For the annotations, the KDD and SIGIR16 datasets collected relevance for all documents and usefulness for clicked documents. The SearchSuccess dataset collected relevance and usefulness only for clicked documents. The relevance and usefulness annotations are different assessments reflecting two aspects of the documents (Mao et al., 2016). The relevance is assessed by external experts based on the topical relation between the information and the search task. The relevance score is from 0 "not relevant" to 3 "very relevant". The usefulness is annotated by the users about their perceived usefulness of the result, and the score is from 1 "not useful at all" to 4 "very useful. Users' satisfaction feedback on a session was also collected in these three datasets, on a scale of 5, from 1 "low satisfaction" to 5 "high satisfaction". The TREC14 dataset contains only relevance annotations on a scale of 6, including -2 "spam" and 0 to 4 (from "not relevant" to "navigational"). The user interactions in the TREC14 dataset are quite sparse compared to the other three datasets, and only 641 out of 1257 sessions contain at least one click. We filtered out sessions with zero clicks from the TREC14 datasets as we intended to evaluate action-based performance in search sessions. In addition, the KDD19, SIGIR16, and SearchSuccess datasets were collected in a lab-controlled environment for user studies with college students as participants, and the queries are in Chinese. The TREC14 dataset containing English queries was collected for a different purpose as an information retrieval evaluation instead of the user study. The search tasks are general topics in the KDD19, SIGIR16, and TREC14 datasets, while the SearchSuccess dataset contains search tasks in different domains.

Search performance evaluation, early prediction, and recommendation simulation

Figure 1 illustrates the study process, which involves evaluating user search performance, predicting the performance levels, and simulating collaborative query recommendations. We first designed a set of online-offline combined measures and used a clustering analysis on these measures to group users by their performance levels. We then investigated the predictability of performance levels through early prediction by search behaviors and compared the queries between high- and low-performance groups as the simulated collaborative query segment recommendation.

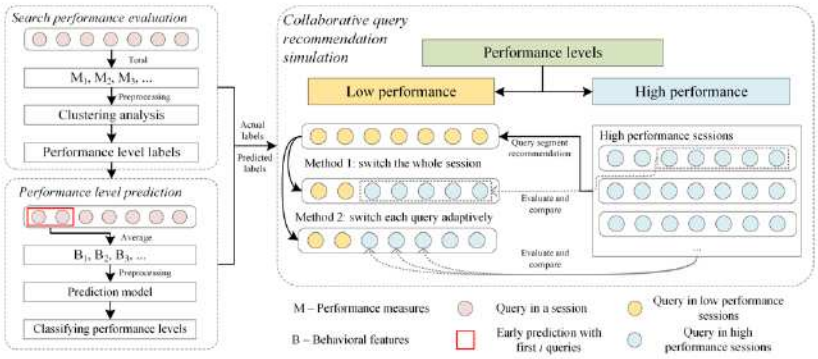


Figure 1. The overall framework of performance evaluation, early prediction, and recommendation simulation.

Search performance measures and evaluation

In the *Search performance evaluation* step in Figure 1, we extracted features of users' behaviors and document annotations from the search session and created online-offline combined measures based on those features and annotations. The online measures were calculated from users' online search behaviors, and the offline measures were calculated based on the annotations of the documents on the search engine result pages (SERPs), such as relevance, usefulness, and discounted cumulative gain (DCG) scores (Dupret, 2011). We combined the online and offline measures to evaluate users' behaviors and grouped the measures into gain- and cost-based measures, which are listed in Table 2. We selected these measures from a large set and removed highly correlated measures. These measures are the sum values at the session level and can represent different aspects of users' gain and cost in completing the search task. We did not involve dwell time on relevant or useful documents as it was unclear when a user had high efficiency in locating required information in the document or spent a long time exploiting abundant relevant information.

Measures		Description
Gain based	DCGrel (use)	Relevance(or usefulness)-based DCG score
	ClicksRel (Use)	Number of relevant (or useful) documents clicked
	ClicksKeyRel	Number of key relevant documents clicked
Cost based	ClicksIrrel (Nonuse)	Number of irrelevant (or non-useful) documents clicked
	DwellIrrel (Nonuse)	Dwell time on irrelevant (or non-useful) documents
	MissRel	Number of relevant documents not been clicked

Table 2. Online-offline combined performance measures grouped by gain and cost.

After obtaining the online-offline combined measures for each session, we conducted a correlation analysis to investigate the relationships between these measures and users' satisfaction state. We used Spearman's rank as the statistical test method since satisfaction is ordinal data. We incorporated measures previously examined (e.g., DCGrel and DCGuse) regarding the correlation with user satisfaction at the query level (Chen et al., 2017; Mao et al., 2016) with other measures we created in this research. We performed the analysis on the KDD19, SIGIR16, and SearchSuccess datasets since the TREC14 dataset does not contain user satisfaction labels. In addition, we utilized the Benjamini–Hochberg method to control the false discovery rate for multiple comparisons (Thissen et al., 2002).

To evaluate users' performance based on measures with multiple aspects of gain and cost, we used K-means clustering analysis on these measures to group search sessions with similar performance characteristics. We normalized all measures in Table 2 and involved them in the clustering analysis. The number of clusters is determined by the elbow method. Then, we visualized the cluster distributions on projections of the gain-cost measure pairs and investigated the patterns for each cluster to identify clusters with characteristics of high performance. We expected to find clusters of sessions with high gain and low cost, characterizing users in these clusters as high performance, and users in clusters of sessions with high cost or low gain as low performance.

