Information Retrieval Journal (2023) 26:10
https://doi.org/10.1007/s10791-023-09426-1

Constructing and meta-evaluating state-aware evaluation metrics for interactive search systems

Marco Markwald[1] · **Jiqun Liu**[2] · **Ran Yu**[1,3]

Received: 7 November 2022 / Accepted: 28 September 2023 / Published online: 31 October 2023
© The Author(s) 2023

Abstract
Evaluation metrics such as precision, recall and normalized discounted cumulative gain have been widely applied in *ad hoc* retrieval experiments. They have facilitated the assessment of system performance in various topics over the past decade. However, the effectiveness of such metrics in capturing users' in-situ search experience, especially in complex search tasks that trigger interactive search sessions, is limited. To address this challenge, it is necessary to adaptively adjust the evaluation strategies of search systems to better respond to users' changing information needs and evaluation criteria. In this work, we adopt a taxonomy of search task states that a user goes through in different scenarios and moments of search sessions, and perform a meta-evaluation of existing metrics to better understand their effectiveness in measuring user satisfaction. We then built models for predicting task states behind queries based on in-session signals. Furthermore, we constructed and meta-evaluated new state-aware evaluation metrics. Our analysis and experimental evaluation are performed on two datasets collected from a field study and a laboratory study, respectively. Results demonstrate that the effectiveness of individual evaluation metrics varies across task states. Meanwhile, task states can be detected from in-session signals. Our new state-aware evaluation metrics could better reflect in-situ user satisfaction than an extensive list of the widely used measures we analyzed in this work in certain states. Findings of our research can inspire the design and meta-evaluation of user-centered adaptive evaluation metrics, and also shed light on the development of state-aware interactive search systems.

Keywords Information retrieval · Search state · Evaluation metrics · Search state prediction

✉ Ran Yu
ran.yu@uni-bonn.de

Marco Markwald
marco.markwald@cs.uni-bonn.de

Jiqun Liu
jiqunliu@ou.edu

[1] University of Bonn, Bonn, Germany

[2] The University of Oklahoma, Norman, OK, USA

[3] Lamarr Institute for Machine Learning and Artificial Intelligence, lamarr-institute.org, Bonn, Germany

1 Introduction

The *batch-style* evaluation approach integrated with one or two unified evaluation metrics (e.g. precision, recall, normalized discounted cumulative gain) has been widely applied in a large body of *ad hoc* retrieval tasks and evaluation experiments (Chen et al., 2017; Harman, 2011). While employing one unified measure across different queries may facilitate the comparison of system performance in varying topics and search contexts, it may undermine the effectiveness of evaluation experiments in capturing users' actual search experiences, especially in prolonged, interactive search sessions (Cole et al., 2009). When engaging in *complex search tasks* that involve ill-defined, ambiguous goals, users often go through varying *cognitive states*, seek to fulfill different *search intentions*, and thereby evaluate system performances differently under varying queries (Liu et al., 2020; Sarkar et al., 2020). Under these circumstances, intelligent search systems will need to adaptively adjust the evaluation and re-ranking strategies to better respond to users' changing information needs and search obstacles under overarching motivating tasks.

To achieve this, researchers need to deploy and meta-evaluate *state-aware* evaluation metrics that can reliably connect the evaluation of search system performance with users' actual experiences in search interactions and partially address the limitations of traditional offline evaluation procedures (Liu, 2022). Furthermore, it is critical to design and implement new evaluation metrics that can achieve better performance than existing metrics in capturing search satisfaction levels under specific states and search scenarios. The knowledge and techniques learned through state-aware evaluation research and practice will allow researchers to better capture the nuances hidden in the cognitive process behind IR and to develop more fine-grained user models to support adaptive ranking and search recommendations.

To address the research gap discussed above, our study sought to predict the varying *task states* that a user goes through in different complex search tasks based on observable search signals and to identify appropriate evaluation metrics that best reflect user satisfaction under each state. Taking a step forward, we also explored and meta-evaluated new evaluation metrics that could better reflect in-situ user satisfaction than all existing measures. To obtain robust and potentially generalizable evaluation results, our meta-evaluation experiments were conducted based on datasets collected from both controlled lab and naturalistic search settings.

Going beyond traditional multi-system evaluation setups built upon one or two cross-session unified metrics, our study connects IR evaluation to dynamic task states and makes the following contributions:

- Our study demonstrates the relationships between in-situ user satisfaction and different evaluation metrics, and shows that the best-performing metrics vary across different task states within individual task-based search sessions.
- Based on observable search behavior and textual features that can be collected from the backend, we developed machine learning-based classifiers that can predict users' task states during search sessions. These predictive models can serve as the basis for adaptive and even proactive search recommendations and evaluations.
- In addition to comparing and evaluating existing metrics under different states, we also developed new evaluation metrics that can outperform all current metrics under certain task states. Our new metrics could be replicated and reused in a wider range

of search evaluation scenarios and contribute to the enhancement of human-centered IR evaluations.

2 Related works

Many evaluation metrics have been widely used throughout the years, the metrics were built under various assumptions and assessment goals, such as search result relevance, user perceived usefulness, user satisfaction. Based on the information needed for computing the evaluation metrics, we can categorize existing metrics into (1) online metrics, which can be computed based on system log files containing user interaction records, and (2) offline metrics, which rely on external knowledge such as human annotations. In this paper, we investigate a wide range of common evaluation metrics that can be computed on our experimental dataset. The full list of metrics is presented in Sect. 5. In this section, we give an overview of the existing works investigating the relationship between common evaluation metrics, user satisfaction and search states.

2.1 Task states in interactive search sessions

In contrast to the simulated scenarios of ad hoc retrieval tasks, users engaged in complex search tasks often experience the transitions between *task states* and aim to fulfill different *subgoals* or *intentions* at different moments of a search session (Liu et al., 2020; Rha et al., 2016). Mitsui et al. (2016) examined users' search intentions associated with different queries in the same task and developed behavior-based prediction models for identifying users' intentions in real-time. Chen et al. (2021) found that user reformulation is closely related to user intent and incorporated this knowledge into click-based metrics, improving the correlation with user satisfaction. Vuong et al. (2019) introduced a categorization of queries by intention, task goal, and task substance. Similarly also Järvelin et al. (2015) looked at the different types of tasks and suggested that this should be included in IR. Borlund (2016) investigated which type of information is needed in which type of task. Liu et al. (2019b) developed a multilevel model of task-based information seeking and found that users' search tactics and document judgments vary significantly across different intentions and task types. There is also some research investigating the use of user models to improve the correlation with user satisfaction (Moffat et al., 2022; Wicaksono & Moffat, 2020, 2021; Zhang et al., 2020b). Researchers have also extracted four task states, i.e. exploration, exploitation, known-item and evaluation, from participants' in-situ intention annotations, studied the transition patterns between different task states under complex tasks of different types (Liu et al., 2020), and developed state-aware search path recommendation algorithms that can improve the efficiency of search interactions in finding useful information (Liu & Shah, 2022). Urgo and Arguello (2022) applied a state-based approach to investigate the Search as Learning (SAL) process (apply, evaluate and create) and characterized the transitions between different knowledge types (factual, conceptual and procedural) during search. In addition to user-centered task modeling and evaluation, researchers have also adopted a similar state-based approach in offline simulation-based studies and demonstrated the value of leveraging task state information in improving relevance-based ranking performance (Luo et al., 2014).

Since users typically go through multiple task states and intentions during the same search session (Liu et al., 2020; Ruotsalo et al., 2014), the evaluation of search systems

should also be adaptive and customized based on the nature of local task states rather than relying on one or two unified measures across all search queries (e.g., nDCG, Reciprocal Rank) (Liu & Han, 2022). It is also unclear how and to what extent users' criteria and thresholds of usefulness and satisfaction vary across states. To address these gaps, our study seeks to investigate heterogeneity across task states and construct *state-aware* evaluation metrics that best reflect users' in-situ levels of search satisfaction under each state, rather than simply optimizing predefined document relevance metrics. The implementation and meta-evaluation of adaptive evaluation measures will also facilitate the development and evaluation of personalized IR systems and search recommendations.

2.2 Understanding and measuring user satisfaction

User satisfaction has been described in many papers as the golden standard for evaluating the quality of search results (Chen et al., 2017; Jiang et al., 2015; Zhang et al., 2018, 2020c). Many studies have investigated the factors that affect user satisfaction. For example, Jiang et al. (2015) concluded from their study that satisfaction can best be explained as the value of the search outcome compared to the degree of search effort. Liu et al. (2015) investigate whether there is a difference between assessors' and users' judgments of satisfaction. They find that assessors' and users' judgments are moderately correlated. Liu et al. (2018) investigated the differences between user satisfaction and search success for complex search queries. Their experiments indicate that there is a high discrepancy between user satisfaction and search success. These previous studies demonstrate that user satisfaction can be influenced and reflected by many user and system features.

As attempts to measure user satisfaction, existing works study relationships between various metrics and user satisfaction. Chen et al. (2017) conducted a meta-evaluation of a set of existing online and offline metrics on datasets collected from task-based lab studies to study how they correlate with user satisfaction. They found that offline metrics are better aligned with user satisfaction in homogeneous search, while online metrics outperform when vertical results are federated. Zhang et al. (2020a) found that task difficulty influences the correlation between metrics and satisfaction. This shows that how well existing metrics reflect satisfaction varies by task type. Chuklin and de Rijke (2016) developed the CAS model, which combines user clicks and attention behavior on a SERP to capture user satisfaction. Mao et al. (2016) attempted to use expert-annotated usefulness to measure user satisfaction and found that usefulness is strongly correlated with user satisfaction. This observation was also confirmed by Liu et al. (2019b). However, usefulness annotations are query dependent and subjective, the annotation of the same resource cannot be generalized to other queries or users, thus it is not an efficient metric for measuring satisfaction.

