Information Retrieval Journal (2021) 24:240–268
https://doi.org/10.1007/s10791-021-09391-7

Topic-independent modeling of user knowledge in informational search sessions

Ran Yu[1] · Rui Tang[2] · Markus Rokicki[3] · Ujwal Gadiraju[4] · Stefan Dietze[1,3,5]

Received: 10 July 2020 / Accepted: 2 March 2021 / Published online: 16 March 2021

Abstract

Web search is among the most frequent online activities. In this context, widespread informational queries entail user intentions to obtain knowledge with respect to a particular topic or domain. To serve learning needs better, recent research in the field of interactive information retrieval has advocated the importance of moving beyond relevance ranking of search results and considering a user's knowledge state within learning oriented search sessions. Prior work has investigated the use of supervised models to predict a user's knowledge gain and knowledge state from user interactions during a search session. However, the characteristics of the resources that a user interacts with have neither been sufficiently explored, nor exploited in this task. In this work, we introduce a novel set of resource-centric features and demonstrate their capacity to significantly improve supervised models for the task of predicting knowledge gain and knowledge state of users in Web search sessions. We make important contributions, given that reliable training data for such tasks is sparse and costly to obtain. We introduce various feature selection strategies geared towards selecting a limited subset of effective and generalizable features.

Keywords Human–computer interaction · Search as learning · Knowledge gain · SAL · Online learning

Part of this work is financially supported by the Leibniz Association, Germany (Leibniz Competition 2018, funding line "Collaborative Excellence", project SALIENT [K68/2017]).

✉ Ran Yu
ran.yu@gesis.org

Stefan Dietze
stefan.dietze@gesis.org

[1] GESIS, Leibniz Institute for the Social Sciences, Cologne, Germany

[2] Ping An Technology, Beijing, China

[3] L3S Research Center, Leibniz University Hannover, Hannover, Germany

[4] Delft University of Technology, Delft, Netherlands

[5] Institute for Computer Science, Heinrich-Heine-University Dusseldorf, Dusseldorf, Germany

1 Introduction

People ubiquitously use Web search to find a variety of information and satisfy a wide range of information needs. Search sessions are commonly categorized into *navigational*, *transactional* or *informational* ones (Broder 2002). Informational search sessions involve an inherent learning intent, i.e. the desire of a user to acquire knowledge or information with respect to a particular topic, assumed to be present on one or more Web pages. In this context, the individual relevance of search results is strongly dependent on the current knowledge state of the corresponding user.

Recent research at the intersection of information retrieval and learning theory has recognized the importance of learning scopes and focused on observing and detecting learning needs during Web search. Eickhoff et al. (2014) investigated the correlation between several query and search session-related metrics and learning progress. Collins-Thompson et al. (2016) investigated the effectiveness of user interaction with respect to learning outcomes. In addition, Zhang et al. (2011) have shown that data obtained online during the search process provides valuable indicators about the domain knowledge of a user.

While the importance of learning as an implicit element of Web search has been established, recent work from Gadiraju et al. (2018) has explored the correlation between Web search behavior and a user's knowledge state and knowledge gain (i.e., a user's learning performance). Yu et al. (2018) presented an approach and model for the prediction of knowledge state as well as knowledge gain of a user using a range of behavioral signals captured during online search sessions. The proposed features pertain to queries, sessions or behavioral traces, including mouse movements and navigational activities. Supervised models were proposed using a dataset obtained through crowdsourcing of search tasks and corresponding knowledge tests. The findings from previous works demonstrate that knowledge gain/state can be predicted from user behavior throughout search sessions.

However, research so far has been constrained by limited and very specific feature sets. Insights into the generalizability of predictive models across topics are still shallow. This is particularly concerning in the light of recent work (Gadiraju et al. 2018), which has found that the correlation between search behavior and search topic is stronger than the correlation between search behavior and the corresponding knowledge indicators (knowledge gain, knowledge state).

Building on such prior works, this work introduces a novel set of Web resource-centric features and investigates their impact on the knowledge gain/state prediction task. Web resource features consider characteristics of resources a user interacts with, such as their linguistic tone, their complexity or structural aspects of an HTML page. We make valuable contributions given that reliable training data for such tasks is sparse and costly to obtain. The feature space of potentially relevant features is large: 179 distinct features (109 web resource features, 70 user behavior features) are investigated in total in our work. Thus, we introduce various feature selection strategies geared towards selecting a limited subset of effective and generalizable features by considering feature correlation with knowledge gain/state, topic-dependency of feature performance and feature redundancy.

Our experiments using data obtained through crowdsourcing studies demonstrate that resource features are essential to improve the prediction performance of knowledge gain/state in search sessions. The supervised models that we propose in this work outperform the state-of-the-art and show an average F1-score improvement by 25.5%, and an increase in accuracy by 23.2% on average across different prediction tasks. In summary, this work and provides the following novel contributions:

- *Novel feature sets* We introduce and experimentally evaluate a novel set of features (109 features in total) extracted by analyzing Web resource content for the task of knowledge state (KS) and knowledge gain (KG) prediction, which extend state-of-the-art models.
- *Feature analysis* We conduct comprehensive feature analysis assessing both generalisability of features across search topics as well as their overall effectiveness in the aforementioned prediction tasks. Findings from this analysis can inform future work for user modeling in search sessions in various ways. Moreover, our analysis can be leveraged to build computationally efficient models through a limited set of effective features.
- *Feature selection approach* In order to cope with the wide variety and large number of features in the presence of very sparse training data, we introduce a novel approach for feature selection which combines feature correlation with target variables (KG/KS) as well as the topic-dependency of feature performance. By doing so, we identify best performing features in cross-topic prediction settings and facilitate generalisable models.
- *Improved prediction models* We evaluate our features and feature selection approach by building supervised classifiers which outperform state-of-the-art baselines for the knowledge gain/state prediction on *unseen* topics. On average, our improved models outperform the previous state-of-the-art baseline (Yu et al. 2018) by 20.6, 39.9, and 16% (average F1 score) in the tasks of knowledge gain, pre-knowledge state, and post-knowledge state prediction, respectively.

It is worth noting that these contributions expand the state-of-the-art in the field, including our own prior work (Yu et al. 2018) which addresses a similar task.

Potential applications of our work include the consideration of a user's knowledge state during the retrieval and ranking step as part of state-of-the-art Web search. Our findings are equally relevant for the classification and guidance of search behavior in learning-oriented search scenarios, for example, in class rooms, libraries or work environment.

2 Related works

Understanding learning in web search Some previous works have focused on studying the correlation between learning progress and user activity features and resource features. Bhattacharya and Gwizdka (2019) investigated the relationship between users' search and eye gaze behaviours and their learning performance based on a lab study (n = 30). Gadiraju et al. (2018) described the use of knowledge tests to calibrate the knowledge of users before and after their search sessions, quantifying their knowledge gain. They investigated the impact of information needs on the search behavior and knowledge gain of users. Eickhoff et al. (2014) investigated the relation between a list of features extracted from search activities and Search Engine Result Pages (SERPs) corresponding to a search session with learning needs related to either procedural or declarative knowledge. Collins-Thompson et al. (2016) studied the influence of query types on knowledge gain, finding that intrinsically diverse queries lead to increased knowledge gain. In another work, Bhattacharya and Gwizdka (2018) studied the relation between eye-tracking measures and users' knowledge change. They found that the reading behaviors of high knowledge gain users and low knowledge gain users differ significantly. Liu and Song (2018) investigated the influence of three different types of learning resource on users' learning outcome in search sessions.