Predicting user performance in the session

After investigating the characteristics of these clusters and identifying high-performance patterns, we developed a prediction model to classify search sessions as either high-performance or not by predictor features of users' search behaviors. The performance level is a binary label assigned to sessions based on whether they are in the high-performance cluster. The predictor features include behavioral features from the datasets, and the descriptions are listed in Table 3. We also normalized the data and handled imbalanced data. We developed and compared different classification models, including logistic regression, random forest classifiers, and feedforward neural networks. The classification models were chosen for this study because they are commonly used in machine learning to classify data into different categories. Logistic regression is a simple and widely used model that is effective when the outcome is binary. Random forest classifiers are ensemble learning models that combine multiple decision trees to increase accuracy and reduce overfitting. Feedforward neural networks are powerful models that can learn complex relationships between features and outcomes. The model was evaluated by accuracy, precision, recall, and f1 score.

Features	Description
Mouse and keyboard based	
QueryLength	Number of terms used in an issued query.
UniqueTerm	Number of unique terms used in an issued query.
NewTerm	Number of new unique terms used in an issued query.
QuerySim	Similarity between the current query and the previous query.
ClickCount	Number of clicks.
AvgClickRank	Average rank of clicked results.
Clicks@3/5/5+	Number of clicks between ranks 1-3, 1-5, or below rank 5.
ClickDepth	The deepest or lowest rank of the clicked result.
ActionCount	Number of actions (page click, scroll, query formulation).
ScrollDist	Total scrolling distance.
HoverCount	Number of mouse hovers on results.
Dwell time based	
SERPtime	Total dwell time on search engine result page (SERP).
AvgContent	Average dwell time on content pages.
TotalContent	Total dwell time on content pages.
QueryDwellTime	Total dwell time within a query segment.
TimeFirstClick	Time delta between the start of a session and the first click.
TimeLastClick	Time delta between the start of a session and the last click.

Table 3. Behavioral features.

After comparing the performances of three model settings on prediction with features in the whole session, we chose the best model setting to investigate the early prediction problem by evaluating the model performance at different query orders in the search sessions. From different query orders, the behavioral features are aggregated by the average values. For example, if predicting at the second query, we use average values of behaviors in the first two queries as shown in Figure 1. The models trained with data at different query orders were evaluated to determine how early the researchers could predict user search performance in a session.

Query recommendation simulation

With search performance measures and prediction models, the third step of our research is to verify the effectiveness of our performance evaluation and prediction framework. We investigated how high-performance users differ from others in terms of their search behavior, specifically in terms of their query segments, including the query, retrieved

documents, and click interactions on the documents (Hendahewa & Shah, 2017). As users with high search performance can issue better queries that retrieve more relevant or useful documents on SERPs (Moraveji et al., 2011; Odijk et al., 2015; Salmerón et al., 2005), we wanted to determine whether users with estimated high performance in our study also have better query segments than users with low predicted performance.

To answer RQ3, we simulated a query segment recommendation process to compare the query results between the high-performance sessions and other sessions, shown in the *Collaborative query recommendation simulation* part in Figure 1. The recommendation simulation is developed on two facets: first, given the performance labels, evaluating the recommended query segments, and second, given behavioral features, predicting the performance levels and evaluating the recommended query segments. By comparing the results between these two facets and the original query segments, we can validate the effectiveness of our performance evaluation process in differentiating high- and low-performance users and the feasibility of predicting the performance level with behavior features. We used two methods for recommendation: (1) recommending the entire search path based on the similarity of the i-th query and (2) recommending query segments based on each query's intention. The first method simulates the recommendation of the whole search path from the high-performance group (Hendahewa & Shah, 2017), and the second method designed in this study simulates the adaptive recommendation which is relevant to user search intention and needs.

The query intention was categorized based on the typology of query reformulation types proposed in Rha et al. (2016) and some query intentions are listed in Table 4. In this study, we simplified the query intentions into four main types based on query reformulation types, including generalization, specialization, word substitution, and new, because of the limited data size for specific search tasks. Based on the characteristics of these query intentions, we classified the user's original query intention by comparing the current query and the previous one query. For the adaptive query recommendation simulation, we first collected queries from high-performance queries with the same intention for the recommendation list. Within the recommendation list, we selected the top five queries by similarity with the current query. We evaluated recommended query segments and used the original sessions as the baseline. The evaluation metrics include discounted cumulative gain (DCG) (Dupret, 2011), normalized discounted cumulative gain (nDCG) (Wang et al., 2013), and mean reciprocal rank (mRR) (Chapelle et al., 2009) on both relevance-based scores and usefulness-based scores.

Intention type	Definition	Examples
Generalization	At least one term in common in two queries; the second query contains fewer terms than the first query	campus bike store → bike store
Specialization	At least one term in common in two queries; the second query contains more terms than the first query	bike store → bike store near me
Substitution	At least one term in common in two queries; the second query has the same length as the first query but contains some terms not in the first query	bike store → motorcycle store
New	No common terms in the two queries	bike store → motorcycle dealership

Table 4. Query intention types (Rha et al., 2016).

RESULTS

RQ1: User performance measures and evaluation

After creating the performance measures for each dataset, we did a correlation analysis to investigate the relationships between the performance measures and users' search satisfaction. Table 5 presents Spearman's ranks of measures with users' satisfaction. The cost-based measures, such as ClicksIrrel and MissRel, have negative correlations with satisfaction in these three datasets. This negative impact is brought by users' costs during the search process (Mao et al., 2016). However, some gain-based measures also showed negative correlations with users' satisfaction, such as DCGrel and ClicksRel, which may be related to users' efforts to reach a high gain level. In general, cost-based measures had more significant impacts than gain-based measures in the KDD19 and SIGIR16 datasets, which share similar measures of significant relationships due to the same search tasks and similar data collection procedures (M. Liu, Liu, Mao, Luo, & Ma, 2018). However, for the SearchSuccess dataset, only two cost-based measures significantly related to satisfaction, and ClicksUse was the only measure showing a positive correlation among the three datasets. The relationships between the measures and users' satisfaction indicate the tradeoff effects in the gain and cost measures. Moreover, the differences between the SearchSuccess dataset and the other two datasets suggest that some performance measures have different impacts on users' satisfaction in datasets with different experimental settings, and it is critical to analyze these performance measures from multiple aspects.

