State-Aware Meta-Evaluation of Evaluation Metrics in Interactive Information Retrieval

Jiqun Liu
University of Oklahoma
Norman, OK, USA
jiqunliu@ou.edu

Ran Yu
University of Bonn
Bonn, Germany
ran.yu@uni-bonn.de

ABSTRACT

In interactive IR (IIR), users often seek to achieve different goals (e.g. exploring a new topic, finding a specific known item) at different search iterations and thus may evaluate system performances differently. Without state-aware approach, it would be extremely difficult to simulate and achieve real-time adaptive search evaluation and recommendation. To address this gap, our work identifies users' *task states* from interactive search sessions and *meta-evaluates* a series of online and offline evaluation metrics under varying states based on a user study dataset consisting of 1548 unique query segments from 450 search sessions. Our results indicate that: 1) users' individual task states can be identified and predicted from search behaviors and implicit feedback; 2) the effectiveness of mainstream evaluation measures (measured based upon their respective correlations with user satisfaction) vary significantly across task states. This study demonstrates the implicit heterogeneity in user-oriented IR evaluation and connects studies on complex search tasks with evaluation techniques. It also informs future research on the design of state-specific, adaptive user models and evaluation metrics.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**.

KEYWORDS

interactive information retrieval, state-aware evaluation

ACM Reference Format:
Jiqun Liu and Ran Yu. 2021. State-Aware Meta-Evaluation of Evaluation Metrics in Interactive Information Retrieval. In *Proceedings of the 30th ACM Int'l Conf. on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3459637.3482190

1 INTRODUCTION

A large body of information retrieval (IR) research have evaluated IR system effectiveness in ad hoc retrieval tasks using a batch-style approach. In Cranfield experiments, despite the differences in task nature and users' intentions, researchers often examine the performance of IR systems with the same set of evaluation metrics (e.g. precision, recall, nDCG) across a variety of scenarios (cf. [4]). Although this standardized evaluation approach might be suitable for IR evaluations under simple factual tasks, it fails to capture the variations across different task states and information seeking intentions in complex search tasks [1, 5]. In interactive IR (IIR), users usually aim to achieve different goals (e.g. exploring a new topic, finding a specific known item) in different search iterations and thus may evaluate system performances differently [14, 16]. Without a state-aware approach, it would be difficult, if not entirely impossible, to push IR personalization forward and achieve real-time adaptive search evaluation and recommendation [2, 8, 9]. To address this gap, our work predicts users' *task states* in search sessions and *meta-evaluates* a series of online and offline metrics under varying states. For each state, we seek to identify the good-performing measure(s) that reflect users' in-situ search satisfaction.

Going beyond ad hoc retrieval tasks, we conducted our analysis on a user study dataset consisting of 1548 *query segments* from 450 *search sessions*. A *query segment* starts with a query, includes all search interactions associated with the query (e.g. browsing the search result list, clicking results, reading landing pages) and ends before the next query. A *search session* consists of all search interactions associated with a relatively complex search task and includes multiple query segments. Based on the results from analysis, we seek to answer three research questions (RQs):

- **RQ1**: What are the task states in task-based search sessions?
- **RQ2**: To what extent can we predict task states from online and offline features of varying types?
- **RQ3**: To what extent do the correlations between evaluation metrics and user satisfactions vary across different task states?

Among the three RQs, RQ1 speaks to the foundation for state-based analysis, and RQ2 serves as a feasibility analysis for applying state-aware evaluation approach in real-time search environment. With RQ1 and RQ2 being addressed, RQ3 calls for a state-aware meta-evaluation of different online and offline evaluation metrics. The contributions of our research are threefold:

- We identify task states from query and task information and develop machine learning (ML) classifiers to predict states from varying types of measures extracted from search logs.
- We meta-evaluate a series of evaluation metrics and demonstrate the heterogeneity in evaluation across different task states.
- This work connects studies on complex search tasks with IR evaluation techniques and inspires future research on design of state-specific user models and evaluation metrics in IIR.

2 DATASET AND METHOD

To answer the RQs above, we conducted analysis based on a user study dataset that consists of 450 search sessions and 1548 query

CIKM '21, November 1–5, 2021, Virtual Event, Australia.

ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3482190

segments collected under nine exploratory search tasks [15]. In addition to the search logs from which we could extract online metrics, this dataset also contains users' in-situ usefulness (or "task relevance") annotations (the contribution of the document to completing the search task) and crowdsourced relevance annotations (the topical relevance in each document-query pair). These annotations allowed us to develop offline metrics at both task and query levels. We adopted user satisfaction as the ground truth measure.

2.1 Task State Annotation

To answer RQ1, two experienced IR researchers manually annotated the task state associated with each query segment based on the state typology developed by [14]. According to [14], there are several different task states in search sessions: 1) *Exploration state*: users explores unknown topics and seeks to open new search paths. The queries tend to be general and short (e.g. *group activities weekend*); 2) *Exploitation state*: users may have a clear topic in mind and try to follow the current search path and keep exploiting the information patch at hand. At this state, users' queries tend to be focused on one topic, but not specific enough for identifying target items (e.g. *five people hiking nearby weekend*); 3) *Known-item state*: users know exactly what item(s) (e.g. location, concept, product, name) they are looking for. The queries tend to be very specific and the target item(s) are usually obvious in the queries and first visited documents (e.g. *Location of Yosemite Falls Trail*). Note that during annotation process, the two annotators found that the fourth state from [14], *Learn and Evaluate* state, does not applicable or identifiable in this dataset as all nine search tasks are factual and do not involve explicit learning processes. Also, annotating learning and evaluating activities would require fine-grained cognitive level features and explicit feedback regarding user intentions [12, 13].

Following similar approaches from past IIR research [3, 14], the two annotators annotated 10% of the data together in a back-to-back fashion in three rounds (55 unique query segments in each round), and discussed and resolved the disagreements after each round. Then, one of the IR researchers finished the annotation for the rest of the datasets. The task state labels generated in this step were used as the ground truth labels for state prediction (RQ2) and grouping variable for state-aware meta-evaluation (RQ3).

2.2 Predicting States from Searching

As the response to RQ2, we approach the problem of predicting search state with supervised models for classification. More specifically, for a given query, we classified it into one of the 3 classes as defined in Section 2.1: {exploration, exploitation, known item}. We experimented with models that are widely used in feature based classification tasks, namely, Random Forest (*rf*), Naive Bayes (*nb*), Logistic Regression (*lr*), Support Vector Machine (*svm*), K-Nearest Neighbors (*knn*) and Decision Tree (*dt*). We have also considered various Recurrent Neural Network (RNN)-based models that were applied on short texts classification (queries in our case) in prior works, however, due to the characteristics of our data and task, the performance of such models are not satisfactory, we are still exploring for a more suitable model structure for this task.

Given that task state prediction can be conducted at different steps in a search session, depending on the application scenario of the predicted search state, the available contextual information that can be used for the prediction varies. For instance, the prediction need to be done immediately after the query action of user if the search state is to be used for ranking the search result of the current query, in this case, only query related information are available. If the prediction is for supporting offline analysis, then all types of interactions and offline metrics could be considered for feature extraction. Based on the availability of features throughout the search pipeline, we group them into 3 categories as follows.

Table 1: Features considered for the search state prediction.

notation	description
query-related	
NewTerm	#new unique terms used in a issued query
QuerySim	Similarity between current query and the previous query:QuerySim = # shared UniqueTerm/# total UniqueTerm
QueryOrder	The order of the current query within the associated session
QueryLength	#terms used in an issued query
online	
session_end	is the last query segment of the session? (Y=1, N=0)
QueryDwellTime	total dwell time within the query segment
MOUSE_MOVE_count	#mouse moves
HoverCount	#mouse hovers on results
ScrollDist	total scrolling distances
ActionCount	total number of actions (e.g.#mouse hovers on results)
Clicks@{3,5,10}	#clicks among the top 3, 5, 10 results
AvgClickRank	average rank of clicked results
Time{First,LAST}Click	dwell time from query formulation to {first,last} click
ClickCount	number of clicks
TotalContent	total dwell time on content/landing pages
AvgContent	average dwell time on content pages
SERPtime	total dwell time on SERP
ClickDepth	the deepest or lowest rank of clicked result
offline	
click_precision_query	#relevant pages (query level) clicked/total #pages clicked
click_precision_task	#useful pages (task level) clicked/total #pages clicked
Query_Cost-Benefit-2	#relevant pages clicked/ClickDepth
Query_Cost-Benefit-3	(#relevant pages clicked/ClickDepth)*SERPtime
Task_Cost-Benefit-2	#useful pages clicked/ClickDepth
Task_Cost-Benefit-3	(#useful pages clicked/ClickDepth)*SERPtime
QueryRelDocCount	#relevant documents (query relevance score>0).
QueryKeyDocCount	#key documents (query relevance score>1).
TaskRelDocCount	#useful documents (task relevance/usefulness score>1).
TaskKeyDocCount	#key documents (task relevance/usefulness score>2).
QueryNDCG@{3,5,10}	NDCG from rank 1 to 3, 5, 10 (query level)
TaskNDCG@{3,5,10}	NDCG from rank 1 to 3, 5, 10 (task level/usefulness based)
QueryPrecision@{3,5,10}	The proportion of relevant pages (query level) in top {3,5,10}
TaskPrecision@{3,5,10}	The proportion of useful pages (task level) in top {3,5,10}
AveQueryRelScore	average relevance score of content pages clicked. (query level)
AveTaskRelScore	average usefulness score of content pages clicked. (task level)