Existing works also studied the relation between the characteristics of search tasks and users' learning outcome. Vakkari (2016) provided a structured survey of features indicating learning needs as well as user knowledge and knowledge gain throughout the search process. By matching the learning tasks into different learning stages of Anderson and Krathwohl's taxonomy (Anderson et al. 2001), Jansen et al. (2009) studied the correlation between search behaviors of 72 participants and their learning stage. White et al. (2009) investigated the difference between the behavior of domain experts and non-experts in seeking information on the same topic. In a recent work, Roy et al. (2020) investigated at which time during a search session does learning occur, and found that the learning curve is largely influenced by a user's prior knowledge on the searched topic. Kalyani and Gadiraju (2019) explored this direction further by designing search tasks that fit into the different learning stages of the revised Bloom's taxonomy. Through knowledge tests before and after each search session, they found significant impact of the learning stage on a user's search behavior and knowledge gain. Liu et al. (2019) adopted mind map to capture user's knowledge change process and hence identified four types of knowledge change styles.

Studies on exploratory search have also investigated a similar set of search behaviors that influence the learning outcome. Hagen et al. (2016) investigated the relation between the writing behavior and the exploratory search pattern of writers. The authors revealed that query terms can be learned while searching and reading.

The aforementioned works consider a limited set of features or address specific learning scenarios and learning types. In this paper, we use a dataset that simulates real-world information search process and present an analysis of the relation between a large number of features and quantifiable knowledge gain across topics.

Modeling user knowledge in web search. Authors have also proposed to use features extracted from search activity to measure the user's knowledge state in an online learning environment. In a closely related work, Yu et al. (2018) proposed to use user interaction features to build classification models to predict user knowledge state and knowledge gain in search sessions. They extracted features from query term, SERP, browsing behavior and mouse movement, and selected features according to their inter-correlation and their correlation to the prediction goal. They showed that the extracted features can provide useful signals for predicting user knowledge.

Syed and Collins-Thompson (2017) proposed to optimize the learning outcome of the vocabulary learning task by selecting a set of documents that consider the keyword density and domain knowledge of the learner. Furthermore, they explored the possibility of using regression models and features extracted from user accessed document content to predict user knowledge change on vocabulary learning tasks. Experimental results indicate that document content features are effective for predicting user knowledge (Syed and Collins-Thompson 2018). Gwizdka and Chen (2016) proposed to assess learning outcomes in search environments by correlating individual search behaviors with corresponding eye-tracking measures.

Zhang et al. (2011) explored using search behavior as an indicator for the domain knowledge of a user. Through a small study ($n = 35$), they identified features such as the average query length or the rank of documents consumed from the search results as being predictive. Further, Cole et al. (2013), observed that behavioral patterns provide reliable indicators about the domain knowledge of a user, even if the actual content or topics of queries and documents are disregarded entirely. Other works have focused on detecting task difficulty in search environments based on user activity, where the subjective assessment of task difficulty is highly correlated to the user's domain knowledge (Li and Belkin 2008; Gwizdka and Spence 2006). Gwizdka and Spence (2006) showed that a searcher's perception of task difficulty is a subjective factor that depends on the domain knowledge and some other individual traits. Arguello (2014) proposed to use logistic regression to predict task difficulty in a search environment. The author used search tasks created by Wu et al. (2012), which contain task difficulty assessments on multiple dimensions, and collected search data through a crowdsourcing platform.

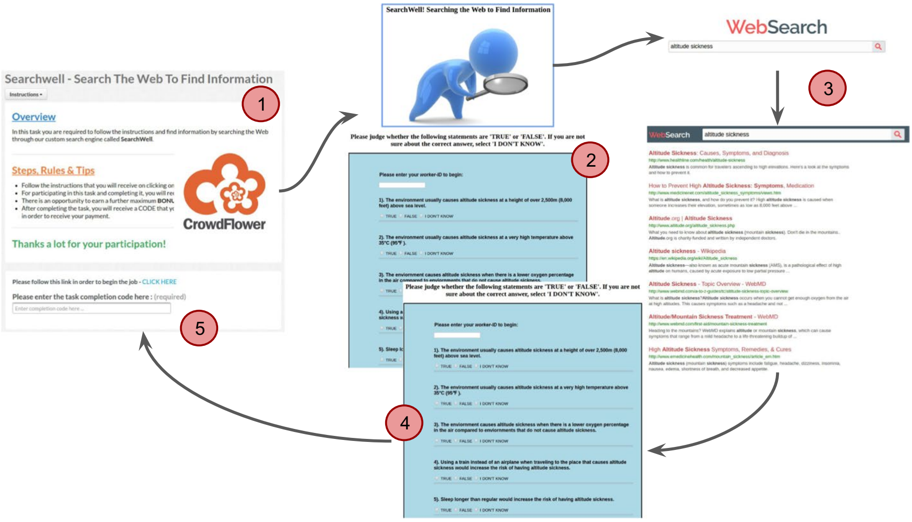
The aforementioned works ignored the influence of learning intent on the features. Previous work (Gadiraju et al. 2018) showed that the learning intent has a strong effect on user behavior. Given that the learning intent of the user in real search environments is diverse and hard to anticipate without explicit knowledge, in order to build generalizable prediction models it is necessary that the task dependency of features is taken into consideration. In this paper, we aim at developing generalizable models using topic independent features extracted from both learning resource and user interactions perspective.

3 Tasks

We adopt a definition as defined by prior work (Yu et al. 2018): an intentional learning-related *search session* comprises the sequence of a user's actions with respect to satisfying her learning intent in a Web search environment through informational queries. A user's sequence of actions begins with an initial Web query and includes browsing through the search results, click and scroll activity, navigation via hyperlinks, query reformulations, and so forth. We refer to such intentional learning-related *search session* as "session" in the remainder of this paper for simplicity.

Let s be a search session starting at time t_i and ending at time t_j aimed at satisfying a particular information need, that is, a learning intent ι of user u. In this work, we study the knowledge indicators (*KIs*): pre-knowledge state (pre-KS) $k(t_i)$, post-knowledge state (post-KS) $k(t_j)$ and knowledge gain (KG) $\Delta k(t_i, t_j)$ during time period $[t_i, t_j]$. This paper aims at extending the understanding of user knowledge (change) in the informational search process and build topic independent models (with respect to users' learning intents), to predict the aforementioned knowledge indicators. More specifically, this work addresses the following tasks:

- **[T1]** Understanding the relation between Web resource features and a user's knowledge state (*KS*) and knowledge gain (*KG*). The features we considered are described in Sect. 5.
- **[T2]** Understanding the topic-specificity of individual features, i.e. dependency between feature performance and information needs (topics), investigating feature selection strategies geared towards selecting effective and topic-independent features for modeling *KIs*. The investigated features include the document features described in Sect. 5, as well as the user interaction features studied by previous works (Gadiraju et al. 2018; Yu et al. 2018).
- **[T3]** Build generalizable models that can be used in real-world search environments on unseen topics for predicting the *KIs* from both behavioural and resource-centric features. We aim to classify a specific *KI*, e.g. knowledge gain $\Delta k(t_i, t_j)$, with respect to a particular information need. For the sake of this work, a user's knowledge state is defined by the user's capability to correctly respond to a set of questions about the corresponding information need. A user's knowledge gain is defined as the improvement of user capability (accuracy) to correctly respond to a set of



Searchwell - Search The Web To Find Information

Instructions ▴

Overview

In this task you are required to follow the instructions and find information by searching the Web through our custom search engine called **SearchWell**.