To investigate user performance in search tasks with multiple cost-gain-based measures, we employed clustering analysis on those measures and grouped the search sessions. The number of clusters was set at five, determined by the Elbow method based on a tradeoff criterion between the cluster number and the mean sum of squared distances to centers (Kodinariya & Makwana, 2013). As the clustering analysis is based on multiple measures, it is difficult to

examine the session distributions on all measures directly. Therefore, we visualized cluster distributions on several projections of individual gain-cost measure pairs (e.g., ClicksRel-ClicksIrrel, ClicksUse-ClicksNonuse). Figure 2 shows examples of the cluster distributions on the projections of gain-cost pairs for the KDD19 and the TREC14 datasets. For the SIGIR16 dataset and the SearchSuccess dataset, the cluster patterns and distributions are similar to those in the KDD19 dataset. We further summarized the cluster patterns in Table 6. In general, the five clusters captured in clustering analysis can be aggregated into two groups based on the gain level, including high gain and low to medium gain. Within each group, the clusters in the four datasets share similar patterns. For example, all four datasets have one or two clusters in the high gain group, such as cluster 0 in Figure 2. These high-gain groups are based on both relevance and usefulness measures for the datasets with usefulness annotations. However, the sessions in this cluster are not distinguishable by cost-based measures, and the cost is variously distributed. For the low to median gain group, some clusters also contain a few sessions with higher gain on specific measures, such as several sessions in cluster 1 in Figure 2(a) and cluster 3 in Figure 2(c).

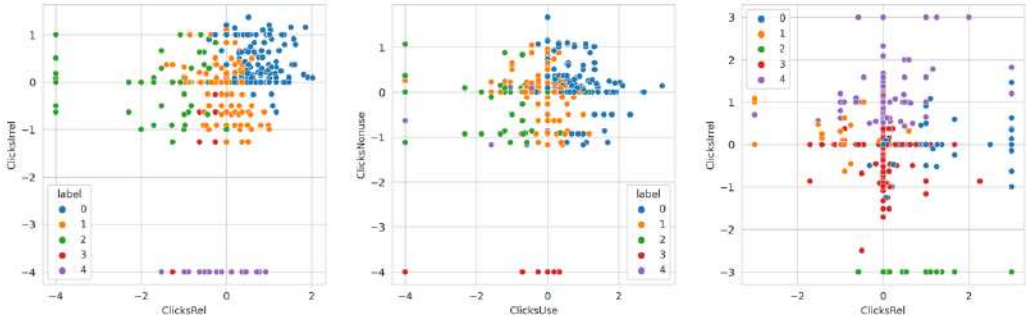
Spearman's r		KDD19	SIGIR16	SearchSuccess
Gain based	DCGrel	**-0.126****	**-0.333****	**-0.252****
	DCGuse	-0.018	-0.035	-0.001
	ClicksRel	-0.105	**-0.280****	**-0.231****
	ClicksKeyrel	0.021	**-0.220****	**-0.159***
	ClicksUse	0.08	0.027	**0.175***
Cost based	ClicksIrrel	**-0.274****	**-0.311****	-0.035
	ClicksNonuse	**-0.382****	**-0.402****	**-0.233****
	DwellIrrel	**-0.218****	**-0.278****	-0.032
	DwellNonuse	**-0.340****	**-0.275****	-0.107
	MissRel	-0.101	**-0.314****	**-0.265****

*Note: values in **boldface** indicate that the correlation is significant: * corrected p<0.05, ** corrected p<0.01).*

Table 5. Correlations between the session satisfaction and the performance measures.

Dataset	High gain		Low to medium gain	
	# Cluster	# Sessions	# Cluster	# Sessions
KDD19	1	134	4	326
Search Success	1	108	4	58
SIGIR16	1	104	4	121
TREC14	2	99	3	542

Table 6. Cluster ID and session numbers by the level of gain.



(a) KDD19 ClicksRel VS. ClicksIrrel

(b) KDD19 ClicksUse VS. ClicksNonuse

(c) TREC14 ClicksRel VS. ClicksIrrel

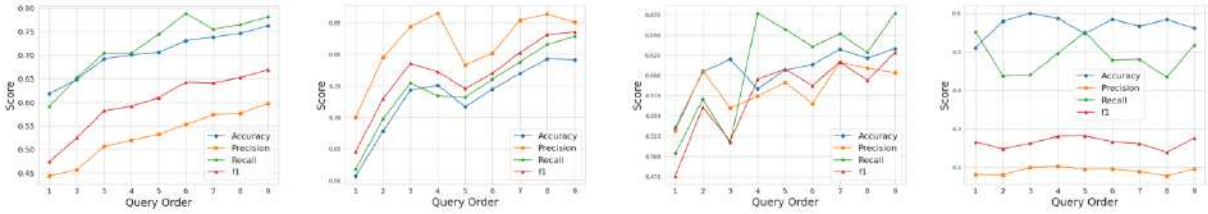
Figure 2. Examples of cluster distributions on projections of gain-cost measure pairs for the KDD19 and the TREC14 datasets: the clusters are mainly separated by the gain level, but there are fewer distinguishable patterns among clusters in terms of the cost.