- **Query-related.** This category includes features that can be computed immediately after the querying behavior of a user.
- **Online.** Features in this category are extracted based on information that are available in search system log.
- **Offline.** Offline metrics are computed based on information from the prior two categories and annotations based on external knowledge, e.g. human assessments of relevance and usefulness.

To simulate the prediction task under different scenarios, we build classifiers using following three combinations: query features only (*query*), query and online features computed based on search system log (*online+query*), and features from all categories (*all*).

2.3 State-aware Meta-Evaluation

To answer RQ3 and illustrate the heterogeneity in user-oriented search evaluation, we measured metric performance or effectiveness using *Pearson's r* between user satisfaction and each evaluation

metric under three task states. By comparing the correlation coefficients, we seek to investigate 1) how do the performances of evaluation metrics varying across different task states? 2) Under which state does each metric gain better evaluation performance?

3 RESULTS

This section reports the results from our analysis and experiments. Replication studies are essential, especially for IR evaluations [6, 11]. To facilitate this effort, we provide links to our research materials here (i.e. annotation spreadsheet[1], codes and data sources [15]).

3.1 Task States

To address RQ1, we annotated task states based on users' search interactions (e.g. query features, clicked documents, overall task context) following the procedure introduced in Section 2.1. The average annotation agreement rate between the two IR researchers was 63%. After back-to-back annotations, the two annotators revisited the annotated records, discussed the cases where they had disagreements, and resolved them by 1) further clarifying the definitions of each task state, 2) discussing the rationale behind each label, and 3) revisiting the task descriptions associated with the annotated query segments. Then, one of the annotators completed all annotation works. Among all query segments, we identified 301 exploration state segments, 514 exploitation state segments, and 733 known-item segments. Note that the specific state distributions might vary across different task types, search contexts, and datasets.

3.2 State Prediction

To answer RQ2, we experimented with the techniques introduced in Section 2.2 and present the evaluation results below.

Experimental setup. The ML models we experimented with were implemented using the Python scikit-learn library[2]. We perform 5-folds cross validation over the experimental dataset. That is, in each iteration we use 80% of the data for training and 20% of the data for testing. Accuracy was used as the scoring function during the training process. The classification performance is evaluated by computing precision, recall and F1 scores for each class, as well as their macro average across three classes, and overall accuracy.

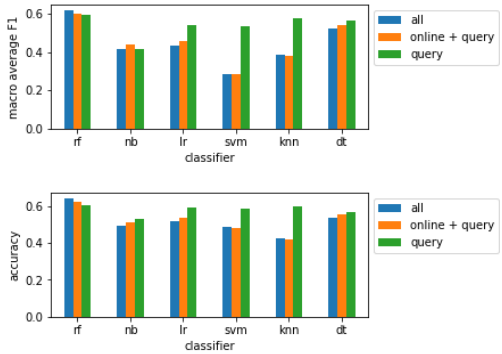

Figure 1: Task state prediction.

Classifiers and impact of feature sets. Figure 1 shows the macro average F1 score and overall accuracy of all classifiers that we have experimented with. As shown in the figure, Random Forest (*rf*) classifier has achieved the best performance in terms of both average F1 score and accuracy.