Despite efforts to measure user satisfaction through other metrics, existing works have not succeeded in finding an effective and efficient approach. Based on our investigation in existing works and datasets, we found the following potential reasons: previous works that attempted to measure user satisfaction did not consider the impact of task states; evaluation metrics created by combining existing metrics were fitted to a set of homogeneous data and investigated only a small set of features. Therefore, in this work, we investigate a larger set of features and take task state into account when analyzing existing metrics and fitting new evaluation metrics. In addition, we experiment on both lab study data and field study data to observe the impact of different data collection setups.

3 Task definition and research questions

In this paper, we consider the search task state for each individual query activity, which is defined by the sequence of a user's actions starting with querying the Web, followed by browsing the search results, browsing the clicked Web resources, and clicking and scrolling activities.

3.1 Evaluation metrics

To ensure a fair and thorough analysis, we researched the most commonly used evaluation metrics for general IR systems. Based on the availability of the data needed to calculate each metric throughout the search process, we grouped them into 3 categories as follows:

- *Query-based metrics:* Metrics that can be calculated immediately after a search query is executed.
- *Online metrics:* Metrics that can be calculated based on system log files that record user interactions (e.g. mouse movements, clicks, timestamps of interactions, etc.). Query-based metrics are a subset of online metrics.
- *Offline metrics:* Metrics that rely on external knowledge, such as annotations based on human judgement, e.g., the relevance score of web documents in a search results list.

The full list of metrics we analyze in this paper and their descriptions are given in Sect. 5.

3.2 Task states

Among the taxonomies discussed in Sect. 2.1, we found the taxonomy proposed by Liu et al. (2020) to be the most appropriate for the analysis in this paper, as it has been conceptually developed and empirically validated with both external labeling and clustering results under *complex search tasks* involving prolonged search sessions and covering task states and user intentions of varying complexity at different moments of search. Also, compared to existing taxonomies, Liu et al. (2020)'s task state taxonomy achieves a better balance between capturing the nuances of user intentions and being practically useful in participants' annotations. The taxonomy does not involve overly abstract or broad categories (cf. informational queries in Broder's taxonomy (Broder, 2002)) and distinguishes different search focuses or task states (e.g., exploring a new topic or domain versus evaluating collected information items) without requiring a detailed, cognitively challenging annotation process (cf. Rha et al. (2016)'s taxonomy).

Based on the labels and clustering results from two controlled lab studies, Liu et al. (2020) found that a querying activity can be assigned to one of the following 4 states:

- *Exploration state:* The user wants to explore an unknown topic in this state. He uses general and short queries (e.g. "sports activities").
- *Exploitation state:* The user knows exactly what topic he is looking for in this state. He follows his search path and looks for different pages that might provide relevant information (e.g."Football pitch nearby").

- *Known-item state:* The user knows exactly what his goal is and is looking for a specific page or information (e.g. "Location of the Football pitch in Friesdorf").
- *Learn and evaluate state:* In the fourth state, the user not only wants to passively absorb information, but also wants to evaluate search results or expand his knowledge. As in the known-item state, the user is looking for specific information (e.g., "Difference between the soccer fields in Friesdorf and Bornheim").

3.3 User satisfaction

In this work, the level of user satisfaction refers to the extent to which a system informationally satisfies a user's search goal(s) under the associated task. The satisfaction scores we use in this work are annotated by users directly.

We aim to investigate state-aware evaluation metrics for search systems in terms of user satisfaction. We approach the problem by answering the following research questions:

RQ1: To what extent do existing evaluation metrics reflect user satisfaction under varying task states?
RQ2: Can we detect the task state of a query activity using in-session signals that can be collected automatically during a search session without explicit feedback and labels?
RQ3: Can we construct new evaluation metrics that better reflect user satisfaction under a specific task state?

4 Experimental data

4.1 Datasets

In order to study the characteristics of more diverse search sessions and to improve the generalizability of the results of this work, we consider two datasets collected under different setups: one from the field study (*TianGong*) and one from a lab study (*KDD*).

4.1.1 TianGong

This dataset was published by Zhang et al. (2020c). The authors conducted a field study that lasted for one month with 30 participants (13 females and 17 males) whose ages ranged from 18 to 41. The participants installed a browser extension to track their search activities. The Participants rate their satisfaction with the search result for each search query on a 5-point Likert scale, with 0 for dissatisfied and 4 for very satisfied. After the study, nine external annotators rated the relevance of the documents with respect to the corresponding query on a scale of 0–3, where 0 means a document is irrelevant and 3 means a document is very relevant. We use *TianGong* to refer to this dataset in the rest of this paper for simplicity.

The *TianGong* dataset contains 3875 queries. Each query is associated with logs containing the query string, the corresponding SERP, mouse movements (clicks and scrolls), switching between SERP and other browsed pages, and the corresponding timestamps or dwell times of the above activities. On average, 55 actions are recorded in the search logs for each query.

4.1.2 KDD19

This dataset was collected in a laboratory study and was published by Liu et al. (2019b). Fifty undergraduate students (24 female, 26 male) were recruited from the campus, ranging in age from 18 to 27. All participants were familiar with the basic use of web search engines and used them on a daily basis. Nine search tasks were given to each participant. Similar to the *TianGong* dataset, participants rated their satisfaction with the search result corresponding to each search query on a 5-point Likert scale, with 1 being dissatisfied and 5 being very satisfied. To obtain the relevance ratings for each document, the authors used a crowdsourcing platform. Each crowd worker was given a "query-document" pair. Then they were asked to assign a relevance score (0–3) to each document, 0 if they think the document is not relevant or a spam webpage, 1 if there is only a small amount of information in the document related to the query, 2 if there is important information related to the query in this document, 3 if the document should be a top result in the SERP because the content is dedicated to the query. A total of 1548 queries with search logs were recorded. On average, there are 188 actions per query. We use *KDD* to refer to this dataset in the rest of this paper for simplicity.

The distribution of the annotated search satisfaction and document relevance of the two datasets is shown in Fig. 1. The difference in the study setup can potentially explain the

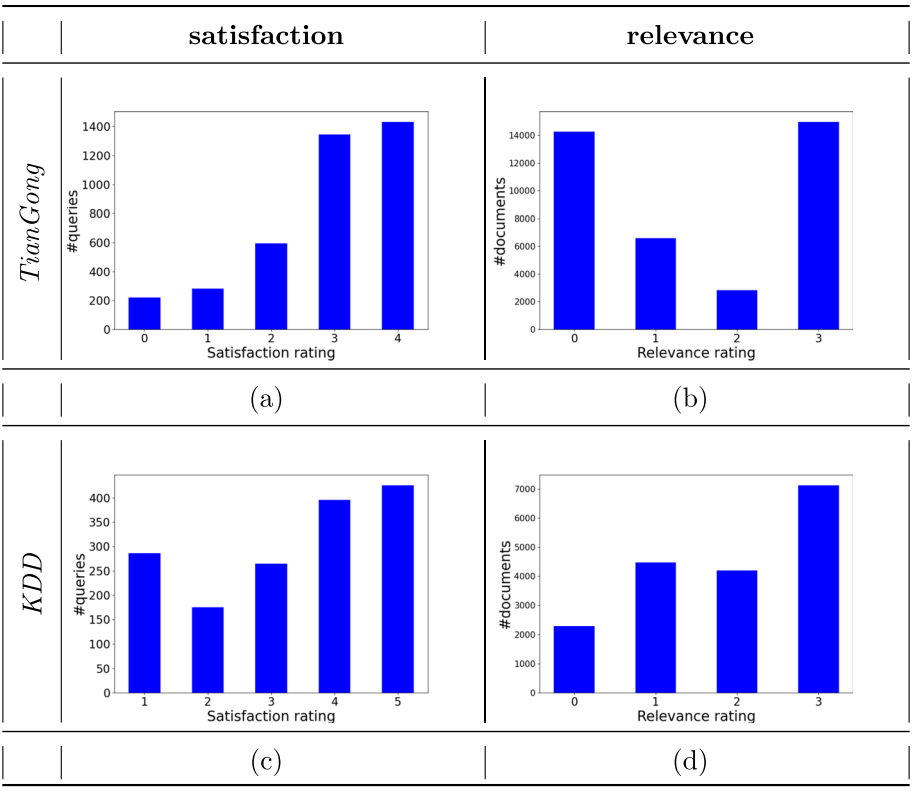


Fig. 1 Distribution of user satisfaction (query level) and document relevance (document level) based on existing annotations in the TianGong and KDD19 datasets

difference in the relevance and satisfaction distribution of the two datasets, i.e. the *TianGong* dataset has a higher percentage of irrelevant documents while having higher satisfaction. In the field study, where the tasks are not clearly defined, users are likely to be satisfied if some relevant information can be found with the self-formulated queries, while in the lab study, with a clear task in mind and queries that can be extracted from the task description, more relevant web resources are recalled, but users are not satisfied as long as the current task goal is not completed.

4.2 Task state annotation

The task state of the *KDD* dataset was published by Liu and Yu (2021). We applied the same coding frame used in their paper to the *TianGong* dataset. More specifically, the task states used in the annotation task, including: (1) Exploration state—users explore unknown topics and seek to open new search paths; (2) Exploitation state—users may have a clear topic in mind and try to follow the current search path and continue to exploit the information patch at hand; (3) Known-item state—users know exactly what item(s) they are looking for. Queries tend to be very specific, and the target item(s) are usually obvious in the queries and the first documents visited; (4) Learning and evaluation state—users try to evaluate, extract and synthesize useful knowledge from retrieved documents and pages. At this state, they tend to have long, specific queries involving multiple subtopics and items, and move between and compare multiple documents. Two annotators annotated a subset of the data together in three rounds (100 unique queries in each round), discussing and resolving disagreements after each round. In the second and third round of annotation, the agreement between the two annotators in each round is both above 70%, the Cohen's Kappa is above 0.559. Then one of the annotators finished the annotation for the rest of the dataset. The distributions of the task state labels in the two datasets are shown in Fig. 2.