Overall, the search sessions are consistently clustered by the variance of gain-based measures in all four datasets, but there are fewer distinguishable patterns among clusters in terms of cost. Although the pattern of low cost and high gain is not captured in the clustering analysis exactly, sessions with more queries and relevant or useful documents for the related task as the overall gain level can be considered *high-performance sessions*. Those high-performance sessions outperform other sessions not on single metrics but on a combination of multiple metrics (e.g., relevance-based, usefulness-based, and behavioral metrics). Although the high cost of these sessions seems to correlate with the high gain, a large number of query segments from the high-performance session cluster are still useful for collaborative query recommendation to support other users in low or medium-gain clusters, especially those with

lower estimated performance and who struggle searching for useful information. Therefore, we aggregated and labeled the clusters in the high-gain group as high-performance clusters and others in the group of low to median gain as low-performance clusters. This classification serves as the basis for following early prediction of user performance and the simulation of search recommendations.

RQ2: Early prediction of user performance

From the last step, we obtained high-performance clusters with sessions of more queries and more relevant documents and differentiated them from low-performance ones. Built upon RQ1, we developed classification models to identify high-performance sessions with user behavioral features to classify all sessions into two categories based on the cluster labels. Based on the preliminary results of the performance with different model settings, we chose feedforward neural networks as the model with the best performance. Figure 3 presents the prediction performance of neural network models at different query sequences in a session (e.g., initial query, 2nd query, 3rd query, etc.) for the four datasets. For the KDD19, SearchSuccess, and SIGIR16 datasets, the model performances gradually increased as search sessions proceeded with the increasing query order shown in Figures 3(a), 3(b), and 3(c). The elbow of the prediction point was observed at the third or fourth query, and the performance became relatively stable when the query number reached seven. This trend reflected the positive correlation between the query number and model performance for the whole session. However, during the initial few queries, the model's ability to accurately identify the performance group that a session belonged to was limited, as the variations in user behaviors were insufficient. As the session progressed, the fluctuations in user behavior became more consistent with the average values of the previous three queries.



(a) KDD19 (b) SearchSuccess (c) SIGIR16 (d) TREC14

Figure 3. Model performance at different query orders.

Other than the similar trend for the three datasets, the model's overall performance for the SearchSuccess dataset is higher than the models for the other two datasets, especially comparing the precision score. Noted that the SearchSuccess dataset contains tasks in different domain knowledge and participants from relevant domains. The higher performance indicates that user behaviors might have larger variances in the task of a specific domain than in the tasks of general topics. The TREC14 dataset showed more fluctuations with nonlinear variations and non-monotonic changes in the model's performance. These fluctuations may be caused by the different characteristics of the dataset and less consistent variance in the fewer behavioral features. There were some false-positive predictions among models, indicating that some low-performance sessions were misclassified.

In general, the developed classification models can effectively identify high-performance search sessions based on user behavioral features. The results showed that the model performance gradually increased along with the query order for most datasets, except for the TREC14 dataset. However, false-positive predictions were common among the models, highlighting the need for further evaluation of the model's performance. Therefore, we need to investigate if the sessions of predicted high performance are actually better than others with lower estimated performance for the same task. This part is explained in the next subsection.

RQ3: Simulated search recommendation

In the final stage of our research, we aimed to apply the knowledge gained from RQ1 and RQ2 to support users by employing early prediction models to simulate and evaluate search path recommendations. Our evaluation metrics included relevance-based and usefulness-based DCG, nDCG, and mRR. We compared the performance of the original query segments (Original) with that of two query recommendation methods: switching the query segments for the entire search path (Entire) and adapting each query segment based on the query intention (Adaptive). Furthermore, we compared the results based on the high-performance target labels without prediction (the Actual performance labels) with those based on early prediction results (Pred@k: with the labels predicted at the k-th query). Table 7 presents the query segment recommendation results for each dataset. The values are average scores of recommended or original query segments. Higher values for DCG or nDCG and lower values for mRR indicate better performance. In general, the results show that the recommended query segments demonstrated significant improvements over the original query segments in almost all relevance-based and usefulness-based scores with actual labels and predicted labels across four datasets. With the actual performance label, one or both recommendation methods can perform better than the original query segments. With the predicted label, some

metric scores showed greater improvements than those with actual labels. Additionally, in contrast to the increasing performance of the prediction model with more queries in RQ2, the query recommendation performance of the prediction point at the first query is similar to that of the prediction point at the second or third query. This finding highlights the potential of early prediction in supporting users.