Steps, Rules & Tips

- Follow the instructions that you will receive on clicking on
- For participating in this task and completing it, you will re
- There is an opportunity to earn a further maximum **BONU**
- After completing the task, you will receive a CODE that yo in order to receive your payment.

CrowdFlower

Thanks a lot for your participation!

Please follow this link in order to begin the job - CLICK HERE

Please enter the task completion code here : (required)

Enter completion code here ...

SearchWell! Searching the Web to Find Information

Please judge whether the following statements are 'TRUE' or 'FALSE'. If you are not sure about the correct answer, select 'I DON'T KNOW'.

Please enter your *worker-ID* to begin:

1). The environment usually causes altitude sickness at a height of over 2,500m (8,000 feet) above sea level.

○ TRUE ○ FALSE ○ I DON'T KNOW

2). The environment usually causes altitude sickness at a very high temperature above 35°C (95℉).

○ TRUE ○ FALSE ○ I DON'T KNOW

3). The enviornment causes altitude sickness when there is a lower oxygen percentage in the air compared to enviornments that do not cause altitude sickness.

○ TRUE

4). Using a

sickness w

○ TRUE

5). Sleep lo

○ TRUE

WebSearch

altitude sickness

WebSearch altitude sickness

Altitude Sickness: Causes, Symptoms, and Diagnosis
http://www.healthline.com/health/altitude-sickness
Altitude sickness is common for travelers ascending to high elevations. Here's a look at the symptoms and how to prevent it.

How to Prevent High **Altitude Sickness: Symptoms**, Medication
http://www.medicinenet.com/altitude_sickness_symptoms/views.htm
What is **altitude sickness**, and how do you prevent it? High **altitude sickness** is caused when someone increases their elevation, sometimes as low as 8,000 feet above ...

Altitude.org | **Altitude Sickness**
http://www.altitude.org/altitude_sickness.php
What you need to know about **altitude sickness** (mountain **sickness**). Don't die in the mountains.. **Altitude**.org is charity-funded and written by independent doctors.

Altitude sickness - Wikipedia
https://en.wikipedia.org/wiki/Altitude_sickness
Altitude sickness—also known as acute mountain **sickness** (AMS), is a pathological effect of high **altitude** on humans, caused by acute exposure to low partial pressure ...

Altitude Sickness - Topic Overview - WebMD
http://www.webmd.com/a-to-z-guides/tc/altitude-sickness-topic-overview
What is **altitude sickness?Altitude sickness** occurs when you cannot get enough oxygen from the air at high altitudes. This causes symptoms such as a headache and not ...

Altitude/Mountain Sickness Treatment - WebMD
http://www.webmd.com/first-aid/mountain-sickness-treatment
Heading to the mountains? WebMD explains **altitude** or mountain **sickness**, which can cause symptoms that range from a mild headache to a life-threatening buildup of ...

High **Altitude Sickness** Symptoms, Remedies, & Cures
http://www.emedicinehealth.com/mountain_sickness/article_em.htm
Altitude sickness (mountain **sickness**) symptoms include fatigue, headache, dizziness, insomnia, nausea, edema, shortness of breath, and decreased appetite.

Please judge whether the following statements are 'TRUE' or 'FALSE'. If you are not sure about the correct answer, select 'I DON'T KNOW'.

Please enter your *worker-ID* to begin:

1). The environment usually causes altitude sickness at a height of over 2,500m (8,000 feet) above sea level.

○ TRUE ○ FALSE ○ I DON'T KNOW

2). The environment usually causes altitude sickness at a very high temperature above 35°C (95℉).

○ TRUE ○ FALSE ○ I DON'T KNOW

3). The enviornment causes altitude sickness when there is a lower oxygen percentage in the air compared to enviornments that do not cause altitude sickness.

○ TRUE ○ FALSE ○ I DON'T KNOW

4). Using a train instead of an airplane when traveling to the place that causes altitude sickness would increase the risk of having altitude sickness.

○ TRUE ○ FALSE ○ I DON'T KNOW

5). Sleep longer than regular would increase the risk of having altitude sickness.

○ TRUE ○ FALSE ○ I DON'T KNOW

Fig. 1 Workflow of participants in the experimental setup orchestrating informational search sessions

test questions about the corresponding information need. The classification goal is introduced in Sect. 4.

4 Dataset

In order to address the aforementioned tasks (Sect. 3), we adopt an existing dataset[1] which has been used for our previous works on understanding and predicting user knowledge state based on user interaction features (Gadiraju et al. 2018; Yu et al. 2018). We have extended the study using the same task setup to include 1100 search sessions conducted by crowd workers spanning across 11 information needs for different topics randomly selected from the *TREC 2014 Web Track*[2] dataset. This includes knowledge assessment data before and after each of the 100 search sessions per information need. In this section, we present the data collection process in details as described in Gadiraju et al. (2018).

4.1 Study design

We recruited participants from CrowdFlower,[3] a premier crowdsourcing platform. At the onset, workers were informed that the task entailed ‘searching the Web for some information’. Workers willing to participate were redirected to our external platform, *SearchWell*. Figure 1 presents the workflow of participants in the experimental setup orchestrating informational search sessions, which consists of 5 steps: (1) Workers are recruited from the

[1] https://github.com/PL8aDSah9l/AnalysingKnowledgeGain.

[2] http://www.trec.nist.gov/act_part/tracks/web/web2014.topics.txt.

[3] Formerly http://www.crowdflower.com/, now https://www.figure-eight.com/.

CrowdFlower platform, and those willing to participate are redirected to *SearchWell*. (2) Participants are asked to answer a few questions (*knowledge test*) regarding a topic; this is used to calibrate their knowledge before the search session. (3) Participants indulge in an informational search session to satisfy a well-defined information need. (4) Participants are asked to complete a post-session test that is identical to the calibration test. (5) Participants receive a completion code, which they enter on CrowdFlower to claim their reward.

Workers were first asked to respond to a few questions (technically referred to as ‘*items*’) corresponding to a particular topic without searching the Web for answers. The questions took the form of statements pertaining to a topic, and workers had to select whether the statement was ‘TRUE’, ‘FALSE’, or ‘I DON’T KNOW’ in case they were not sure. In this way, we calibrated the knowledge of users corresponding to a given topic. To encourage the workers to respond without external consultation, we informed them that their responses to these questions would not affect their pay. We also encouraged workers to provide responses to the best of their knowledge and avoid guessing. The results of this pre-test were used to calibrate the knowledge of the workers with respect to the topic. We describe the topics and how the knowledge tests were created in the following Sect. 4.2. On completing the knowledge calibration test, workers were presented with their actual task.

Workers were presented an *information need* corresponding to the topic of the calibration test they completed. They were told to use the *SearchWell* platform to search the Web and satisfy their information need. To incentivize workers towards realistic attempts to learn about the topic, we informed them that they will have to complete a final test on the topic to successfully finish the task. Furthermore, workers were conveyed the message that depending on their accuracy on the final test they could earn a bonus payment. We subsequently logged all the activities of the workers (mouse movements, key presses and clicks) within the *SearchWell* platform. Workers were allowed to begin the final test anytime after a search session, which is when a link to the final test was made available. Workers were encouraged to proceed to the next stage only once they felt that their information need was satisfied and when they were ready for the post-session test. On completing the post-session test, workers received a unique code that they could enter on CrowdFlower to claim their reward.