We found that the distribution of task states is unbalanced in both datasets. In the *TianGong* dataset, the exploitation state has the highest number of queries. In the *KDD* dataset, there are more known-item searches compared to other states, which could be due to the setup of the lab study, where search tasks are given and therefore the goals are more straightforward compared to natural search sessions. Similar observations can be made on the *TianGong* dataset as Liu and Yu (2021) made on the *KDD* dataset, that the last two states are hardly distinguishable based on the queries. Therefore, we use the same approach to merge the known-item and evaluation states in the experiments, and refer to the merged

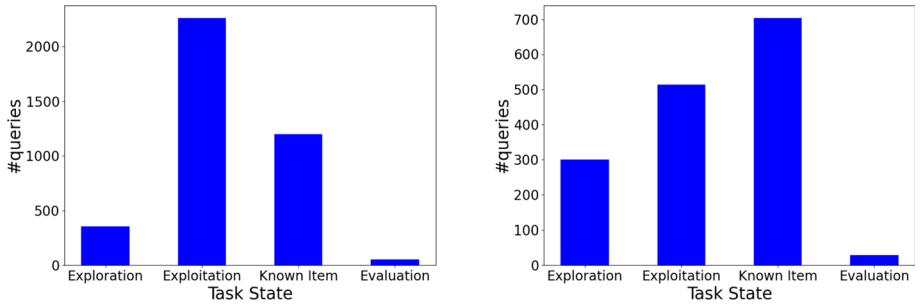


Fig. 2 Task states distribution in TianGong and KDD datasets

Table 1 Query-based metrics

Notation	Description
NewTerms	Number of new terms in a query that are not in the previous query
QuerySim	Proportion of shared terms with last query
QueryOrder	The order of the current query within the associated session
QueryLength	Number of terms in a query

Table 2 Online metrics

Notation	Description
ActionCount	Number of actions associated with the query
MouseMoveCount	Total number of mouse moves
ScrollDistanceance	Total scroll distance
MaxScroll	Maximum of scroll distance for query
TimeTo{First, Last}Click	Dwell time to first/last click in milliseconds
#Clicks	Total number of clicks for query
{Highest, Avg, Lowest} ClickRank	Highest/average/lowest rank of clicked result
{Max, Min, Mean}RR	Maximum/minimum/mean for the reciprocal ranks of the clicks
PLC	#Clicks divided by LowestClickRank
SessionEnd	If the query is the last query in session
TotalContentTime	Total time on landing and content pages in milliseconds

state as known-item in later sections. Finally, we obtain the ground truth label of task states for the two datasets, where there are 301, 514, 733 queries in the *KDD* dataset and 356, 2266, 1252 queries in the *TianGong* dataset for the exploration, exploitation, and known-item states, respectively.

5 Analysis of existing evaluation metrics

In this section, we will investigate *RQ1* using the annotated data as described in Sect. 4 to see how well existing evaluation metrics reflect user satisfaction and explore whether there are differences for task states.

5.1 Evaluation metrics

As described in Sect. 3, we group the metrics into three categories according to the availability of information for their computation. The query-based category includes features that can be computed immediately after a user's query behavior; the considered metrics and their descriptions are shown in Table 1. The list of online metrics that are extracted based on information in the search system log is shown in Table 2. Offline metrics are listed in Table 3. Note that we focus on query level evaluation, so all features are computed for each individual query. For query-based features, the terms appeared in previous queries and the order of the query has been considered as contextual information for the feature calculation of the current query.

Table 3 Offline metrics

Notation	Description
MaxR, MeanR, MinR	Maximum, Mean and Minimum value of relevance for result pages in query
CG@k	Cumulative gain, $k \in \{3, 5, 10\}$
DCG@k	Discounted cumulative gain, $k \in \{3, 5, 10\}$
NDCG@k	Normalized discounted cumulative gain, $k \in \{3, 5, 10\}$
RBP(x)	Rank-biased precision, $x \in \{0.1, 0.5, 0.8, 0.95\}$
ERR	Expected reciprocal rank
Precision@k	Proportion of relevant pages in top k ranked results, $k \in \{3, 5, 10\}$
AvgClickRel	Average relevance of clicked pages
ClickPrecision	Proportion of clicked relevant pages
QueryCost-Benefit-2	Number of clicked relevant pages/LowestClickRank
QueryCost-Benefit-3	(Number of clicked useful pages/LowestClickRank)*SERPtime
RelDocCount1	Number of documents with query relevance > 1
RelDocCount2	Number of documents with query relevance > 2

5.2 Correlation analysis

To understand the relationship between the evaluation metrics introduced in Sect. 5.1, we compute the Pearson correlation between the user satisfaction score and each individual metric on both datasets for each task state. The results are shown in Table 4.

5.2.1 Results on TianGong dataset

The results indicate that the metrics perform differently under different task states, there is no single metric that achieves the highest correlation across all task states. The gap between different metrics is quite large, with the highest correlations being 0.422, 0.265, 0.267, and the median correlations being 0.252, 0.192, 0.194 on the *exploration*, *exploitation*, and *known-item* states, respectively. In terms of the metrics that have the highest correlation with user satisfaction in each task state, *MaxR* achieved the highest scores in the exploitation and known-item states, as in these two states users have a clearer goal in mind and are likely to be satisfied by the most relevant result. While for the exploration state, users consider more search results to get a better overview of the topic, therefore *RBP* and *DCG* based metrics achieve the highest correlation as they consider the relevance and rankings of multiple results.

The offline metrics generally have a higher correlation with user satisfaction in the *exploration* state than in the other task states. For example, for the CG@3 we have a correlation of 0.405 in the *exploration* state and only 0.232 and 0.233 for the other two task states. This is probably because users in the *exploration* state have less prior knowledge about the topic and are satisfied if documents of general relevance are returned. On the other hand, a user in the *known-item* state knows exactly what he is looking for. For example, if a user is only looking for a specific formula, all search results are likely to have higher relevance scores and would have less impact on user satisfaction, as other factors such as efficiency and quality of the web resource may be more important.

There is no clear pattern for online and query metrics. For example, PLC has the highest correlation with user satisfaction in the *known-item* state, while SessionEnd has the highest

Table 4 Correlation analysis results

	Metric	TianGong				KDD			
		All	Exploration	Exploitation	Known-item	All	Exploration	Exploitation	Known-item
Query	NewTerms	−0.007	−0.120*	−0.003	−0.020	*0.094***	0.087	0.166**	0.040
	QuerySim	−*0.119***	−0.042	−0.102**	−0.108**	−0.086**	−0.065	−*0.195***	−0.069
	QueryOrder	0.077**	0.051	*0.148***	*0.147***	−0.072**	−0.072	−0.012	−*0.204***
	QueryLength	−0.052**	−*0.116**	−0.052	−0.078**	0.064*	*0.097*	0.070	0.008
Online	ActionCount	−0.130**	−0.096	−0.155**	−0.157**	−0.088**	0.121*	−0.058	−0.202**
	MouseMoveCount	−0.093**	−0.069	−0.122**	−0.125**	−0.017	0.110	0.007	−0.076*
	ScrollDistance	−0.167**	−0.113*	−0.168**	−*0.164***	−0.140**	0.109	−0.081	−*0.279***
	Max Scroll	−*0.184***	−0.158**	−*0.193***	−*0.187***	−0.100**	0.133*	−0.059	−0.208**
	TimeToFirstClick	−0.054**	0.008	−0.113**	−0.109**	−0.006	0.167**	−0.033	−0.034
	TimeToLastClick	0.039*	0.098	0.045	0.043	0.152**	0.138*	0.274**	0.115**
	#Clicks	0.074**	0.169**	0.057*	0.052	0.213**	0.336**	0.268**	0.118**
	AvgClickRank	−0.030	0.031	−0.105**	−0.101**	0.059*	0.113	0.149**	−0.032
	LowestClick Rank	−0.020	0.060	−0.100**	−0.095**	0.108**	0.230**	0.179**	0.004
	HighestClick Rank	−0.006	0.101	−0.053	−0.050	0.077**	0.157**	0.085	0.030
	MaxRR	0.170**	0.176**	0.139**	0.142**	0.288**	*0.394***	0.328**	0.199**
	MinRR	*0.208***	*0.210***	*0.196***	0.200**	*0.338***	*0.432***	*0.356***	*0.263***
	MeanRR	0.170**	0.174**	0.140**	0.144**	0.067**	0.085	0.113*	0.007
	PLC	*0.220***	*0.222***	*0.224***	*0.225***	*0.354***	*0.413***	0.336**	*0.323***
	SessionEnd	0.164**	*0.252***	0.084**	0.087**	*0.303***	*0.394***	*0.343***	0.223**
	TotalContentTime	−0.001	0.085	−0.009	−0.016	*0.303***	*0.394***	*0.343***	0.223**
Offline	CG@3	0.281**	0.405**	0.232**	0.233**	0.413**	0.353**	0.534**	0.320**
	CG@5	0.275**	0.393**	0.233**	0.234**	0.456**	0.457**	0.545**	0.351**
	CG@10	0.250**	0.341**	0.214**	0.217**	0.453**	0.493**	0.517**	0.355**
	DCG@3	*0.291***	*0.420***	0.240**	0.240**	0.419**	0.401**	0.520**	0.313**
	DCG@5	0.290**	0.417**	0.242**	*0.243***	0.458**	0.463**	*0.546***	0.351**
	DCG@10	0.273**	0.380**	0.231**	0.233**	*0.477***	*0.503***	0.545**	*0.376***

Table 4 (continued)