With respect to the features used, *rf* achieves better performance when using the most number of features (feature group *all*), while the majority of classifiers exhibits better results while only 4 query-related features are used. One potential reason could be that *rf* is less prone to overfitting compared to other tested classifiers. In the remainder of this section, we rely on the best performing classifier (i.e. *rf*) for further discussion. For *rf*, the overall accuracy as well as the average F1 score increases with the increasing number of features ($accu(all) > accu(online feature+query) > accu(query)$). This shows that the online and offline evaluation metrics we computed provide indicative signals for search state.

Search states. As shown in Table 2, when comparing between classes, the *known-item* class achieves the highest recall and F1 score across all feature configurations. This can be explained by the scoring function (i.e. accuracy) for model training, i.e. the class *known-item* has the largest number of samples, hence is favored when optimizing the model towards higher accuracy. The class *exploitation* result in the lowest recall among the three classes for all feature configurations, after having a closer look at the confusion matrix, we noticed that there are over 50% of the samples in class *exploitation* been classified as *know-item*. Based on this result and the experience gained from the annotation process, we found that distinguishing between *exploitation* and *known-item* is a challenging task for both human annotators and machine learning models. This indicates the necessity of developing more fine-grained task state taxonomy and corresponding ground truth dataset at larger scale.

Feature importance. The importance of all features is shown in Figure 2. The feature importance is generated through *rf* classifier based on the mean decrease in impurity (MDI) [7]. The four query-related features are ranked the highest among all features, which indicates that queries carry strong signals for tasks state detection. We consider this as a very positive finding as it suggests the possibility of applying the task state detection at very early stage of search, which enables applications such as state-aware retrieval, ranking and proactive recommendation. Two *offline* features are ranked consecutively starting at rank 5 before all *online* features, it provides evidence that these *offline* features carry stronger signals for the prediction of search state. Meanwhile, starting from position 7, the relative importance of feature category is less visible, i.e.*online* and *offline* features alternate in the list.

3.3 State-Aware Meta-Evaluation

As our response to RQ3, Table 3 presents the Pearson's correlation scores between each evaluation metric listed in Table 1 and user satisfaction under different task states. Through meta-evaluating the metrics against satisfaction score, we found that the extent to which a metric reflects user satisfaction is sensitive to the specific task state under which the metric is evaluated.

Overall, our result demonstrates that the performance of query-related and online features (measured by Pearson's r) has large variations across different task states in terms of the coefficient value, direction of the association and statistical significance. For instance, query dwell time has a significantly stronger correlation

[1] https://github.com/ran-yu/State-Aware-Evaluation-Metrics-in-IIR
[2] https://scikit-learn.org/stable/

Table 2: Performance of Random Forest classifiers.

	Exploration			Exploitation			Known-item			Macro average			
Feature categories	P	R	F1	P	R	F1	P	R	F1	P	R	F1	accu
all	**0.751**	0.581	**0.655**	**0.514**	0.399	0.449	**0.668**	**0.835**	**0.742**	**0.644**	**0.605**	**0.616**	**0.641**
online + query	0.652	0.591	0.620	**0.514**	0.405	0.453	0.664	0.789	0.721	0.610	0.595	0.598	0.623
query	0.628	**0.628**	0.628	0.498	**0.434**	**0.464**	0.657	0.716	0.685	0.594	0.593	0.592	0.605

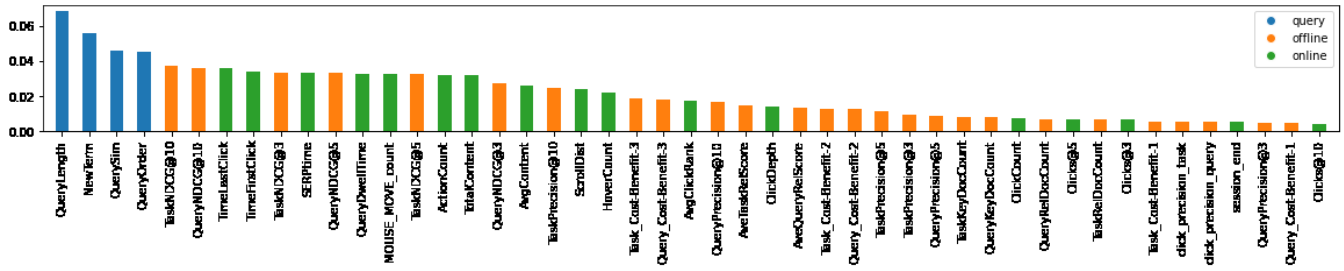

Figure 2: Feature importance.