We restricted the participation to workers from English-speaking countries to ensure that they understood the task and instructions adequately (Gadiraju et al. 2017). To ensure reliability of the resulting data, we restricted the participation to *Level-3 workers*[4] on CrowdFlower.

4.2 Topics:defining information needs

We constructed a corpus of topics representing varying scopes of information needs (with some relatively broader than others). Topics were selected randomly from the *TREC 2014 Web Track* dataset,[5] and corresponding information needs were defined accordingly. In all cases, the knowledge of users before beginning an informational search session was assessed using pre-tested and evaluated *knowledge tests*. Knowledge

[4] *Level-3 contributors* on CrowdFlower comprise workers who completed over 100 test questions across hundreds of different types of tasks, and have a near perfect overall accuracy. They are workers of the highest quality on CrowdFlower.

[5] http://www.trec.nist.gov/act_part/tracks/web/web2014.topics.txt.

Table 1 Topics and corresponding information needs presented to participants in the informational search sessions

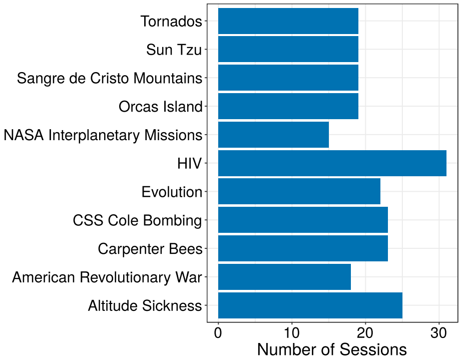
Topic	Information need
1. Altitude sickness	In this task you are required to acquire knowledge about the symptoms, causes and prevention of altitude sickness. (20 items)
2. American revolutionary War	In this task, you are required to acquire knowledge about the 'American Revolutionary War'. (10 items)
3. Carpenter bees	In this task, you are required to acquire knowledge about the biological species 'carpenter bees'. How do they look? How do they live? (10 items)
4. Evolution	In this task, you are required to acquire knowledge about the theory of evolution. (12 items)
5. NASA interplanetary missions	In this task, you are required to acquire knowledge about the past, present, and possible future of interplanetary missions that are planned by the NASA. (20 items)
6. Orcas Island	In this task you are required to acquire knowledge about the Orcas Island. (20 items)
7. Sangre de Cristo mountains	In this task, you are required to acquire knowledge about 'Sangre de Cristo' mountain range. (10 items)
8. Sun Tzu	In this task, you are required to acquire knowledge about the Chinese author Sun Tzu - about his life, his writings, and his influence to the present day. (15 items)
9. Tornado	In this task, you are required to acquire knowledge about the weather phenomenon that is called 'tornado' (20 items)
10. USS cole bombing	In this task, you are required to acquire knowledge about the 2000 terrorist attack that came to be known as the 'USS Cole bombing'. (10 items)
11. HIV	In this task, you are required to acquire knowledge about the transmission, prevention and consequences of HIV. (45 items)

tests are scientifically formulated tests that measure the knowledge of a participant on a given topic (for example, the HIV knowledge test Carey et al. 1997).

Table 1 presents the topics and corresponding information needs considered for orchestrating the informational search sessions. Except for topic *11. HIV*, knowledge on all given topics was measured using knowledge tests comprising of between 10 and 20 items. Knowledge test of the topic *11. HIV* contains 45 items that are filtered from the list of items created by Carey et al. (1997) based on the pilot test as described below. The answer options were in all cases 'TRUE', 'FALSE', and 'I DON'T KNOW'. The differences in the number of items reflects our attempt to feature varying scopes of information needs; relatively narrow (e.g., *Carpenter Bees*—10 items) as well as broad (e.g., *NASA Interplanetary Missions*—20 items). In the construction of all scales, an item pool comprising of more items than finally used was constructed. After a pilot test with 100 distinct participants recruited via CrowdFlower for each of the 10 topics, items that proved to be either too easy (e.g., more than 80% correct answers) or too hard/ambiguous (e.g., more false than true answers) were discarded.

We built *SearchWell* on top of the Bing Web Search API. We logged user activity on the platform including mouse movements, clicks, and key presses, using PHP/Javascript and the jQuery library.

Fig. 2 Number of search sessions pertaining to each topic and the associated information need after filtering



4.3 Data cleaning

To ensure reliability of responses and the resulting behavioral data logged during the search sessions, we filtered sessions using the following criteria.

(1) From 1100 workers, we filtered out 263 workers who did not complete the post-session test, or did not issue at least 1 search query.
(2) We further filtered out 89 workers who selected the same option; either ‘YES’, ‘NO’ or for all items in the calibration test or the post-session test.
(3) From the rest of the sessions, we filtered out 58 workers who did not click on any results on the SERPs. Since the aim of this work is to further the understanding of how the knowledge state of a user evolves in the process of interacting with the Web resources, we discard those users who did not access any Web pages.
(4) From the remaining sessions, we filtered 457 workers who interacted with at least 1 non-English resource. The rationale behind considering this filter was that many of the features (see Sect. 5) we extracted from the Web resource content rely on certain dictionaries, which are currently only available for the English language or not comparable across different language versions.

After applying all the aforementioned filters, we retain 233 search sessions, with 1.361 queries and 2.622 clicks per session on average. Figure 2 shows the number of Web search sessions corresponding to each topic. The topic “HIV” has 31 sessions, i.e. the largest number of sessions. The topic “NASA” has only 15 sessions. The mean number of Web search sessions for each topic is 21.18.

Knowledge state and knowledge gain classes The pre (post)-knowledge score of a user in search sessions corresponding to a topic is measured as the percentage of the correct answers on the knowledge test that a given user has completed. Correspondingly, the knowledge gain is measured as the difference between a user’s pre- and post-search session knowledge score.

For the classification tasks described in [T3], we follow the same approach as used in previous work (Yu et al. 2018), i.e. a *Standard Deviation Classification* approach to obtain three classes of learners with regard to their level of knowledge. Assuming approximately normal distributions of the respective test scores (X) for the different topics, we transformed the test scores into Z-scores with a mean of 0 and a Standard Deviation (SD) of 1 (standardization). We then used statistically defined intervals (low: $X < -0.5\,\mathrm{SD}$;

Table 2 Knowledge state and knowledge gain classes created based on thresholds of *mean* ± *0.5 SD*

Task	Mean	SD	Low	Moderate	High
Pre-knowledge state	0.36	0.255	87	52	94
Post-knowledge state	0.66	0.174	61	95	77
Knowledge gain	0.23	0.208	84	84	65

moderate: $-0.5\,\mathrm{SD} < \mathrm{X} < 0.5\mathrm{SD}$; high: $0.5\,\mathrm{SD} < X$) for the classification of the learners into roughly equal groups with low, moderate, or high pre-KS. The same procedure was repeated for post-KS and KG. Table 2 shows the resulting numbers of learners for the respective classes and underlying statistics.

5 Feature extraction

We approach the problem of predicting *KI* as described in Sect. 3 with supervised classification models, where details about the applied models are given in Sect. 6.1. Each session s is represented by a feature vector $\mathbf{v} = (f_1, f_2, \dots, f_n)$; where the features considered are described in the following subsections.