Metric	TianGong				KDD			
	All	Exploration	Exploitation	Known-item	All	Exploration	Exploitation	Known-item
NDCG@3	0.216**	0.334**	0.236**	0.149**	0.361**	0.408**	0.444**	0.342**
NDCG@5	0.197**	0.301**	0.215**	0.137**	0.343**	0.470**	0.396**	0.353**
NDCG@10	0.183**	0.243**	0.199**	0.136**	0.183**	0.482**	0.184**	0.374**
MaxR	***0.308*****	*0.418***	***0.265*****	***0.267*****	0.384**	***0.529*****	0.373**	0.278**
MinR	0.087**	0.116*	0.061*	0.061*	0.150**	0.169**	0.236**	0.075*
MeanR	0.251**	0.341**	0.217**	0.219**	0.321**	0.31**	0.441**	0.208**
RBP (x=0.1)	0.279**	0.401**	0.229**	0.230**	0.343**	0.441**	0.391**	0.219**
RBP (x=0.5)	*0.295***	***0.422*****	*0.243***	*0.244***	0.434**	0.453**	0.518**	0.320**
RBP (x=0.8)	0.278**	0.387**	0.233**	0.235**	***0.483*****	0.494**	***0.559*****	***0.385*****
RBP (x=0.95)	0.258**	0.353**	0.219**	0.222**	0.448**	0.491**	0.502**	0.363**
ERR	0.177**	0.253**	0.126**	0.126**	0.286**	*0.516***	0.262**	0.041
Precision @3	0.254**	0.361**	0.218**	0.220**	0.381**	0.249**	0.468**	0.344**
Precision @5	0.246**	0.359**	0.214**	0.216**	0.404**	0.308**	0.487**	0.353**
Precision @10	0.225**	0.311**	0.199**	0.202**	0.417**	0.325**	0.491**	*0.375***
AvgClickRel	0.251**	0.341**	*0.256***	0.219**	*0.470***	0.496**	*0.546***	0.361**
ClickPrecision	0.225**	0.311**	0.199**	0.202**	0.344**	0.227**	0.460**	0.271**
QueryCost-Benefit-2	0.225**	0.310**	0.201**	0.204**	0.367**	0.236**	0.483**	0.302**
QueryCost-Benefit-3	0.058**	0.112*	0.060*	0.056*	0.288**	0.303**	0.397**	0.198**
RelDocCount1	0.244**	0.321**	0.192**	0.197**	0.319**	0.475**	0.342**	0.214**
RelDocCount2	0.233**	0.316**	0.192**	0.194**	0.402**	0.437**	0.454**	0.338**

The top k strongest correlations in each column are highlighted in italic for each feature category, where $k = 1, 3, 3$ for query, online, and offline features. The first and second strongest correlations in each column are also bolded and underlined, respectively. Column "all" shows results computed based on all queries without distinguishing task states. $^{*}p < 0.05$, $^{**}p < 0.01$

correlation for the *exploration* state. PLC is higher when a user gets there with fewer clicks on the top results. A user in the *exploitation* and *known-item* states has a clearer idea of the goal and is likely to be more satisfied if a result is found quickly. In the *exploration* state, the user does not have a clear goal and therefore has to try different queries until a document satisfies his information needs and he reaches the end of the session. There are negative correlations in the online metrics and only positive correlations in the offline metrics because the offline metrics are computed based on relevance annotations, so the more relevant the documents in the search result, the higher the offline metrics and the more satisfied the user is. For online metrics, on the other hand, the intuitions are more varied across metrics.

5.2.2 Results on KDD dataset

Similar to the *TianGong* dataset, it can be concluded that how well a metric reflects user satisfaction depends on the task state. The gap between the correlations of different metrics is also high, with the highest correlations being 0.529, 0.559, and 0.385, and the median correlations being 0.325, 0.343, and 0.219, on the *exploration*, *exploitation*, and *known-item* states, respectively. However, with respect to the best metric for measuring user satisfaction in different task states, the results are different from the *TianGong* dataset. The highest correlation in the *exploration* state is achieved by *MaxR*, while in the *exploitation* and *known-item* states it is achieved by *RBP (x=0.8)*. By observing both datasets, we think that this may be caused by the different study setup. The KDD dataset is collected from a lab study where the search goals are given in the task description, the initial challenge is to formulate an appropriate query rather than exploring the different aspects of a topic. In this case, a highly relevant result in the *exploration* state that helps formulate the next query would satisfy the user's search intent, which explains why *MaxR* has a higher correlation compared to other metrics that consider more search results. Since the given tasks usually have more than one sub-goal, users consider several results to cover all the information needs in the exploitation and known-item states, resulting in the *RBP* metric having higher correlations with user satisfaction compared to *MaxR*.

The correlation between online metrics and user satisfaction is overall higher on the *KDD* dataset than on the *TianGong* dataset. As mentioned in Sect. 4, users in the lab study exert more effort per query than in the field study, resulting in more user interactions. This may cause the online metrics to be more informative on the *KDD* dataset. MinRR even outperforms some offline metrics, resulting in the highest correlation for *exploration* state and *exploitation* state among all online metrics.

5.2.3 Implications

With the result of the correlation analysis, we can answer *RQ1*. First, different existing evaluation metrics reflect user satisfaction to different extents under the same task states. Taking exploration state as an example, Table 4 shows that on the KDD dataset, the offline metric *MaxR* has a correlation of 0.529 with user satisfaction, while the online metric *ActionCount* has a correlation of only 0.121. Second, the same metric reflects user satisfaction differently under different task states. For example, the SessionEnd metric achieves a correlation of 0.252 in the *exploration* state on the *TianGong* dataset, it has a correlation of only 0.084 and 0.087 in the *exploitation* and *known-item* states, respectively. We can find another example of this in the *KDD* dataset. Here we see a drop in correlation for the MaxRR metric of 0.432

in the *exploration* state, 0.356 in the *exploitation* state, and only 0.263 in the *known-item* state. Meanwhile, we have found stronger correlations to user satisfaction for offline metrics compared to online metrics. We also see substantial differences in the correlation of metrics across different datasets. This suggests that the way the search task is set up has a strong impact on how users search and evaluate search results, and therefore conclusions drawn from laboratory studies alone may be biased when applied to the real world scenario. Overall, desipite that some metrics achive moderate correlation with user satisfaction, there is still a large gap in using existing metrics for measuring user satisfaction, which demonstrates the neccasity of investigation on new metrics in this respect.

6 Search state detection

Our analysis in Sect. 5 demonstrates that the evaluation metrics reflect user satisfaction to different degrees under different task states. In order to use this result for a more precise evaluation of user satisfaction, we first need to answer *RQ2*: can we detect the task state of a query activity using in-session signals that can be collected automatically? In this section, we present our approach for detecting task states, which we formulate as a classification task, i.e., classifying a query into one of the defined task states. We have experimented with both feature-based machine learning models (Sect. 6.1) and deep learning-based models (Sect. 6.2).

6.1 Feature-based machine learning models

We consider four of the most commonly used feature-based classification models in our experiments, namely logistic regression (LR), k-nearest neighbors (KNN), support vector machines (SVMs), and random forest (RF). The features used by these models are:

6.1.1 Query-based features

We computed several sets of features based on query related information as follows.

- We consider all metrics in Table 1 to be descriptive features.
- *Term frequency* We also consider the original terms in the query string. The terms are represented as a term frequency vector.
- *Readability scores* Query complexity has been found to evolve during the search process (Eickhoff et al., 2014) and thus may provide clues to the search state of the current query. In this work, we compute a set of query readability and complexity scores and use them as features. Many readability scores and complexity metrics have been proposed over the years. In this work, we consider the most commonly used ones, according to the findings in (Eltorai et al., 2015) and (Zhou et al., 2017):
  - The Flesch Reading Ease (FRES) (Flesch, 1979) is computed based on sentence and word length to measure whether a text is in plain English. The higher the number, the easier the text is to read. It is computed as shown in Eq. 1.

$$FRES = 206.835 - \left(1.015 * \frac{\#Words}{\#Sentences}\right) - \left(84.6 * \frac{\#Syllables}{\#Words}\right) \quad (1)$$

– The Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975) measures the education (equivalent to U.S. grade level) required to understand a text. It takes into account the relative number of words per sentence and the number of syllables per word. The higher the result, the easier the text is to read. The calculation is shown in the Eq. 2.

$$FKGL = \left(0.39 * \frac{\#Words}{\#Sentences}\right) + \left(11.8 * \frac{\#Syllables}{\#Words}\right) - 15.59 \tag{2}$$

– The Gunning Fog Index (GFI) introduces the concept of complex words. A complex word is defined as a word with more than three syllables. In addition to the relative proportion of complex words, the length of sentences is also considered. If the result is over 20, the text is considered difficult to read. A text with a score of 5 is readable (Eltorai et al., 2015).

$$GFI = 0.4 * \left[\left(\frac{\#Words}{\#Sentences}\right) + 100 * \left(\frac{\#ComplexWords}{\#Words}\right)\right] \tag{3}$$

– The SMOG Index (SMOG) adopts the concept of complex words (Mc Laughlin, 1969) and measures how many years of education the average person needs to understand a text.

$$SMOG = 1.043 * \sqrt{\#ComplexWords * \left(\frac{30}{\#Sentences}\right)} + 3.1291 \tag{4}$$

– The Automated Readability Index also provides a grade. It was developed for the U.S. Air Force to determine how readable text is as it is typed. Senter and Smith chose the number of characters per word and words per sentence to assess readability. Regression analysis is used to determine the weight of the parts was based on 24 books labeled with readability (Zhou et al., 2017).

$$ARI = 4.71 * \left(\frac{\#Characters}{\#Words}\right) + 0.5 * \left(\frac{\#Words}{\#Sentences}\right) - 21.43 \tag{5}$$

– The Cole-Liau Index is based on a regression analysis of 36 150-word passages with cloze percentages. The authors included the average number of letters per 100 words and the number of sentences per 100 words. The result represents the U.S. grade level of the reader's reading skills (Zhou et al., 2017).

$$\begin{aligned} CLI = &\left[0.0588 * \left(\frac{\#Characters}{\#Words} * 100\right)\right] \\ &- \left[0.296 * \left(\frac{\#Sentences}{\#Words} * 100\right)\right] - 15.8 \end{aligned} \tag{6}$$

6.1.2 Online metrics as features

User interaction signals are easily obtained from the search engine log and can potentially be indicators of task state. Therefore, in addition to query-based features, we also use online metrics (see Table 2) computed from in-session signals as features.