Dataset			KDD19						SearchSuccess						
Label position	Query type	# Query	Relevance based			Usefulness based			# Query	Relevance based			Usefulness based		
			DCG	nDCG	mRR	DCG	nDCG	mRR		DCG	nDCG	mRR	DCG	nDCG	mRR
Actual	Original	555	7.02	0.85	2.93	2.67	0.67	3.78	157	3.56	0.67	5.65	0.42	0.22	9.35
	Entire	549	**7.60**	**0.87**	2.78	3.30	0.66	**3.57**	116	5.43	**0.78**	**4.62**	1.00	0.48	**5.28**
	Adaptive	542	7.19	0.86	**2.65**	**3.31**	**0.68**	3.59	154	**5.45**	0.77	4.99	**1.56**	**0.54**	5.59
Pred@1	Original	711	6.87	0.85	2.94	2.61	0.65	4.01	212	4.12	0.71	5.13	0.66	0.33	8.10
	Entire	582	**7.61**	**0.86**	**2.41**	3.20	0.68	4.22	164	5.19	0.68	9.38	1.24	0.49	5.93
	Adaptive	688	7.33	0.85	2.52	**3.77**	**0.74**	**3.11**	207	**5.54**	**0.74**	6.86	**1.96**	**0.56**	**5.64**
Pred@2	Original	453	6.82	0.85	2.87	2.57	0.64	4.00	182	3.89	0.69	5.47	0.64	0.29	8.68
	Entire	392	**7.48**	**0.87**	3.31	3.24	**0.74**	**3.17**	135	4.45	0.70	6.66	1.29	0.55	**4.62**
	Adaptive	440	7.29	0.85	**2.81**	**3.66**	0.71	3.53	179	**5.22**	**0.76**	5.69	**1.61**	**0.57**	5.22
Pred@3	Original	329	6.73	0.85	3.03	2.70	0.66	3.91	159	2.95	0.66	5.29	0.45	0.23	9.01
	Entire	282	**7.24**	**0.85**	3.32	3.03	**0.73**	**3.16**	107	4.61	0.70	6.71	1.18	0.50	**5.11**
	Adaptive	319	7.11	0.83	**2.89**	**3.70**	0.71	3.62	153	**5.41**	**0.73**	6.34	**1.62**	**0.54**	5.31
Dataset			SIGIR16						TREC14				*Note: Within the Label position column, "Actual" denotes the performance group divided based on the performance label created in RQ1, and "Pred@k" denotes the performance label predicted at the k-th query. The TREC14 dataset has no usefulness annotations, so the results only include relevance-based metrics. In the metrics columns, values in **boldfaced** indicate that the recommendation method outperforms the original query segments significantly with $p<0.05$ and has better performance than the other method.*		
Label position	Query type	# Query	Relevance based			Usefulness based			# Query	Relevance based					
			DCG	nDCG	mRR	DCG	nDCG	mRR		DCG	nDCG	mRR			
Actual	Original	283	7.06	0.93	2.58	1.70	0.55	5.24	821	0.29	0.46	8.68			
	Entire	269	8.16	**0.93**	**2.00**	**2.16**	**0.60**	**4.72**	658	0.89	**0.61**	6.93			
	Adaptive	279	**8.29**	0.92	2.93	1.84	0.51	5.58	586	**1.01**	0.53	**6.72**			
Pred@1	Original	467	7.32	0.93	2.42	1.78	0.54	5.15	605	0.31	0.43	8.70			
	Entire	388	7.48	0.89	3.41	2.08	0.55	6.85	537	0.40	0.41	8.32			
	Adaptive	443	**8.04**	0.92	2.99	**2.35**	**0.57**	5.63	509	**0.50**	0.43	**8.29**			
Pred@2	Original	339	6.91	0.93	2.44	1.56	0.52	5.31	562	0.37	0.40	8.70			
	Entire	298	7.40	0.92	2.44	1.96	0.49	6.30	477	0.48	**0.49**	8.31			
	Adaptive	310	**8.46**	0.91	2.85	**2.31**	**0.57**	**5.09**	431	**0.52**	0.43	**8.25**			
Pred@3	Original	246	6.93	0.93	2.54	1.58	0.53	5.48	376	0.42	0.37	8.56			
	Entire	218	8.20	**0.94**	**2.25**	2.51	**0.65**	**4.28**	317	0.55	0.34	8.15			
	Adaptive	227	**8.30**	0.92	2.50	**2.59**	0.64	4.29	302	**0.73**	**0.43**	**7.78**			

Table 7. Results of query recommendation simulation.

Additionally, we compared the usefulness-based and relevance-based scores of the recommended query segments and found that while the usefulness-based scores showed improvements, the relevance-based scores did not always do, as measured by nDCG and mRR in the SIGIR16 dataset. This discrepancy may be due to differences in the assessment processes used for the two types of scores. Users assess the usefulness of the documents they click on, and those with more clicks tend to have more usefulness assessments, resulting in higher usefulness-based scores and groups of high-performance sessions. Furthermore, in the SearchSuccess dataset, the relevance assessments were only for clicked documents, and this assessment may lead to bias in evaluating the query performance (e.g., users in high-performance groups might find more documents at lower ranks), affecting the improvements in relevance-based scores, particularly mRR, of recommended query segments.

DISCUSSION AND CONCLUSION

Research Findings

With the theoretical foundations of user online search performance, online search behaviors, and proactive information retrieval, we proposed a framework for evaluating users' search performance, early predicting their performance with behavioral features, and assisting low-performance users with query segment recommendations. For the three RQs, we have the following answers:

RQ1: Characterizing multidimensional search performance. We combined users' online behaviors with relevance- and usefulness-based offline metrics to measure users' search behaviors by gain and cost. Previous research usually focused on one aspect of performance in search tasks, such as effective results, efficient interactions, knowledge increment, and increases in search skills and domain knowledge (Moraveji et al., 2011; Savenkov & Agichtein, 2014; Wildemuth, 2004). Our approach is unique in that we use various features and assessments to measure the total gain and effort to complete the task, providing a more comprehensive understanding of user performance. When analyzing the clustering results, we found that users in the high-gain group had high levels of relevance and usefulness assessments with more interactions in the session, including more queries and clicks for more relevant or useful documents. It indicates that their performance requires high effort, search skills, and experience. However, we also found that most measures negatively impact users' satisfaction. This finding suggests that search systems

should focus on both search success and user satisfaction by considering multiple performance measures and optimizing the search results with minimum search effort (M. Liu, Liu, Mao, Luo, Zhang, et al., 2018).

On the other hand, we observed that some low-cost users in the medium gain group, especially the low-cost outliers, might also be considered high-performance users. These users submitted only one query in the session but clicked on relevant or useful documents with few cost interactions. They might be familiar with the task topic and can quickly find the information they need and leave the session, demonstrating high efficiency in completing the task (Kelly & Cool, 2002; Odijk et al., 2015; Shiri & Revie, 2003). However, we caution that their search strategies might not be appropriate for other users with less knowledge of the task topics.

RQ2: Model performances at different query orders. We evaluated models built with varying methods in early predicting search performance. The results indicate that the query order influences model performance in the KDD19, SearchSuccess, and SIGIR16 datasets. Based on the performance trend, it suggests predicting the third or fourth query might be the optimal prediction point. The results indicate it is feasible to predict user performance as an implicit search state and it is possible to early predict and intervene to assist low-performance users in search sessions (J. Liu et al., 2020; J. Liu & Shah, 2022; Sarkar et al., 2020). Besides, the fluctuation of model performance after the fourth query indicates that users might leave the session before the fourth query in a real search scenario and different user behavior patterns between short and long sessions. It suggests investigating the behavioral features and applying the early prediction in the short and long sessions separately (Hassan et al., 2014).