Table 3: State-aware meta-evaluation (Pearson's r).

Eval. Metrics	Exploration	Exploitation	Known-item
NewTerm	0.095	**0.153****	**0.107****
QuerySim	-0.088	**-0.201****	**-0.084***
QueryLength	0.096	0.070	0.014
QueryDwellTime	**0.359****	**0.346****	**0.147****
Mouse_Move_count	0.110	0.007	**-0.085***
HoverCount	**0.203****	0.032	**-0.132****
ScrollDist	0.056	**-0.140****	**-0.242****
ActionCount	**0.120***	-0.058	**-0.206****
Clicks@3	**0.256****	**0.244****	**0.114****
Clicks@5	**0.322****	**0.261****	**0.107****
Click@10	0.105	**0.095***	0.005
AvgClickRank	-0.077	-0.044	**-0.161****
TimeFirstClick	0.053	0.060	**-0.088***
TimeLastClick	**-0.168****	-0.017	**-0.234****
ClickCount	**0.353****	**0.261****	**0.144****
TotalContent	**0.377****	**0.416****	**0.188****
AvgContent	**0.427****	**0.353****	**0.196****
SERPtime	0.035	-0.001	-0.011
ClickDepth	**-0.547****	**-0.432****	**-0.371****
Click_precision_query	**0.382****	**0.406****	**0.340****
Click_precision_task	**0.389****	**0.416****	**0.358****
Query_cost_benefit-2	**0.237****	**0.231****	**0.226****
Query_cost_benefit-3	0.113	**0.098***	**0.118****
Task_cost_benefit-2	**0.249****	**0.253****	**0.231****
Task_cost_benefit-3	**0.118***	**0,108***	**0.117****
QueryRelDocCount	**0.313****	**0.283****	**0.168****
QueryKeyDocCount	**0.351****	**0.358****	**0.223****
TaskRelDocCount	**0.321****	**0.297****	**0.172****
TaskKeyDocCount	**0.358****	**0.351****	**0.213****
QueryNDCG@3	**0.370****	**0.327****	**0.189****
QueryNDCG@5	**0.485****	**0.319****	**0.167****
QueryNDCG@10	**0.522****	**0.364****	**0.233****
TaskNDCG@3	**0.366****	**0.303****	**0.215****
TaskNDCG@5	**0.485****	**0.303****	**0.216****
TaskNDCG@10	**0.523****	**0.363****	**0.279****
QueryPrecision@3	**0.442****	**0.370****	**0.122****
QueryPrecision@5	**0.504****	**0.416****	**0.135****
QueryPrecision@10	**0.531****	**0.445****	**0.166****
TaskPrecision@3	**0.407****	**0.437****	**0.207****
TaskPrecision@5	**0.506****	**0.477****	**0.251****
TaskPrecision@10	**0.534****	**0.504****	**0.267****
AvgQueryRelScore	**0.492****	**0.515****	**0.400****
AvgTaskRelScore	**0.483****	**0.496****	**0.395****

*:p<.05, **:p<.01. statistically significant correlations with user satisfaction are boldfaced. The strongest correlation under each metric is highlighted in gray.

with user satisfaction under exploration state, compared to the result from known-item state. Similar patterns were also found under other behavioral features, such as *ClickCount*, *AvgContent*, and *ClickDepth*. In addition, although the number of mouse hovering and total actions are positive indicators of user satisfaction under exploration state, they negatively correlate with satisfaction level under known-item state. Also, there are a variety of online evaluation metrics that are significantly associated with user satisfaction under only one state, such as *TimeFirstClick*, *AvgClickRank*, and *Click@10*, suggesting that users may not be sensitive to these actions or "costs" under the other two task states. These intrinsic heterogeneity would not have been captured if we did not evaluate the metrics under separate task states.

Compared to the metrics above, offline evaluation metrics in general have stronger correlations with user satisfaction, with a variety of statistically significant coefficient scores being greater than 0.5. This echoes the results from [4]'s IR evaluation experiments where researchers found strong associations between offline metrics and user satisfaction in homogeneous search (ten blue links). In addition, our result indicates that the offline metrics (especially the query- and task-level nDCG and precision measures) have close associations with user satisfaction in exploration state. In contrast, these metrics are not quite effective in the situations where users know exactly what they are looking for. Also, under exploitation state, we noticed that *AvgQueryRelScore*, *AvgTaskRelScore* and *Click_precision* as relatively simple measures tend to better correlate with user satisfaction compared to other complex evaluation metrics. The findings discussed above can enhance our understanding of the implicit variations and heterogeneity in IR evaluation.