6.2 BERT-based language models

Detecting the search state early in a search session, i.e., when query terms are entered, can enable search systems to adjust ranking optimizations accordingly. Therefore, in addition to the feature-based classification approach that uses various signals in a search session as introduced in Sect. 6.1, we tried to apply the advanced language models to understand the semantics in query terms for detecting search state. Models based on Bidirectional Encoder Representation from Transformers (BERT) have been applied to many natural language processing tasks and have achieved superior performance (Devlin et al., 2018). Based on it, we develop a two-step pipeline: a pre-training step with unlabeled data, and a fine-tuning step with task-specific labeled data.

There are two different tasks in pre-training. The first task is to train the Masked Language Model (MLM). Existing systems have taken a unidirectional approach by examining the language from left to right. This approach has been further developed in the BERT model by examining the context bidirectionally. To do this, 15% of the input tokens are masked. The system then predicts the words behind the masked tokens. The second task in pre-training is Next Sentence Prediction (NSP), where the probability that sentence B follows sentence A is determined. In the fine-tuning phase, task-specific data and labels are provided to the model. The texts are encoded with pre-trained embeddings, which are then fed into an output layer for classification. The task state labels are used to train the classification model. The BERT model we realized is based on the implementation of the multi-layer bidirectional transformer encoder by Vaswani et al. (2017) from the tensor2tensor (Vaswani et al., 2018) library. The model used in this work corresponds to the $BERT_{BASE}$ model with 12 layers, a hidden size of 768 and 12 self-attention heads (Devlin et al., 2018).

6.3 Experimental evaluation of task state prediction

To evaluate the prediction performance of the models, we compute the standard precision (P), recall (R) and F1 score for each class. To evaluate the overall result, we compute the accuracy (acc) and the macro average of precision, recall and F1 score. We perform 10-fold cross-validation on each of the two experimental datasets. The evaluation results of the search state prediction models obtained on both datasets are shown in Table 5.

To answer *RQ2*, we observed in the result that the applied models achieved over 59.8% accuracy on both datasets, demonstrating that the signals we chose are effective and that task states can be predicted from user interactions. Comparing different models, the BERT-based model using query terms achieved the best performance in terms of both accuracy and average F1 score on both datasets. Among the results of the BERT model, the highest F1 score is achieved in the *known-item* state, while *exploration* and *exploitation* are harder to distinguish.

With respect to the different datasets, we notice a generally lower performance on *TianGong dataset*. After investigating the original dataset, we think that the reasons for the low performance of the state detection models on *TianGong* dataset, especially for the *exploration* state, are due to the high topic diversity and the smaller number of samples. In terms of topic diversity, the *KDD* dataset is collected under the setup that all search activities are related to the 9 predefined topics, while the *TianGong* dataset is collected in a field study, the topics are very diverse, the overlap of search topics among

Table 5 Evaluation result of search state prediction

			Exploration			Exploitation			Known-item			Macro average			All
	Feature	Model	P	R	F1	P	R	F1	P	R	F1	P	R	F1	Accu
KDD	Q	LR	0.598	0.510	0.548	0.579	0.294	0.386	0.592	0.837	0.693	0.590	0.547	0.542	0.588
		KNN	0.550	0.623	0.581	0.489	0.440	0.462	0.702	0.708	0.703	0.580	0.590	0.582	0.604
		SVM	0.512	0.426	0.464	0.715	0.030	0.058	0.530	0.904	0.667	0.586	0.453	0.396	0.530
		RF	0.618	0.633	0.625	0.522	0.450	0.481	0.683	0.737	0.708	0.608	0.607	0.604	0.624
		BERT	0.485	0.800	0.604	0.582	0.542	0.561	0.806	0.711	0.755	0.624	0.684	**0.640**	**0.658**
	Q+O	LR	0.523	0.028	0.053	0.469	0.202	0.260	0.497	0.879	0.634	0.496	0.370	0.315	0.483
		KNN	0.261	0.259	0.257	0.351	0.345	0.347	0.509	0.514	0.510	0.374	0.372	0.371	0.405
		SVM	0.412	0.059	0.102	0.364	0.061	0.103	0.487	0.939	0.640	0.421	0.353	0.282	0.475
		RF	0.695	0.587	0.635	0.576	0.475	0.518	0.704	0.829	0.760	0.658	0.631	**0.638**	**0.668**
TianGong	Q	LR	0.200	0.006	0.012	0.602	0.949	0.736	0.534	0.109	0.181	0.445	0.355	0.309	0.597
		KNN	0.262	0.162	0.198	0.614	0.753	0.676	0.449	0.308	0.364	0.442	0.407	0.413	0.557
		SVM	0.000	0.000	0.000	0.584	0.999	0.737	0.150	0.001	0.002	0.245	0.333	0.246	0.583
		RF	0.249	0.068	0.106	0.605	0.753	0.671	0.451	0.343	0.389	0.435	0.388	0.389	0.557
		BERT	0.088	0.273	0.133	0.781	0.740	0.760	0.694	0.641	0.667	0.521	0.551	**0.520**	**0.693**
	Q+O	LR	0.000	0.000	0.000	0.588	0.997	0.740	0.000	0.000	0.000	0.196	0.332	0.247	0.587
		KNN	0.298	0.212	0.247	0.659	0.762	0.707	0.511	0.409	0.453	0.489	0.461	**0.469**	**0.598**
		SVM	0.000	0.000	0.000	0.593	1.000	0.744	0.000	0.000	0.000	0.198	0.333	0.248	0.593
		RF	0.351	0.068	0.114	0.615	0.825	0.704	0.486	0.296	0.368	0.484	0.396	0.395	0.584

Bolded numbers indicate the highest values of that metric when using the corresponding set of features

In the 'Feature' column, Q indicates the use of only query terms (BERT) or query-related features; Q+O indicates the use of both query and online metric-based features

participants is very small. Meanwhile, the *exploration* state has the least number of samples, i.e. 356, 2266, 1252 samples in*exploration, exploitation* and *known-item* state respectively in *TianGong* dataset. This resulted in the model not being fully trained. This can also explain that after adding online features, the performance of the model decreased even further, as there are not enough samples to train a robust model. For the feature-based classifier on the KDD dataset, adding online features on top of query-based features improves the performance of the models, suggesting that user interaction with the search engine can provide signals that indicate the task state.

7 Construction of state-aware evaluation metrics

We found that evaluation metrics correlate differently with user satisfaction under different task states and that it is possible to predict task states based on in-session signals. We now try to develop new evaluation metrics that can better assess user satisfaction and answer *RQ3*.

7.1 Task formulation

As a preliminary attempt, we aim at creating explainable state-aware evaluation metrics. Hence we choose to use a linear regression model to combine existing metrics and features to better measure satisfaction. The basic linear regression model for multiple features is as follows:

$$y = \beta_0 x_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

$X = \{x_0, x_1, \ldots, x_p\}$ is the set of features we are using. The βs are the coefficients for each feature, ϵ is the bias, and y is our target, i.e., user satisfaction. We perform least squares on this formula and try to optimize the βs. The method we use to perform the linear regression is implemented in the scikit-learn library. To construct the new evaluation metrics by linear regression, we consider the query, online and offline evaluation metrics as introduced in Sect. 5.1 and the readability scores computed from query terms as described in Sect. 6.1 as input features. We start by fitting the linear model with all considered features on the two datasets and for each search state, respectively.

7.2 Experimental results

We fitted new metrics under two settings: (1) general metrics ($new - all$) that do not distinguish between task states, and (2) state-specific metrics ($new - st$) that are trained on state-specific data. To fit the linear model, we split the data into 70% for model fitting and 30% for evaluation. To improve the interoperability and efficiency of the fitted linear model, we applied the sequential forward selection strategy (John et al., 1994; Last et al., 2001). The highest correlations obtained under different setups are shown in Table 6.

Based on the results, we can answer *RQ3*—compared to existing metrics in the general setting and in different task states, we observe that the highest correlations (bolded

Table 6 Highest correlation of evaluation metrics in each setup

		All		Exploration			Exploitation			Known-item		
Dataset	Feature	New-all	Existing	New-all	New-st	Existing	New-all	New-st	Existing	New-all	New-st	Existing
TianGong	On	**0.292**	0.220	0.186	**0.282**	0.252	0.283	**0.408**	0.224	**0.335**	0.330	0.225
	On+off	**0.466**	0.308	**0.528**	0.515	0.422	0.486	**0.495**	0.265	0.412	**0.414**	0.267
KDD	On	**0.555**	0.354	0.569	**0.629**	0.432	**0.556**	0.551	0.356	**0.523**	0.467	0.323
	On+off	**0.595**	0.483	0.588	**0.681**	0.529	0.634	**0.674**	0.559	0.543	**0.621**	0.385

In the 'feature' column, 'on' represents using online features, 'on+off" represents using both online and offline features. The highest correlation in each state and feature group is bolded

in Table 6) are all achieved by new metrics, demonstrating that user satisfaction can be better measured by combining online signals and existing metrics. The general metrics (*new − all*) achieved higher correlation with user satisfaction compared to the existing metrics in most cases, with one exception (underlined in table 6) on the *TianGong* dataset. One possible reason is that the *TianGong* dataset contains fewer online signals, more diverse topics and user behavior, while the task state distribution is unbalanced. Therefore, without enough meaningful signals to distinguish the characteristics of different search states, the trained metrics are better at capturing easier patterns in the larger class *known-item* state, resulting in less predictive power for user satisfaction in other search states. Comparing the general and state-specific metrics, the state-specific metrics outperformed the general metrics in more cases, even with less training data. Results suggest that having state-specific metrics for certain task states can be useful to better measure user satisfaction.