RQ3: Simulation of query segment recommendation. We simulated the collaborative query recommendation by first predicting the session group at an early point and then recommending queries from high-performance sessions for the same task. As we expected the queries from users in the high-performance sessions could help users in other sessions, and the results show consistent improvements with the actual performance labels. In addition, we chose two methods of query segment recommendation and compared varying approaches to proactive information retrieval. Our proposed method might better satisfy users' needs by recommending query segments related to their query intentions instead of recommending the entire search path as the recommender expected in previous research (Bhatia et al., 2016; Hendahewa & Shah, 2017; Sarkar et al., 2020). Overall, the improvements in the query recommendation simulation show the validity of the previous two steps in this research. The performance evaluation step can divide sessions with high performance, and the prediction step is able to predict those sessions with only behavioral features at an early point in sessions. With the last step of query recommendation, our research completes the framework of evaluating and predicting users' search performance and assisting low-performance users.

Limitations and Future Research

The study has certain limitations that suggest the need for further research. Firstly, the user search performance needs to be defined across multiple dimensions beyond the sum value of each measure. The study can modify and add new factors to discuss different aspects of search performance, such as time-based gain or cost, gain/cost ratio, and increment values. The actual user performance can be set at the tradeoff point between clearance and one hundred percent completion by adjusting the weights of measures based on a better balancing of gain and cost. Secondly, the study is based on simulated research and requires further investigation of user characteristics in different performance groups to develop new performance measures and extract useful features. Future work can involve more user performance labels as evaluation baselines, direct explicit feedback on recommended search paths, and behavioral features in ablation studies to analyze feature importance. Thirdly, the study used effective but simple query segment recommendation methods, and future research can consider more advanced methods, such as AI-enabled information retrieval. User studies can be conducted to investigate how users perform given recommended query segments in real search scenarios, and the activity path in recommended query segments can also be investigated to improve re-ranking performance. Furthermore, query recommendation needs to consider users' preferences on topically diversified results at different moments of search (J. Liu & Han, 2022).

In conclusion, this research proposed a cost-gain-based framework for evaluating users' search performance, early predicting their performance with search behavioral signals, and supporting low-performance users by query segment recommendation. The results show the feasibility of early predicting and improving user performance in search sessions. Despite the limitations, this research contributes to understanding and characterizing multiple aspects of user performance in complex search tasks. It suggests further research to improve the multidimensional evaluation of search performance, user characteristics, and proactive information retrieval methods.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation (NSF) Award IIS-2106152 and a grant from the Seed Funding Program of the Data Institute for Societal Challenges, the University of Oklahoma. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

REFERENCES

Arguello, J. (2014). Predicting Search Task Difficulty. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 8416 LNCS (pp. 88–99). https://doi.org/10.1007/978-3-319-06028-6_8

Azzopardi, L., Thomas, P., & Craswell, N. (2018). Measuring the utility of search engine result pages: An information foraging based measure. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 605–614. https://doi.org/10.1145/3209978.3210027

Bailey, E. (2017). Measuring Online Search Expertise [The University of North Carolina at Chapel Hill]. In *ProQuest Dissertations and Theses*. https://www.proquest.com/dissertations-theses/measuring-online-search-expertise/docview/1952045722/se-2?accountid=15393

Bennett, P. N., Radlinski, F., White, R. W., & Yilmaz, E. (2011). Inferring and using location metadata to personalize web search. *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 135–144. https://doi.org/10.1145/2009916.2009938

Bennett, P. N., White, R. W., Chu, W., Dumais, S. T., Bailey, P., Borisyuk, F., & Cui, X. (2012). Modeling the impact of short- and long-term behavior on search personalization. *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 185–194. https://doi.org/10.1145/2348283.2348312

Bhatia, S., Majumdar, D., & Aggarwal, N. (2016). Proactive Information Retrieval: Anticipating Users' Information Need. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 9626, pp. 874–877). https://doi.org/10.1007/978-3-319-30671-1_84

Chapelle, O., Metlzer, D., Zhang, Y., & Grinspan, P. (2009). Expected reciprocal rank for graded relevance. *Proceeding of the 18th ACM Conference on Information and Knowledge Management - CIKM '09*, 621. https://doi.org/10.1145/1645953.1646033

Chen, Y., Zhou, K., Liu, Y., Zhang, M., & Ma, S. (2017). Meta-evaluation of Online and Offline Web Search Evaluation Metrics. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 15–24. https://doi.org/10.1145/3077136.3080804

Cronen-Townsend, S., Zhou, Y., & Croft, W. B. (2002). Predicting query performance. *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 299–306. https://doi.org/10.1145/564376.564429

Dumais, S., Jeffries, R., Russell, D. M., Tang, D., & Teevan, J. (2014). Understanding User Behavior Through Log Data and Analysis. In *Ways of Knowing in HCI* (pp. 349–372). Springer New York. https://doi.org/10.1007/978-1-4939-0378-8_14

Dupret, G. (2011). Discounted Cumulative Gain and User Decision Models. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 7024 LNCS (pp. 2–13). https://doi.org/10.1007/978-3-642-24583-1_2

Fox, S., Karnawat, K., Mydland, M., Dumais, S., & White, T. (2005). Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems*, *23*(2), 147–168. https://doi.org/10.1145/1059981.1059982

Halvey, M., & Jose, J. M. (2012). Bridging the gap between expert and novice users for video search. *International Journal of Multimedia Information Retrieval*, *1*(1), 17–29. https://doi.org/10.1007/s13735-012-0007-3

Hanauer, D. A., Wu, D. T. Y., Yang, L., Mei, Q., Murkowski-Steffy, K. B., Vydiswaran, V. G. V., & Zheng, K. (2017). Development and empirical user-centered evaluation of semantically-based query recommendation for an electronic health record search engine. *Journal of Biomedical Informatics*, *67*, 1–10. https://doi.org/10.1016/j.jbi.2017.01.013

Harvey, M., Hauff, C., & Elsweiler, D. (2015). Learning by example: Training users with high-quality query suggestions. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 133–142. https://doi.org/10.1145/2766462.2767731

Hassan, A., White, R. W., Dumais, S. T., & Wang, Y.-M. (2014). Struggling or exploring? Disambiguating long search sessions. *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, 53–62.