5.1 Web resource features

We introduce 109 Web resource features in total. Given space limitations, we discuss only a subset of features. The full set of features are publicly accessible at the following link.[6]

Document complexity features The assumption behind the document complexity related features is that, the higher a user's knowledge state is on a topic, it is more likely that the user prefers documents with higher complexity. As previously reported (Eickhoff et al. 2014), the number of words (*c_word*) can be an indicator for content complexity. Moreover, long words (*c_char*) are more likely to be specific and indicative of complex vocabularies than short words. Similarly, long sentences (*c_sentence*) have been found to indicate higher resource complexity than short sentences.

The syntactic structure of a document, which is represented by the ratio of the number of nouns, verbs, adjectives, or *other words* (i.e. words that are not verb, noun or adjective) to the total words ($c\_\{noun, verb, adj, oth\}$), is likely to suggest the intention and complexity of its content (Heilman et al. 2007).

The grade level readability index can be used to measure the readability of a document by computing a score based on the number of the syllables in words. The assumption is that it requires a higher education level to read a document with a higher score (Horne and Adali 2017). We compute three different readability grades: Gunning Fog Grade[7] (*c_gi*), SMOG (Mc Laughlin 1969) (*c_smog*) and Flesch–Kincaid Grade (Kincaid et al. 1975) (*c_fk*). Using the age-of-acquisition (AoA) dictionary proposed by Kuperman et al. (2012) that contains a listing of more than 30,000 English words along with the age at which

[6] https://github.com/hwtroow8/TIMoUK.

[7] http://gunning-fog-index.com/.

native speakers typically learn the term, we compute the age-of-acquisition across all words on Web pages (*c_aoa*), which provides another indicator of document complexity.

HTML structural features Previous works (Syed and Collins-Thompson 2018) have investigated the influence of images on user's learning outcome in Web search, they found a positive correlation between a relevant image and KG, and a negative correlation between the total number of images and KG. Here we do not distinguish between the relevant and irrelevant images due to the current lack of an automated approach that can be applied in a real-world search environment. We hence compute the number of <*img*> elements on Web pages to estimate the number of images (*h_img*) it contains.

Prior work (DeStefano and LeFevre 2007) found a negative association between the number of hyperlinks embedded in a Web page and the user's KG. The assumption is that people may not focus on the content in the presence of too many embedded links. We quantify the number of outbound links by counting the <*a*> elements (*h_link*). The average length of each paragraph (*h_p*) is one of the indicators of the required effort for understanding the resource (DeStefano and LeFevre 2007). The <*ul*> elements embedded (*h_oth_ul*) are often used to present important ideas of the document as an unordered list in a more structured and easily digestible fashion, and thus, may have a positive impact on KG. The <*script*> element is used to define a client-side script (e.g. JavaScript). Based on our observation, different types of websites adopt different styles of using scripts, e.g. Wikipedia uses far fewer scripts than typical commercial websites. We assume that the presence of scripts might be correlated with the possibility of whether a website suits learning-oriented needs, and therefore correlates with KG. We compute the number of scripts (*h_script*) on a Web page to serve as a feature.

Linguistic features We make use of the 2015 Linguistic Inquiry and Word Count (LIWC) dictionaries[8] to compute the features in this category. According to previous work (Horne and Adali 2017), the amount of words on Web pages that are correlated with different psychological processes and basic sentiment can influence a learner's cognitive state. Based on this assumption, we extracted 56 features. Due to space limitations, we only show the features in this list of features that are discussed in the paper in Table 3, where notations begin with *l_* in the *Linguistic* category.

The stylistic features capture grammatical characteristics, text style, and syntax of a document (Horne and Adali 2017). The writing style could affect the readability of a learning resource and the engagement of readers. We compute 35 relevant features using the LIWC dictionary. We show the features that are discussed in the paper from this feature list (features in Table 3 with notations beginning with *ls_* in the *Linguistic* category). All features in this category are named (in Table 3) according to the label generated by LIWC dictionary.

5.2 User behavior features

Apart from the resource content-related features introduced above, we also consider the 70 user behavior-related features that were introduced through prior work (Yu et al. 2018). The user behavior features were extracted according to multiple dimensions of a search session, namely features related to the session, queries, SERP, browsing behavior and mouse movements. The SERP category consists of features extracted from direct interactions with

[8] http://liwc.wpengine.com/.

Table 3 Considered Web resource features and user behavior features (a subset), bold value cells having *p-value* ≥ 0.05

	Feature name	Corr			SDoC			Feature description
		Pre	Post	KG	Pre	Post	KG	
Complexity	c_adj	−0.287	−0.329	**0.076**	0.166	0.230	0.187	Ratio of the number of adjectives to the total word number
	c_aoa	0.265	0.199	−0.157	0.227	0.177	0.245	Average AoA rating of words in each Web page
	c_char	**−0.077**	**−0.084**	**0.024**	0.180	0.187	0.193	Average number of characters per term
	c_fk	−0.174	−0.203	**0.043**	0.236	0.155	0.284	Flesch-Kincaid Grade Readability Index
	c_gi	**−0.062**	**0.092**	0.153	0.257	0.272	0.235	Gunning Fog Grade Readability Index
	c_noun	−0.165	**0.001**	0.203	0.153	0.179	0.118	Ratio of the number of nouns to the total word number
	c_oth	**0.117**	**−0.007**	−0.149	0.143	0.150	0.105	Ratio of the *other word* number to the total word number
	c_sentence	**−0.024**	**−0.006**	**0.024**	0.230	0.181	0.257	Average number of words per sentence
	c_smog	**−0.041**	**0.088**	**0.124**	0.230	0.248	0.202	SMOG Readability Index
	c_uniq_word	**0.015**	0.164	**0.118**	0.202	0.178	0.229	Ratio of the unique word number to the total word number
	c_verb	0.342	0.234	−0.222	0.188	0.235	0.199	Ratio of the number of verb to the total word number
	c_word	−0.191	−0.243	**0.030**	0.196	0.199	0.190	Number of words in each Web page
HTML	h_img	−0.187	−0.273	**0.001**	0.204	0.191	0.191	Number of <img> elements
	h_link	**−0.099**	−0.240	**−0.079**	0.190	0.221	0.192	Number of outbound links
	h_nav_ul	0.206	0.184	**−0.098**	0.249	0.172	0.238	Number of <ul> elements embedded in <nav> elements
	h_oth_ul	−0.243	−0.305	**0.043**	0.218	0.148	0.187	Number of <ul> elements not in <nav> elements
	h_p	−0.281	−0.230	0.151	0.205	0.255	0.238	Average length of each paragraph in <p> elements
	h_script	0.165	**0.067**	−0.145	0.221	0.141	0.245	Number of <script> elements

Table 3 (continued)