Comparing between different feature groups, we observe that on both datasets and in all states, the highest correlations are obtained by using both online and offline features. Overall, the correlation values between the new metrics and user satisfaction are lower on the *TianGong* dataset compared to the *KDD* dataset. The metrics are less stable on the *TianGong* dataset when only online features are used. A possible reason could be that due to the diversity of tasks in the *TianGong* dataset, more training data is needed for the models to learn robust metrics.

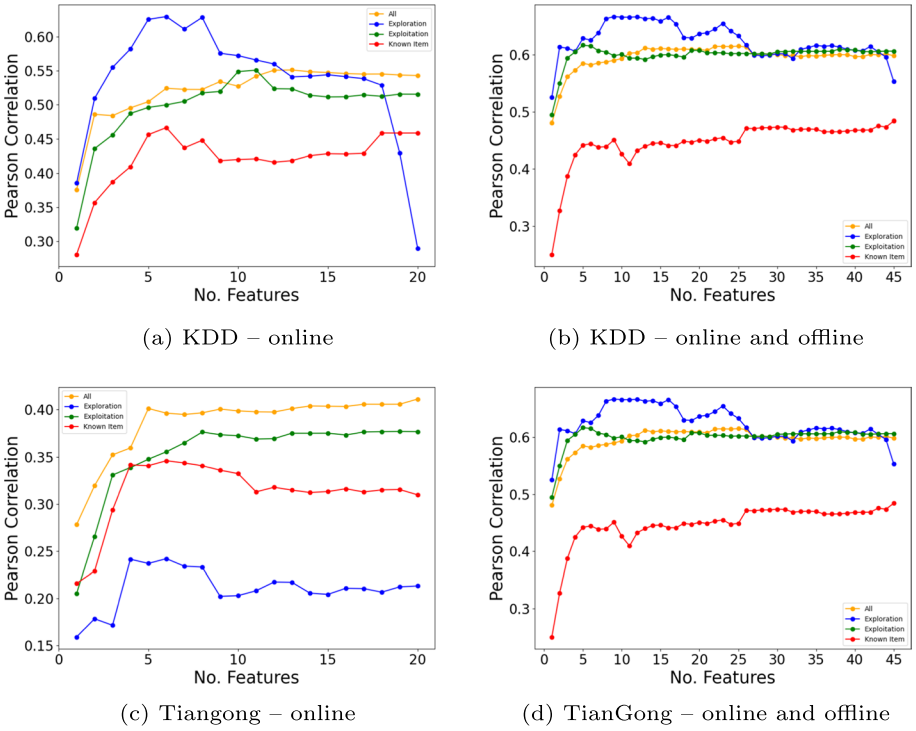
The result of selecting a different number of features is shown in Fig. 3, where the x-axis shows the number of features used by the linear function, and the corresponding value on the y-axis is the correlation between the new metric and user satisfaction on the validation set.

On the *KDD* dataset (Fig. 3a and b), we observe that the correlations between the new metrics and user satisfaction increase up to a point as the number of features increases. This suggests that we can better assess user satisfaction with the search result by considering multiple metrics simultaneously. While for the *exploration* state, the correlation starts to drop rapidly after reaching a certain number of features. The drop in correlation is likely caused by model overfitting, as the *exploration* state has the least number of samples.

Similar trends are shown on the *TianGong* dataset (Fig. 3c and d), where the combination of multiple metrics can better reflect user satisfaction than a single metric. A possible overfitting effect is shown by increasing the number of features for the *exploration* state in both feature settings, i.e. the class with fewer samples compared to the *exploitation* and *known-item* states.

8 Discussion and conclusions

People increasingly rely on search and recommendation technologies to perform information-intensive tasks and make complex decisions. In contrast to simple, fact-finding retrieval tasks, complex search tasks often involve prolonged search sessions and motivate users to achieve different intentions and subgoals across different *task states* in information seeking and search episodes (Mitsui et al., 2016). While different states and intentions may affect the way users evaluate the performance of search systems (Liu et al., 2019b), how to adaptively evaluate systems and employ the appropriate metric(s) that best reflect task states and user satisfaction criteria currently



(a) KDD – online (b) KDD – online and offline

(c) Tiangong – online (d) TianGong – online and offline

Fig. 3 Pearson correlation between new evaluation metrics and user satisfaction. Blue: *exploration* state; Green: *exploitation* state, Red: *known-item* state; All: *without distinguishing task states* (Color figure online)

remains an open challenge. This gap is unlikely to be addressed by applying one or two mainstream "unified metrics" such as precision and nDCG.

To address this challenge, our study goes beyond the mainstream Cranfield-style offline evaluation approach (Voorhees, 2001) and seeks to develop an adaptive evaluation approach by (1) meta-evaluating the performance of different metrics under different task states and (2) developing new metrics that better capture user in-situ satisfaction. In this paper, we first investigate the correlation between existing IR evaluation metrics and user satisfaction under different task states and find that the degree to which a metric can reflect user satisfaction varies across task states. The analysis extends previous work (Liu & Yu, 2021) not only by considering a more complete set of evaluation metrics but also by conducting the analysis on datasets collected from different search scenarios, i.e. both field study and task-oriented lab studies, to mitigate the impact of the experimental setup. In the next step, we experiment with the automatic detection of the task state of a search query based on in-session signals. In addition to the models used in the baseline work (Liu & Yu, 2021), we also applied a BERT-based classification model to query terms and achieved superior performance compared to other feature-based models. To better assess user satisfaction, we constructed a set of new evaluation metrics using linear regression. The results demonstrate that the new metrics can better reflect satisfaction than the existing metrics in the same category, i.e., online or offline metrics. In this study, we

experimentally found the best-combined metric in each configuration. When applying the combined metric to support real-world applications, such as real-time satisfaction prediction or search result re-ranking, the metric could be further reduced depending on the objective and feature availability. For example, in real-world applications, offline metrics are not available, so it may be necessary to construct an evaluation metric based only on online signals.

The characteristics of users, such as their level of knowledge about the topic and their familiarity with search engines, can have a strong impact on their level of satisfaction with different search results. Incorporating user characteristics into the evaluation metric could potentially increase its effectiveness in measuring satisfaction. Limited by the availability of user information and the diversity of users in the available datasets, we did not experiment with user features in this work. Since it is costly to obtain user satisfaction, document relevance, and task state annotations, the main limitation of our work is the size of our experimental dataset. Meanwhile, the task state classes are unbalanced, and some of the classes have few training examples for building the classifiers in Sect. 6. For these reasons, we believe that the power of the features and classifiers is limited. However, our experiments manifested the potential of a state-aware approach in understanding users' intentions and supporting interactive IR activities and presented new metrics that can more accurately characterize users' in-situ experiences. Together, the behavioral features, classifiers, and state-aware evaluation metrics provide a methodological and empirical foundation for more fine-grained user modeling, adaptive search recommendations, and dynamic IR evaluation. In future work, we plan to work on heuristics that enable user search and annotation data collection on a larger scale. We will also seek to verify our results, fine-tune the task state classifiers, and meta-evaluate the new metrics on more diverse datasets, task types, and user populations.

Appendix A: New evaluation metrics

Here we list the formulas of the 12 new evaluation metrics corresponding to Table 6. The new metrics are denoted as $F\_state^{feature}$, where $state \in \{exploration, exploitation, known-item\}$, $feature \in \{online(on), online\&offline(on+off)\}$.

A.1 Metrics fitted on the TianGong dataset

$$\begin{aligned} F\_exploration^{on} = {} & 0.5360 * MinRR - 0.0006 * MaxScroll \\ & + 0.7023 * SessionEnd - 0.1885 * GFI + 2.7479 \end{aligned} \tag{A1}$$

$$\begin{aligned} F\_exploration^{on+off} = {} & -0.0017 * ActionCount - 0.0005 * MaxScroll \\ & + 0.3973 * MinRR - 0.0070 * NewTerms \\ & + 0.5844 * SessionEnd + 0.0765 * DCG@3 \\ & + 0.1992 * NDCG@5 + 0.1774 * MeanR + 2.0224 \end{aligned} \tag{A2}$$

$$
\begin{aligned}
F\_exploitation^{on} = & -10.2405 * MeanRR + 0.8970 * MinRR \\
& - 0.1267 * MaxRR - 0.7848 * AvgClickRank \\
& + 0.0596 * LowestClickRrank + 0.1776 * PLC \\
& - 0.0393 * HighestClickRank + 0.6746 * \#Clicks \\
& - 0.1987 * QuerySim - 0.0302 * ActionCount \\
& + 0.0282 * MouseMoveCount + 0.0003 * ScrollDistanceance \\
& - 0.0007 * MaxScroll + 2.022e - 06 * TimeToFirstClick \\
& + 4.481e - 9 * TimeToLastClick \\
& + 5.500e - 6 * TotalContentTime \\
& + 0.3660 * SessionEnd + 0.2318 * FRES + 1.6966 * FKGL \\
& - 0.9548 * GFI - 0.0329 * SMOG - 6.019e - 6 * ARI \\
& - 0.0077 * CLI - 19.2787
\end{aligned} \tag{A3}
$$