Hassan Awadallah, A., White, R. W., Pantel, P., Dumais, S. T., & Wang, Y.-M. (2014). Supporting Complex Search Tasks. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 829–838. https://doi.org/10.1145/2661829.2661912

Hendahewa, C., & Shah, C. (2015). Implicit search feature based approach to assist users in exploratory search tasks. *Information Processing & Management*, *51*(5), 643–661. https://doi.org/10.1016/j.ipm.2015.06.004

Hendahewa, C., & Shah, C. (2017). Evaluating user search trails in exploratory search tasks. *Information Processing & Management*, *53*(4), 905–922. https://doi.org/10.1016/j.ipm.2017.04.001

Huurdeman, H. C., & Kamps, J. (2015). Supporting the Process: Adapting Search Systems to Search Stages. In *Communications in Computer and Information Science* (Vol. 552, pp. 394–404). https://doi.org/10.1007/978-3-319-28197-1_40

Kelly, D., & Cool, C. (2002). The effects of topic familiarity on information search behavior. *Proceedings of the ACM International Conference on Digital Libraries*. https://doi.org/10.1145/544220.544232

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining of cluster in K-means. *International Journal of Advance Research in Computer Science and Management Studies*, *1*(6), 90–95. https://www.researchgate.net/publication/313554124

Lazonder, A. W., Biemans, H. J. A., & Wopereis, I. G. J. H. (2000). Differences between novice and experienced users in searching information on the World Wide Web. *Journal of the American Society for Information Science*, *51*(6), 576–581. https://doi.org/10.1002/(SICI)1097-4571(2000)51:6<576::AID-ASI9>3.0.CO;2-7

Li, X., de Rijke, M., Liu, Y., Mao, J., Ma, W., Zhang, M., & Ma, S. (2021). Investigating Session Search Behavior with Knowledge Graphs. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1708–1712. https://doi.org/10.1145/3404835.3463107

Liu, F. (2016). The Search Performance Evaluation and Prediction in Exploratory Search. In *ProQuest Dissertations and Theses*. University of California, Los Angeles. https://login.pallas2.tcl.sc.edu/login?url=https://search.proquest.com/docview/1827846345?accountid=13965%0Ahttp://resolver.ebscohost.com/openurl?ctx_ver=Z39.88-2004&ctx_enc=info:ofi/enc:UTF-8&rfr_id=info:sid/ProQuest+Dissertations+%26+Theses+Global&rft_v

Liu, J., & Han, F. (2022). Matching Search Result Diversity with User Diversity Acceptance in Web Search Sessions. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2473–2477. https://doi.org/10.1145/3477495.3531880

Liu, J., Sarkar, S., & Shah, C. (2020). Identifying and Predicting the States of Complex Search Tasks. *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, 193–202. https://doi.org/10.1145/3343413.3377976

Liu, J., & Shah, C. (2019a). Interactive IR User Study Design, Evaluation, and Reporting. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *11*(2), i–75. https://doi.org/10.2200/S00923ED1V01Y201905ICR067

Liu, J., & Shah, C. (2019b). Proactive identification of query failure. *Proceedings of the Association for Information Science and Technology*, *56*(1), 176–185. https://doi.org/10.1002/pra2.15

Liu, J., & Shah, C. (2022). Leveraging user interaction signals and task state information in adaptively optimizing usefulness-oriented search sessions. *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, 1–11. https://doi.org/10.1145/3529372.3530926

Liu, J., & Yu, R. (2021). State-Aware Meta-Evaluation of Evaluation Metrics in Interactive Information Retrieval. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 3258–3262. https://doi.org/10.1145/3459637.3482190

Liu, M., Liu, Y., Mao, J., Luo, C., & Ma, S. (2018). Towards Designing Better Session Search Evaluation Metrics. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 1121–1124. https://doi.org/10.1145/3209978.3210097

Liu, M., Liu, Y., Mao, J., Luo, C., Zhang, M., & Ma, S. (2018). "satisfaction with Failure" or "unsatisfied Success": Investigating the Relationship between Search Success and User Satisfaction. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, 1533–1542. https://doi.org/10.1145/3178876.3186065

Liu, M., Mao, J., Liu, Y., Zhang, M., & Ma, S. (2019). Investigating Cognitive Effects in Session-level Search User Satisfaction. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 923–931. https://doi.org/10.1145/3292500.3330981

Liu, P., & Li, Z. (2012). Task complexity: A review and conceptualization framework. *International Journal of Industrial Ergonomics*, *42*(6). https://doi.org/10.1016/j.ergon.2012.09.001

Luo, J., Zhang, S., & Yang, H. (2014). Win-win search: Dual-agent stochastic game in session search. *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, 587–596. https://doi.org/10.1145/2600428.2609629

Mao, J., Liu, Y., Kando, N., Zhang, M., & Ma, S. (2018). How Does Domain Expertise Affect Users' Search Interaction and Outcome in Exploratory Search? *ACM Transactions on Information Systems*, *36*(4), 1–30. https://doi.org/10.1145/3223045