	Feature name	Corr			SDoC			Feature description
		Pre	Post	KG	Pre	Post	KG	
Linguistic	l_Analytic	−0.469	−0.294	0.329	0.192	0.224	0.172	Number of analytic words
	l_anger	−0.250	−0.151	0.179	0.172	0.210	0.090	Number of anger words
	l_bio	0.564	0.364	−0.386	0.238	0.207	0.281	Number of biological process words
	l_body	0.469	0.336	−0.293	0.236	0.186	0.264	Number of body words
	l_focuspresent	0.514	0.302	−0.376	0.263	0.226	0.240	Number of present focus words
	l_health	0.556	0.351	−0.387	0.186	0.217	0.259	Number of health words
	l_money	−0.156	**−0.026**	0.169	0.130	0.156	0.188	Number of money words
	l_relativ	−0.264	−0.295	0.077	0.226	0.285	0.209	Number of relativity words
	l_relig	−0.121	−0.190	−0.011	0.160	0.292	0.248	Number of religion words
	ls_article	−0.297	−0.294	**0.118**	0.187	0.170	0.174	Number of articles
	l_percept	−0.043	−0.039	0.020	0.240	0.172	0.316	Number of perceptual processes
	ls_conj	0.422	0.184	−0.362	0.208	0.182	0.182	Number of conjunctions
	ls_Dic	0.398	0.164	−0.349	0.213	0.197	0.243	Number of dictionary words
	ls_number	−0.358	−0.227	0.248	0.240	0.238	0.246	Number of numbers
	ls_Quote	−0.339	−0.320	0.147	0.208	0.158	0.260	Number of quotation marks
	ls_you	0.473	0.337	−0.297	0.155	0.247	0.201	Number of you pronouns

Table 3 (continued)

	Feature name	Corr			SDoC			Feature description
		Pre	Post	KG	Pre	Post	KG	
User behavior	b_revisited_ratio	**0.002**	**−0.043**	**−0.038**	0.292	0.180	0.357	Ratio of revisited pages
	b_time_per_q	−0.144	**0.035**	0.205	0.306	0.257	0.264	Average active time on the browsed pages per query
	b_ttl_len	0.397	0.281	−0.251	0.203	0.190	0.258	Average page title length
	m_rank_max	**−0.027**	**0.110**	**0.126**	0.194	0.187	0.315	max mouseover rank in the session
	m_scroll_dist	**−0.030**	**−0.011**	**0.028**	0.278	0.220	0.353	total scroll distance in session
	q_len_first	**0.033**	**0.002**	**−0.039**	0.180	0.298	0.197	First query length
	q_uniqT_first	**0.023**	**0.019**	**−0.012**	0.179	0.283	0.184	Number of unique terms of first query
	q_uniqT_ratio	−0.185	**−0.085**	0.154	0.292	0.304	0.234	$\frac{\text{Number of unique query terms}}{\text{number of query terms}}$
	s_duration	**0.113**	**0.087**	**−0.066**	0.380	0.262	0.351	Duration of the search session of a worker on a given topic
	s_duration_per_q	**0.113**	**0.087**	**−0.065**	0.312	0.234	0.288	Session duration per query

SERP items, while the browsing category consists of features extracted from subsequent user navigation beyond simple SERP clicks. In this section, we describe the motivation behind the extraction of the user behavior features. We listed the user behavior features that are discussed in the remainder of this paper in Table 3. We also give the full list of user interaction features online.[9]

Session features The relation between session duration (*s_duration*) and different stages of learning has been discussed by Jansen et al. (2009). It has been shown that there is a difference in the duration of sessions among the classifications in Anderson and Krathwohl's taxonomy (Anderson et al. 2001). White et al. (2009) also found that the sessions conducted by domain experts were generally longer than non-expert sessions.

Query features Several prior works (Jansen et al. 2009; Arguello 2014; White et al. 2009) have investigated the correlation between query activities in a search session and learning performance. Based on the study by White et al. (2009), the number of queries applied by experts and non-experts show big differences across domains: non-expert users usually run significantly more queries than experts. Jansen et al. (2009) also found that the number of queries applied on learning tasks classified as applying stage was significantly different from other learning stages.

The length of queries has been found to have a strong correlation with learning outcome by Zhang et al. (2011). Their study shows that the average query length and user domain knowledge is correlated with a Pearson correlation score of 0.344.

The complexity of queries has been investigated by Eickhoff et al. (2014), and has been found to evolve during the learning process. We applied the same query complexity measure as in Eickhoff et al. (2014), which is computed based on the dictionary created by Kuperman et al. (2012) that contains a listing of more than 30,000 English words along with the age at which native speakers typically learn the term. The maximum age of acquisition across all query terms is used as query complexity.

Furthermore, the investigation from Arguello (2014) shows that beside the number of total terms, the number of unique terms in the session is strongly correlated with knowledge level on the task, while the number and ratio of stop words do not have a big difference when comparing between search sessions with different levels of domain knowledge.

As we aim at predicting knowledge state change during a session, similarly to the features discussed above, we extract the features – the length of first and last queries, and the number of unique terms of first and last queries, which potentially are indicators of the knowledge level at the beginning and end of the session.

SERP activity features Some activities on SERP have also been investigated by previous works. Specifically, Collins-Thompson et al. (2016) found that the total number of clicks on SERP is strongly correlated with a user's understanding of the topic. The analysis shows that users tend to click more often when having stronger interest in the topic.

The ranking position of the clicked URL on SERP has also been shown to be a strong indicator of user domain knowledge by Zhang et al. (2011). Arguello (2014) the authors discovered that the difficult tasks with which a user is less knowledgeable are associated with more clicks, more clicks on lower ranks, more abandoned queries, i.e. queries without clicks, longer time till first click and longer time till next click.

Browsing features Browsing features such as the number of documents viewed and the average number of documents viewed per query were shown by several previous

[9] https://github.com/hwtroow8/TIMoUK.

works (Eickhoff et al. 2014; Jansen et al. 2009; Arguello 2014; Gwizdka and Spence 2006) to be positively correlated with the knowledge improvement. More detailed features corresponding to the browsing behavior have also been studied, indicating that the more difficult a task is for a user, the higher the ratio of revisited pages (*b_revisited_ratio*) is.

Despite the number of pages visited, the time spent on the accessed pages are found to vary to a large extent between domain expert and non-expert (White et al. 2009). Features corresponding to the time spent on browsing SERP pages were also shown to be effective for predicting the user's assessment of task difficulty (Arguello 2014), which is subject to the user's knowledge state.

We further distinguish the viewed pages into two sets {pages navigated through SERP, pages navigated through non-SERP}, by parsing its ancestor page. Motivated by the features introduced above, we extract the number and the percentage of pages in a session that belong to those two sets respectively.

The content of the accessed Web documents strongly influence the user's learning outcome. White et al. (2009) found that domain experts encountered different and more diverse domains (number of distinct domains) than domain novices. Several other document content related features: page size, title length have also been found to evolve during the learning process (Eickhoff et al. 2014). Based on the assumption that domain experts and novices have different capabilities of choosing learning resources, for instance, experts are able to recognize useful documents without query terms presented in the page title, we computed features based on the overlap between page title and query. The page URL as a complementary source containing hints about a page's content has also been considered in the feature extraction process.

Mouse features Features in the *Mouse* category are indicators of quantity and quality of user interactions with a knowledge source and were also shown to be effective for predicting the user's assessment of task difficulty (Arguello 2014).

5.3 Feature selection strategies

For the classification tasks, we consider all 109 resource content-related features and 70 user behavior-related features as described in Sect. 5, denoted as F. However, due to the difficulty of obtaining user knowledge assessment data, the scale of training/testing data is limited. Hence, feature selection is important for building reliable models, and in particular, to avoid overfitting. The goal of this step is to select a set of features $F' \subseteq F$ that can produce the best performing model for the prediction of a specific knowledge indicator. In this section, we introduce 3 feature selection strategies. For the sake of simplicity, we refer to all knowledge indicators, i.e. pre-KS $k(t_i)$, post-KS $k(t_j)$ or KG $\Delta k(t_i, t_j)$ as KI in the following.