$$
\begin{aligned}
F\_exploitation^{on+off} = & -0.0272 * ActionCount + 0.0269 * MouseMoveCount \\
& + 0.0003 * ScrollDistance - 0.0008 * MaxScroll \\
& + 9.6017 * MeanRR + 0.4054 * MinRR \\
& - 1.7495 * MaxRR - 0.7617 * AvgClickRank \\
& + 0.0847 * LowestClickRank + 0.3343 * PLC \\
& - 0.0550 * HighestClickRank + 0.4838 * \#Clicks \\
& - 0.0049 * NewTerms - 0.2170 * QuerySim \\
& + 0.3975 * SessionEnd + 0.0750 * FRES + 0.5696 * FKGL \\
& - 0.1836 * GFI - 0.0504 * SMOG + 0.0001 * ARI \\
& - 0.0069 * CLI + 2.7225 * CG@3 - 8.2646 * CG@10 \\
& - 5.9392 * DCG@3 + 26.7016 * DCG@10 + 0.0247 * NDCG@3 \\
& + 0.2429 * MaxR + 3.6601 * MeanR + -5.0641 * RBP(x = 0.1) \\
& - 6.8486 * RBP(x = 0.5) - 47.4918 * RBP(x = 0.8) \\
& + 29.0866 * RBP(x = 0.95) - 0.0586 * ERR \\
& + 0.0175 * Precision@3 - 0.2638 * Precision@5 \\
& + 24.7966 * Precision@10 + 0.1349 * RelDocCount1 \\
& + 0.1157 * RelDocCount2 - 11.6368 * ClickPrecision \\
& - 11.6368 * QueryCostBenefit2 + 3.6601 * AvgClickRel \\
& - 5.1123
\end{aligned} \tag{A4}
$$

$$
\begin{aligned}
F\_known - item^{on} = & 0.5708 * PLC - 0.2833 * QuerySim \\
& - 0.0030 * ActionCount - 0.0006 * MaxScroll \\
& - 1.344e - 05 * TimeToFirstClick \\
& + 6.420e - 6 * TotalContentTime \\
& - 0.0026 * FRES - 0.0084 * CLI \\
& + 3.1930
\end{aligned} \tag{A5}
$$

$$\begin{aligned} F\_known - item^{on+off} = & -0.0001 * ScrollDistance - 0.0004 * MaxScroll \\ & + 0.4191 * MinRR - 0.2262 * AvgClickRank \\ & + 0.0679 * PLC + 0.0346 * HighestClickRank \\ & - 0.1210 * QuerySim + 0.0234 * FKGL - 0.0371 * SMOG \\ & + 0.0003 * ARI - 0.0091 * CLI + 0.2452 * MaxR \\ & - 0.0775 * MinR + 0.0173 * RelDocCount2 \\ & + 2.2506 \end{aligned} \tag{A6}$$

$$\begin{aligned} F\_all^{on} = & -0.001795 * ActionCount - 0.002198 * MaxScroll \\ & - 2.9299 * MeanRR + 1.0501 * MinRR \\ & - 0.3766 * AvgClickRank - 0.0191 * LowestClickRank \\ & + 0.4014 * PLC + 0.1795 * HighestClickRank \\ & + 0.3193 * \#Clicks - 0.3826 * QuerySim \\ & + 0.7985 * SessionEnd + 0.1325 * SMOG \\ & + 2.0223 \end{aligned} \tag{A7}$$

$$\begin{aligned} F\_all^{on+off} = & -0.0174 * ActionCount + 0.0171 * MouseMoveCount \\ & + 0.0002 * ScrollDistance - 0.0006 * MaxScroll \\ & - 2.8137 * MeanRR + 0.6774 * MinRR - 0.1579 * MaxRR \\ & - 0.5110 * AvgClickRank + 0.0345 * LowestClickRank \\ & - 0.1179 * PLC + 0.2335 * \#Clicks + 0.0010 * NewTerms \\ & - 0.1527 * QuerySim + 0.2915 * SessionEnd + 0.1621 * FRES \\ & + 1.1837 * FKGL + -0.6682 * GFI + 0.0080 * SMOG \\ & + 0.0000 * ARI - 0.0069 * CLI + 23.7345 * CG@3 \\ & + 13.6403 * CG@5 - 33.4357 * CG@10 - 51.3511 * DCG@3 \\ & - 36.4389 * DCG@5 + 118.0293 * DCG@10 + 0.0483 * NDCG@3 \\ & - 0.1431 * NDCG@5 + 0.1519 * NDCG@10 + 0.1604 * MaxR \\ & - 0.0281 * MinR + 3.8755 * MeanR - 13.4478 * RBP(x = 0.1) \\ & + 36.9579 * RBP(x = 0.5) - 246.4284 * RBP(x = 0.8) \\ & + 158.0059 * RBP(x = 0.95) + 0.1633 * ERR \\ & + 0.1528 * Precision@3 - 0.2735 * Precision@5 \\ & + 25.2055 * Precision@10 + 0.1274 * RelDocCount1 \\ & + 0.0840 * RelDocCount2 - 11.9427 * ClickPrecision \\ & - 11.9427 * QueryCostBenefit2 + 3.8755 * AvgClickRel \\ & - 13.7308 \end{aligned} \tag{A8}$$

A.2 Metrics fitted on the KDD dataset

$$\begin{aligned} F\_exploration^{on} = {} & 0.0139 * MaxRR - 0.9012 * AvgClickRank \\ & + 1.0712 * PLC + 0.1727 * HighestClickRank \\ & + 0.3714 * \#Clicks + 0.8835 * SessionEnd + 1.9670 \end{aligned} \tag{A9}$$

$$\begin{aligned} F\_exploration^{on+off} = {} & -0.0002 * ScrollDistance + 0.6656 * PLC \\ & + 0.8788 * SessionEnd - 0.4733 * GFI \\ & + 0.0084 * ARI - 0.3873 * CG@3 + 0.4381 * DCG@5 \\ & + 0.4669 * NDCG@10 + 0.2880 * MaxR + 1.1829 \end{aligned} \tag{A10}$$

$$\begin{aligned} F\_exploitation^{on} = {} & -0.0018 * ActionCount - 0.0011 * MaxScroll \\ & - 1.7031 * MeanRR + 1.0807 * MinRR \\ & + 0.0116 * LowestClickRank + 0.5176 * PLC \\ & + 0.1537 * HighestClickRank + 0.1320 * \#Click \\ & - 0.6024 * QuerySim + 1.0086 * SessionEnd \\ & + 0.1314 * SMOG + 1.5640 \end{aligned} \tag{A11}$$

$$\begin{aligned} F\_exploitation^{on+off} = {} & -0.0013 * ActionCount + 0.0001 * MaxScroll \\ & + 0.8984 * MinRR + 0.1291 * PLC \\ & + 0.1149 * HighestClickRank - 0.3129 * QuerySim \\ & + 0.8929 * SessionEnd + 0.0229 * FKGL \\ & - 0.0242 * GFI + 0.8032 * CG@3 + 0.3682 * CG@10 \\ & - 2.2387 * DCG@3 - 1.3680 * DCG@10 \\ & + 0.4560 * NDCG@3 - 0.3209 * NDCG@5 \\ & - 0.1335 * NDCG@10 - 0.5525 * MeanR \\ & - 0.1330 * RBP(x = 0.1) + 5.3020 * RBP(x = 0.5) \\ & - 0.1488 * ERR - 0.5917 * Precision@3 \\ & + 0.0402 * RelDocCount1 + 1.0027 * ClickPrecision \\ & + 1.0702 * QueryCostBenefit2 + 0.4251 * AvgClickRel \\ & + 1.1105 \end{aligned} \tag{A12}$$

$$\begin{aligned} F\_known - item^{on} = {} & -0.0007 * ScrollDistance - 4.3411 * MeanRR \\ & + 0.1078 * LowestClickRank + 1.6888 * PLC \\ & + 0.6119 * SessionEnd - 0.0035 * FRES + 2.9388 \end{aligned} \tag{A13}$$

$$
\begin{aligned}
F\_known - item^{on+off} = & -0.0008 * ActionCount + -0.0017 * MouseMoveCount \\
& - 0.0003 * ScrollDistance - 0.0029 * MaxScroll \\
& - 0.3455 * MeanRR + 1.3413 * MinRR \\
& - 0.9586 * MaxRR - 0.4687 * AvgClickRank \\
& - 0.0273 * LowestClickRank + 0.1549 * PLC \\
& + 0.1413 * HighestClickRank + 0.5481 * \#Clicks \\
& - 0.0604 * NewTerms - 0.9784 * QuerySim \\
& + 0.4799 * SessionEnd + 0.5366 * FRES \\
& + 3.8820 * FKGL - 2.6873 * GFI \\
& + 0.1093 * SMOG + 0.0006 * ARI + 0.0096 * CLI \\
& - 5.0538 * CG@3 - 5.5758 * CG@5 - 2.0028 * CG@10 \\
& + 10.6242 * DCG@3 + 15.1764 * DCG@5 \\
& + 6.8286 * DCG@10 - 0.2600 * NDCG@3 \\
& + 0.0545 * NDCG@5 + -0.2418 * NDCG@10 \\
& - 0.0284 * MaxR - 0.0424 * MinR - 0.1541 * MeanR \\
& - 5.7473 * RBP(x = 0.1) - 29.4309 * RBP(x = 0.5) \\
& + 0.4429 * RBP(x = 0.8) - 1.1266 * RBP(x = 0.95) \\
& + 0.7724 * Precision@3 - 0.9949 * Precision@5 \\
& + 2.2600 * Precision@10 + 0.0662 * RelDocCount1 \\
& + 0.0489 * RelDocCount2 - 0.8741 * ClickPrecision \\
& - 0.6607 * QueryCostBenefit2 + 0.2771 * AvgClickRel \\
& - 48.5396
\end{aligned}
\tag{A14}
$$

$$
\begin{aligned}
F\_all^{on} = & -0.0017 * ActionCount - 0.0001 * ScrollDistance \\
& - 0.0019 * MaxScroll - 1.1029 * MeanRR \\
& + 1.6565 * MinRR - 1.0159 * MaxRR - 0.6754 * AvgClickRank \\
& + 0.5492 * PLC + 0.1892 * HighestClickRank \\
& + 0.5866 * \#Clicks - 0.3657 * QuerySim \\
& + 0.7161 * SessionEnd - 0.0042 * FRES + 2.6961
\end{aligned}
\tag{A15}
$$

$$
\begin{aligned}
F\_all^{on+off} = & -0.0004 * ScrollDistance + 0.0527 * LowestClickRank \\
& + 0.9360 * PLC + 0.5567 * SessionEnd \\
& + 0.8777 * RBP(x = 0.8) + 1.1917
\end{aligned}
\tag{A16}
$$

Author Contributions Ran Yu and Jiqun Liu initiated the research idea and conducted the preliminary research. Marco Markwald carried out most of the data analysis and experiments presented in this manuscript under the supervision of Ran Yu. All three authors made substantial contributions to the writing of the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability The raw data used in this work is publicly available. Expert annotations produced in this work can be shared upon request.