Mao, J., Liu, Y., Zhou, K., Nie, J.-Y., Song, J., Zhang, M., Ma, S., Sun, J., & Luo, H. (2016). When does Relevance Mean Usefulness and User Satisfaction in Web Search? *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 463–472. https://doi.org/10.1145/2911451.2911507

Moayedikia, A., Yeoh, W., Ong, K.-L., & Boo, Y. L. (2019). Improving accuracy and lowering cost in crowdsourcing through an unsupervised expertise estimation approach. *Decision Support Systems*, *122*, 113065. https://doi.org/10.1016/j.dss.2019.05.005

Moraveji, N., Russell, D., Bien, J., & Mease, D. (2011). Measuring improvement in user search performance resulting from optimal search tips. *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 355–364. https://doi.org/10.1145/2009916.2009966

Morris, M. R. (2008). A survey of collaborative web search practices. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1657–1660. https://doi.org/10.1145/1357054.1357312

Morris, M. R., & Horvitz, E. (2007). SearchTogether: An interface for collaborative web search. *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology*, 3–12. https://doi.org/10.1145/1294211.1294215

Odijk, D., White, R. W., Hassan Awadallah, A., & Dumais, S. T. (2015). Struggling and Success in Web Search. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, *19-23-Oct-*, 1551–1560. https://doi.org/10.1145/2806416.2806488

Payne, J. W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational Behavior and Human Performance*, *16*(2). https://doi.org/10.1016/0030-5073(76)90022-2

Piech, C., Spencer, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., & Sohl-Dickstein, J. (2015). Deep Knowledge Tracing. *Advances in Neural Information Processing Systems*, *2015-Janua*, 505–513. http://arxiv.org/abs/1506.05908

Raiber, F., & Kurland, O. (2014). Query-performance prediction: Setting the expectations straight. *SIGIR 2014 - Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://doi.org/10.1145/2600428.2609581

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. https://doi.org/10.1002/pra2.2016.14505301048

Salmerón, L., Cañas, J. J., & Fajardo, I. (2005). Are expert users always better searchers? Interaction of expertise and semantic grouping in hypertext search tasks. *Behaviour & Information Technology*, *24*(6), 471–475. https://doi.org/10.1080/0144329042000320018

Sarkar, S., Mitsui, M., Liu, J., & Shah, C. (2020). Implicit information need as explicit problems, help, and behavioral signals. *Information Processing & Management*, *57*(2), 102069. https://doi.org/10.1016/j.ipm.2019.102069

Savenkov, D., & Agichtein, E. (2014). To hint or not: Exploring the effectiveness of search hints for complex informational tasks. *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, 1115–1118. https://doi.org/10.1145/2600428.2609523

Shiri, A. A., & Revie, C. (2003). The effects of topic complexity and familiarity on cognitive and physical moves in a thesaurus-enhanced search environment. *Journal of Information Science*, *29*(6). https://doi.org/10.1177/0165551503296008

Smith, C. L. (2017). Domain-independent search expertise: Gaining knowledge in query formulation through guided practice. *Journal of the Association for Information Science and Technology*, *68*(6), 1462–1479. https://doi.org/10.1002/asi.23776

Suzuki, M., & Yamamoto, Y. (2021). Characterizing the Influence of Confirmation Bias on Web Search Behavior. *Frontiers in Psychology*, *12*, 132–141. https://doi.org/10.3389/fpsyg.2021.771948

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics*, *27*(1). https://doi.org/10.3102/10769986027001077

Wang, Y., Wang, L., Li, Y., He, D., Liu, T.-Y., & Chen, W. (2013). A Theoretical Analysis of NDCG Type Ranking Measures. *Proceedings of the 26th Annual Conference on Learning Theory*.

White, R. W., & Morris, D. (2007). Investigating the querying and browsing behavior of advanced search engine users. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 255–262. https://doi.org/10.1145/1277741.1277787

Wildemuth, B. M. (2004). The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology*, *55*(3), 246–258. https://doi.org/10.1002/asi.10367

Willett, P., Wildemuth, B., Freund, L., & Toms, E. G. (2014). Untangling search task complexity and difficulty in the context of interactive information retrieval studies. *Journal of Documentation*.

Yu, R., Gadiraju, U., Holtz, P., Rokicki, M., Kemkes, P., & Dietze, S. (2018). Predicting User Knowledge Gain in Informational Search Sessions. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 75–84. https://doi.org/10.1145/3209978.3210064

Zahera, H. M., El-Hady, G. F., & El-Wahed, W. F. A. (2013). Query Recommendation for Improving Search Engine Results. In *Information Retrieval Methods for Multidisciplinary Applications* (pp. 46–53). IGI Global. https://doi.org/10.4018/978-1-4666-3898-3.ch004

Zamani, H., Croft, W. B., & Culpepper, J. S. (2018). Neural Query Performance Prediction using Weak Supervision from Multiple Signals. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 105–114. https://doi.org/10.1145/3209978.3210041

Zendel, O., Shtok, A., Raiber, F., Kurland, O., & Culpepper, J. S. (2019). Information Needs, Queries, and Query Performance Prediction. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 395–404. https://doi.org/10.1145/3331184.3331253

Zhang, F., Mao, J., Liu, Y., Xie, X., Ma, W., Zhang, M., & Ma, S. (2020). Models Versus Satisfaction. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 379–388. https://doi.org/10.1145/3397271.3401162

Zhang, X., Liu, J., Cole, M., & Belkin, N. (2015). Predicting users' domain knowledge in information retrieval using multiple regression analysis of search behaviors. *Journal of the Association for Information Science and Technology*, *66*(5), 980–1000. https://doi.org/10.1002/asi.23218

Zhou, Y., & Croft, W. B. (2007). Query performance prediction in web search environments. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 543–550. https://doi.org/10.1145/1277741.1277835