$Corr(f_i, KI)$: *feature effectiveness* We compute the Pearson correlation coefficient between each feature and the knowledge indicator $Corr(f_i, KI)$ across all sessions. The correlation scores are shown in Table 3. To ensure effectiveness of features, we select features fulfilling the condition $Corr(f_i, KI) \geq \alpha$ for the prediction task. Performance of the prediction model using features selected based on varied α has been evaluated and corresponding results are presented in Sect. 7.

$SDoC(f_i, KI)$: *generalizability* In order to measure the topic dependency of features, we compute the correlation between a feature and a knowledge indicator for each topic and measure the standard deviation (SD) of the correlation score across topics. The intuition is that a small standard deviation of the correlation between a feature and the respective KI is indicative for a topic independent relationship that may generalize to other topics as well. For simplicity, we will refer to this metric as *SDoC* (Standard Deviation of Correlation) in the remaining of this paper. The computation of *SDoC* of feature f_i is shown in Equation 1.

$$SDoC(f_i, KI) = \sqrt{\frac{\sum_{j=1}^{N}\left(Corr^{(j)}(f_i, KI) - \frac{\sum_{j=1}^{N} Corr^{(j)}(f_i,KI)}{N}\right)^2}{N-1}} \tag{1}$$

where N is the number of topics in the sample data set (here $N = 11$), $Corr^{(j)}(f_i, KI)$ is the correlation between f_i and KI on the sample data corresponding to topic j. To improve the generalizability of the model, we remove topic dependent features using the *SDoC* metric, that is, we keep only the features with $SDoC(f_i) < \beta$ on the respective knowledge indicator.

$Corr(f_i,f_j)$: *feature redundancy* We compute the Pearson correlation coefficient $Corr(f_i,f_j)$ between each pair of features in the feature set. If $Corr(f_i,f_j) \geq \tau$, i.e. features do not appear to be independent from each other, we remove the feature of the feature pair which has a lower $Corr(f_i, KI)$ for the corresponding prediction.

6 Experimental setup

6.1 Approach, configurations and baseline

Classifiers We apply a range of standard models for the classification tasks, namely, Naive Bayes (*nb*), Logistic Regression (*lr*), Support Vector Machine (*svm*) and Random Forest (*rf*). For our experiments, we used the *scikit-learn* library for Python.[10] We tune hyperparameters of the algorithms using grid search within the repeated cross-topic validation setup described in Sect. 6.2.

Feature category In order to evaluate the influence of resource content-related features on the prediction of *KIs*, we compare between the performance of the prediction models using: (1) only user behavior features (feature category *UB*), (2) only Web resource features (feature category *WR*) and (3) using both user behavior and Web resource features (feature category *WR&UB*).

Feature selection strategy We test a range of thresholds for selecting the features according to the strategies introduced in Sect. 5.3. Specifically, for the feature selection based on $Corr(f_i, KI)$, we apply $\alpha \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$; for the selection based on $SDoC(f_i, KI)$, we apply $\beta \in \{0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4\}$; for the selection based on $Corr(f_i,f_j)$, we apply $\tau \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$. The thresholds α, β, τ are treated as hyperparameters of the knowledge prediction model, and is tuned using the repeated cross-topic validation in the model fitting process (see Sect. 6.2). Some combinations of γ, β, τ which reduce the feature set to an empty set are excluded in the experiment.

Baseline We compare our approach against prior work (Yu et al. 2018), in which classifiers were built using user interaction features (Sect. 5.2) to predict KG and post-KS. Their model achieved best performance when using Random Forest as classifier and when applying certain thresholds on the feature-indicator-correlation and the

[10] http://scikit-learn.org.

between-feature-correlation. In the repeated cross validation process of our experiments, we tune the hyperparameters of their model again using grid search to ensure a fair comparison.

6.2 Evaluation method and metrics

In order to estimate the performance and generalisation of pretrained and pretuned models on unseen topics, we conduct a repeated cross-topic validation consisting of an inner loop, where hyperparameters are tuned, and an outer loop, for the actual cross-topic performance assessment. Instead of randomly splitting the experimental dataset into training and testing set, we split the search sessions in our dataset according to the topic of a search session. More specifically, for the repeated cross validation, we run 11 iterations in the outer loop, for each run, we use the session data corresponding to one topic for testing, and the rest of the sessions for training and validation. Similar to the outer loop, the inner loop consists of 10 iterations, for each run, the session data corresponding to one topic is used for validation, the session data corresponding to the remaining 9 topics are used for model training. Hyperparameters of the classifiers as well as the feature selection thresholds α, β and τ are tuned in the inner loop. We evaluate the results according to the following metrics:

- *Accuracy (Accu) across all classes*: percentage of search sessions that were classified with the correct class label.
- *Precision (P), Recall (R), F1 ($F1$) score of class i*: we compute the standard precision, recall and F1 score on the prediction result of each class i.
- *Macro average of precision (P), recall (R), and F1 ($F1$)*: the average of the corresponding score across 3 classes.

7 Results

To evaluate the performance of our approach, we tune the hyperparameters according to the average F1 score through repeated cross-topic validation, as described in Sect. 6. We present the result of the configuration in terms of classifier and feature category that produces the highest average F1 score for each prediction task in Table 4. Our main findings are discussed below.

Overall performance Using our approach, accuracy scores are above 0.464 for all 3 prediction tasks and the average F1 scores are above 0.465. Compared to the baseline, we observe improvements for all 3 prediction tasks, with the highest improvements for the pre-KS prediction task, where the average F1 score is 43.9% higher and the accuracy is 26.8% higher.

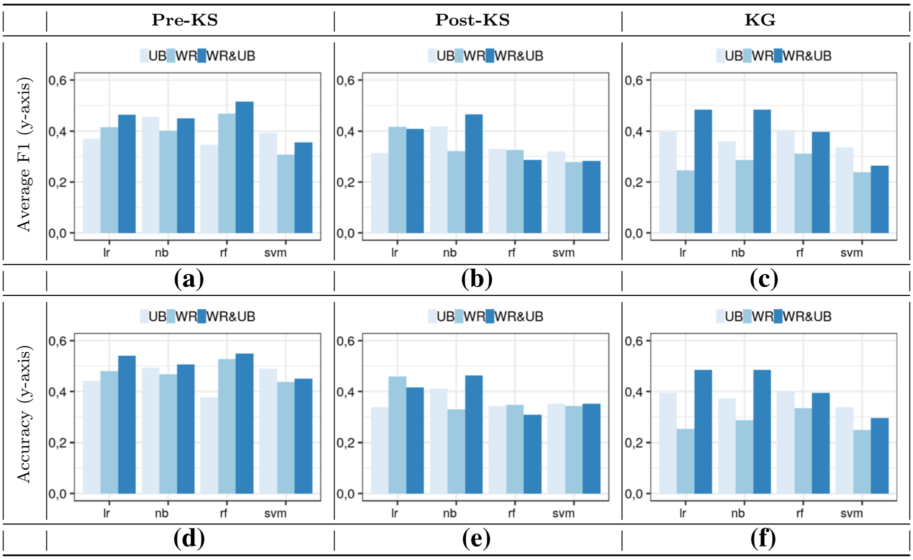
Knowledge indicator classes Compared to the baseline, for pre-KS, our model shows particular improvements in F1 score for the moderate class, indicating that the resource features allow for better identifying medium knowledge state compared to the user behavior features. For post-KS our model shows similar improvements for all three classes. For knowledge gain, our model shows greater improvements for low and moderate KG classes.