Declarations

Conflict of interest I declare that the authors have no competing interests as defined by Springer, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

Ethical approval Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Borlund, P. (2016). Framing of different types of information needs within simulated work task situations: An empirical study in the school context. *Journal of Information Science, 42*(3), 313–323. https://doi.org/10.1177/0165551515625028

Broder, A. (2002). A taxonomy of web search. *ACM Sigir Forum* (pp. 3–10). NY, USA: ACM New York.

Chen, J., Liu, Y., Mao, J., Zhang, F., Sakai, T., Ma, W., Zhang, M. & Ma, S. (2021). Incorporating query reformulating behavior into web search evaluation. In: *CIKM '21: The 30th ACM international conference on information and knowledge management, virtual event, Queensland, Australia,* November 1–5, 2021. ACM, pp 171–180, https://doi.org/10.1145/3459637.3482438

Chen, Y., Zhou, K., Liu, Y., Zhang, M. & Ma, S. (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*, pp. 15–24.

Chuklin, A. & de Rijke, M. (2016). Incorporating clicks, attention and satisfaction into a search engine result page evaluation model. In: *Proceedings of the 25th acm international on conference on information and knowledge management*, pp. 175–184.

Cole, M., Liu, J., Belkin, N. J., Bierig, R., Gwizdka, J., Liu, C., Zhang, J. & Zhang, X. (2009). Usefulness as the criterion for evaluation of interactive information retrieval. In: *Proceedings of the third workshop on human-computer interaction and information retrieval Cambridge*, HCIR, pp. 1–4.

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

Eickhoff, C., Teevan, J., White, R., & Dumais, S. (2014). Lessons from the journey: a query log analysis of within-session learning. In: *Proceedings of the 7th ACM international conference on Web search and data mining*, pp. 223–232.

Eltorai, A. E., Naqvi, S. S., Ghanian, S., Eberson, C. P., Weiss, A. P., Born, C. T., & Daniels, A. H. (2015). Readability of invasive procedure consent forms. *Clinical and Translational Science, 8*(6), 830–833.

Flesch, R. (1979). *How to write plain english*. University of Canterbury

Harman, D. (2011). Information retrieval evaluation. *Synthesis Lectures on Information Concepts, Retrieval, and Services, 3*(2), 1–119.

Järvelin, K., Vakkari, P., Arvola, P., Baskaya, F., Jarvelin, A., Kekalainen, J., Keskustalo, H., Kumpulainen, S., Saastamoinen, M., Savolainen, R., & Sormunen, E. (2015). Task-based information interaction evaluation: The viewpoint of program theory. *ACM Transactions on Information Systems, 33*(1), 31–330. https://doi.org/10.1145/2699660

Jiang, J., Hassan Awadallah, A., Shi, X. & White, R. W. (2015). Understanding and predicting graded search satisfaction. In *Proceedings of the eighth ACM international conference on web search and data mining*, pp. 57–66.

John, G. H., Kohavi, R., Pfleger, K. (1994). Irrelevant features and the subset selection problem. In: *Machine learning proceedings 1994*. Elsevier, pp. 121–129.

Kincaid, J. P., Fishburne, R. P., Jr., Rogers, R. L., et al. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Naval Technical Training Command Millington TN Research Branch: Tech. rep.

Last, M., Kandel, A., & Maimon, O. (2001). Information-theoretic algorithm for feature selection. *Pattern Recognition Letters, 22*(6–7), 799–811.

Liu, J., & Han, F. (2022). Matching search result diversity with user diversity acceptance in web search sessions. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pp. 2473–2477.

Liu, J., & Shah, C. (2022). Leveraging user interaction signals and task state information in adaptively optimizing usefulness-oriented search sessions. In *Proceedings of the 22nd ACM/IEEE joint conference on digital libraries*, pp. 1–11.

Liu, J., & Yu, R. (2021). State-aware meta-evaluation of evaluation metrics in interactive information retrieval. In *Proceedings of the 30th ACM international conference on information and knowledge management*, pp. 3258–3262.

Liu, Y., Chen, Y., Tang, J., Sun, J., Zhang, M., Ma, S., & Zhu, X. (2015). Different users, different opinions: Predicting search satisfaction with mouse movement information. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pp. 493–502.

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. In *Proceedings of the 2018 World Wide Web Conference*, pp. 1533–1542.

Liu, M., Mao, J., Liu, Y., Zhang, M. & Ma, S. (2019b). Investigating cognitive effects in session-level search user satisfaction. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 923–931.

Liu, J., Mitsui, M., Belkin, N. J. & Shah, C. (2019a). Task, information seeking intentions, and user behavior: Toward a multi-level understanding of web search. In *Proceedings of the 2019 conference on human information interaction and retrieval*, pp. 123–132.

Liu, J., Sarkar, S., & Shah, C. (2020). Identifying and predicting the states of complex search tasks. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, pp. 193–202.

Liu, J. (2022). Toward cranfield-inspired reusability assessment in interactive information retrieval evaluation. *Information Processing and Management, 59*(5), 103007.

Luo, J., Zhang, S., & Yang, H. (2014). Win-win search: Dual-agent stochastic game in session search. In *Proceedings of the 37th international ACM SIGIR conference on research and development in information retrieval*, pp. 587–596.

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? In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 463–472.

Mc Laughlin, G. H. (1969). Smog grading-a new readability formula. *Journal of Reading, 12*(8), 639–646.

Mitsui, M., Shah, C., & Belkin, N. J. (2016). Extracting information seeking intentions for web search sessions. In *Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval*, pp. 841–844.

Moffat, A., Mackenzie, J., Thomas, P. & Azzopardi, L. (2022). A flexible framework for offline effectiveness metrics. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Madrid, Spain, July 11–15, 2022. ACM, pp. 578–587, https://doi.org/10.1145/3477495.3531924

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

Ruotsalo, T., Jacucci, G., Myllymäki, P., et al. (2014). Interactive intent modeling: Information discovery beyond search. *Communications of the ACM, 58*(1), 86–92.

Sarkar, S., Mitsui, M., Liu, J., & Shah, C. (2020). Implicit information need as explicit problems, help, and behavioral signals. *Information Processing and Management, 57*(2), 102069.

Urgo, K., & Arguello, J. (2022). Understanding the "pathway'' towards a searcher's learning objective. *ACM Transactions on Information Systems (TOIS), 40*(4), 1–42.

Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A. N., Gouws, S., Jones, L., Kaiser, L., Kalchbrenner, N., Parmar, N., & Sepassi, R. (2018). *Tensor2tensor for neural machine translation*. CoRR abs/1803.07416

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Voorhees, E. M. (2001). The philosophy of information retrieval evaluation. In *Workshop of the cross-language evaluation forum for european languages*, Springer, pp. 355–370.

Vuong, T. T., Saastamoinen, M., Jacucci, G., et al. (2019). Understanding user behavior in naturalistic information search tasks. *Journal of the Association for Information Science and Technology, 70*(11), 1248–1261. https://doi.org/10.1002/asi.24201

Wicaksono, A. F., & Moffat, A. (2020). Metrics, user models, and satisfaction. In *WSDM '20: The Thirteenth ACM international conference on web search and data mining*, Houston, TX, USA, February 3–7, 2020. ACM, pp. 654–662, https://doi.org/10.1145/3336191.3371799

Wicaksono, A. F., & Moffat, A. (2021). Modeling search and session effectiveness. *Information Processing and Management, 58*(4), 102601. https://doi.org/10.1016/j.ipm.2021.102601

Zhang, F., Mao, J., Liu, Y., Ma, W., Zhang, M. & Ma, S. (2020a). Cascade or recency: Constructing better evaluation metrics for session search. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pp. 389–398.

Zhang, F., Mao, J., Liu, Y., Xie, X., Ma, W., Zhang, M. & Ma, S. (2020c). Models versus satisfaction: Towards a better understanding of evaluation metrics. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pp. 379–388.

Zhang, F., Mao, J., Liu, Y., Xie, X., Ma, W., Zhang, M., & Ma, S. (2020b). Models versus satisfaction: Towards a better understanding of evaluation metrics. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, SIGIR 2020, Virtual Event, China, July 25–30, 2020. ACM, pp. 379–388, https://doi.org/10.1145/3397271.3401162

Zhang, F., Zhou, K., Shao, Y., Luo, C., Zhang, M., & Ma, S. (2018). How well do offline and online evaluation metrics measure user satisfaction in web image search? In *The 41st international ACM SIGIR conference on research and development in information retrieval*, pp. 615–624.

Zhou, S., Jeong, H., & Green, P. A. (2017). How consistent are the best-known readability equations in estimating the readability of design standards? *IEEE Transactions on Professional Communication, 60*(1), 97–111.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.