4 CONCLUSIONS AND FUTURE WORK

Users often seek to achieve different goals at different search moments and thus may evaluate system performances differently. Going beyond existing research on IR meta-evaluation [4, 10], our research highlights the heterogeneity in user-oriented IR evaluation and demonstrates the importance of developing state-aware approach to modeling and evaluating search interactions. Our future research will 1) develop a more fine-grained task state taxonomy that could be easily used by non-IR experts, 2) explore new cognitive-level features for state prediction, and 3) construct new metrics that better reflect user satisfaction under certain states.

5 ACKNOWLEDGMENT

This work is partially supported by the National Science Foundation (NSF) grant IIS-2106152.

REFERENCES

[1] Nicholas J Belkin. 2016. People, Interacting with Information. In *ACM SIGIR Forum*, Vol. 49. ACM New York, NY, USA, 13–27.

[2] Nicholas J Belkin, Daniel Hienert, Philipp Mayr, and Chirag Shah. 2018. Data requirements for evaluation of personalization of information retrieval-a position paper. *arXiv preprint arXiv:1809.02412* (2018).

[3] Robert Capra, Jaime Arguello, and Falk Scholer. 2013. Augmenting web search surrogates with images. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 399–408.

[4] Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. Meta-evaluation of online and offline web search evaluation metrics. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*. 15–24.

[5] Michael Cole, Jingjing Liu, Nicholas Belkin, Ralf Bierig, Jacek Gwizdka, C Liu, Jin Zhang, and X Zhang. 2009. Usefulness as the criterion for evaluation of interactive information retrieval. *Proc. HCIR* (2009), 1–4.

[6] Nicola Ferro and Diane Kelly. 2018. SIGIR initiative to implement ACM artifact review and badging. In *ACM SIGIR Forum*, Vol. 52. ACM New York, NY, USA, 4–10.

[7] Hong Han, Xiaoling Guo, and Hua Yu. 2016. Variable selection using mean decrease accuracy and mean decrease gini based on random forest. In *2016 7th ieee international conference on software engineering and service science (icsess)*. IEEE, 219–224.

[8] Jiepu Jiang and James Allan. 2016. Adaptive effort for search evaluation metrics. In *European Conference on Information Retrieval*. Springer, 187–199.

[9] Gareth JF Jones, Nicholas J Belkin, Séamus Lawless, and Gabriella Pasi. 2018. Report on the CHIIR 2018 Workshop on Evaluation of Personalisation in Information Retrieval (WEPIR 2018). In *ACM SIGIR Forum*, Vol. 52. ACM New York, NY, USA, 129–134.

[10] Suneel Kumar Kingrani, Mark Levene, and Dell Zhang. 2018. A meta-evaluation of evaluation methods for diversified search. In *European Conference on Information Retrieval*. Springer, 550–555.

[11] Jiqun Liu. 2021. Deconstructing search tasks in interactive information retrieval: A systematic review of task dimensions and predictors. *Information Processing & Management* 58, 3 (2021), 102522.

[12] Jiqun Liu and Yong Ju Jung. 2021. Interest Development, Knowledge Learning, and Interactive IR: Toward a State-based Approach to Search as Learning. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. 239–248.

[13] Jiqun Liu, Matthew Mitsui, Nicholas J Belkin, and Chirag Shah. 2019. Task, information seeking intentions, and user behavior: Toward a multi-level understanding of web search. In *Proceedings of the 2019 ACM SIGIR Conference on human information interaction and retrieval*. 123–132.

[14] Jiqun Liu, Shawon Sarkar, and Chirag Shah. 2020. Identifying and Predicting the States of Complex Search Tasks. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 193–202.

[15] Mengyang Liu, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Investigating cognitive effects in session-level search user satisfaction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 923–931.

[16] Matthew Mitsui, Jiqun Liu, Nicholas J Belkin, and Chirag Shah. 2017. Predicting information seeking intentions from search behaviors. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1121–1124.