The best performance with respect to both average F1-score and overall accuracy has been achieved for the pre-KS prediction, indicating that predicting the user's knowledge state on the search topic before the search session is a easier task compared to predicting the other two KIs. This is intuitive as the interactions such as input queries and the resource

Table 4 Best performing results of different approaches according to average F1 score

KI	Approach	Feature cat.	clf	Low			Moderate			High			Average			Accu
				P	R	F1	P	R	F1	P	R	F1	P	R	F1	
pre	New baseline	WR&UB	rf	0.600	0.621	0.610	0.296	0.308	0.302	0.652	0.617	0.634	0.516	0.515	**0.515**	**0.549**
		–	rf	0.442	0.529	0.482	0.146	0.115	0.129	0.511	0.479	0.495	0.367	0.374	0.368	0.416
post	New baseline	WR&UB	nb	0.367	0.590	0.453	0.513	0.411	0.456	0.559	0.429	0.485	0.480	0.476	**0.465**	**0.464**
		–	rf	0.320	0.508	0.392	0.417	0.368	0.391	0.519	0.351	0.419	0.418	0.409	0.401	0.399
KG	New baseline	WR&UB	lr	0.578	0.440	0.500	0.425	0.571	0.487	0.500	0.431	0.463	0.501	0.481	**0.483**	**0.485**
		–	rf	0.437	0.369	0.400	0.368	0.381	0.374	0.400	0.462	0.429	0.401	0.404	0.401	0.399

The F1-score and the accuracy for the best performing model of each task are marked in bold



(a) (b) (c)

(d) (e) (f)

Fig. 3 Average F1-score and accuracy for best performing classifier and respective feature category

selection are strongly affected by the user's pre-KS. While the post-KS is dependent on the pre-KS as well as the effort the user spends during the search session. Due to the short duration of the sessions in the ground truth dataset, despite using multiple features (e.g. *s_duration*, *b_time_per_q*) to capture the effort of the user, it is more challenging to distinguish the post-KS and KG classes.

With respect to the prediction performance on different classes, we observe that for the pre-KS prediction, the model performs particularly well for low and high knowledge classes. For the prediction of post-KS and KG, on the other hand, performance differences on different KI classes are less pronounced.

In summary, our approach outperforms the baseline for all prediction tasks and the resource-related features appear to provide useful information for all the prediction tasks and knowledge classes. The performance of the classifiers using different categories of features and feature selection strategies will be discussed more in the remainder of this section.

7.1 Performance of classifiers

Here we compare the performance achieved when using different classification algorithms, combined with the available feature categories, as seen in Fig. 3. As also listed in Table 4, the best performing classifier varies for different prediction tasks. The *rf* classifier achieves the highest average F1-score for pre-KS prediction, outperforming the other classifiers by at least 11.3%. The *nb* classifer achieves the highest average F1-scores for the post-KS prediction task. The *lr* classifier achieves the highest average F1-score for the KG prediction, outperforming the *nb* classifier with a 0.1% score improvement. The result is inconsistent with the finding of previous work (Yu et al. 2018) where *rf* was the best performing classifier for both post-KS and KG prediction. The reasons behind might be:

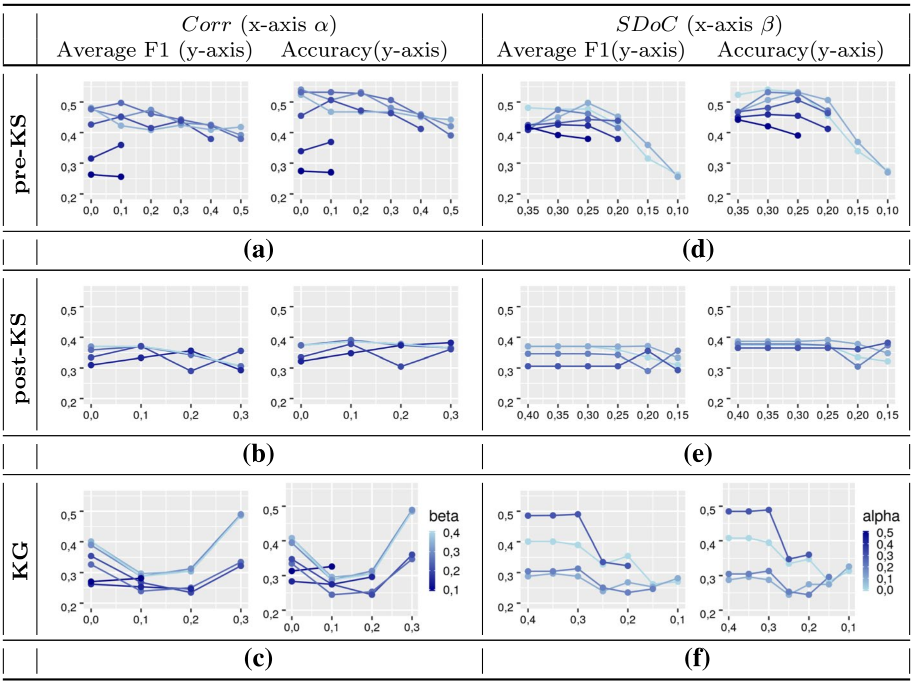


Fig. 4 Classification performance of the different feature selection strategies using the complete set of features and the best performing classifiers for each of the prediction tasks (*rf* for Pre-KS, *nb* for Post-KS, and *lr* for KG). The threshold for the feature redundancy filter is fixed at $\tau = 0.9$

(1) different features, feature selection strategies and experimental setup (i.e. we test the models on unseen topics, we tune the feature selection thresholds as hyperparameters) and (2) the *rf* classifier used by previous work may have been overfitted. This is also supported by the intermediate results produced in the repeated cross-topic validation process, where we observed that the hyperparameters selected by the inner loop do not always produce the best average F1 score for the overall result. Hence, if the parameters are selected based on their performance on the test set, there is a high risk of overfitting. Both *lr* and *nb* are less prone to this effect which was observed to a larger extent for the two more challenging prediction problems.

7.2 Feature category

The highest average F1 scores for all prediction tasks are achieved when using both Web resource and user behavior features. The results indicate that by utilizing signals from both categories, our approach is able to improve the performance of the prediction models.

For pre-KS prediction using the *rf* classifier, using both categories of features only slightly outperforms using resource features only, with the average F1 score being 10.3% higher and the overall accuracy 4% higher. For post-KS prediction, using the *lr* classifier and resource features achieves similarly high accuracy compared to using *nb* and both categories of features. This suggests that the content of the Web resources a user interacted with might carry most of the meaningful signals for post-KS prediction. For KG prediction, none of the configurations using a single category of features achieves comparable results compared to the best performing configuration. The result suggests that both the user interaction and the visited resources have strong influence on user's knowledge gain and each group of features encodes unique information about the learning process.

7.3 Feature selection strategy

To better understand the interaction of feature selection strategies for the individual KIs, we evaluate the impact of settings for *feature effectiveness* ($Corr(f_i, KI) \geq \alpha$) and *generalizability* ($SDoC(f_i, KI) < \beta$) feature selection strategies on model performance. For each of the prediction tasks, we present the results of the best classification model using different feature selection configurations.

In Fig. 4a–c, the x-axis represents α, each line corresponds to a specific β, and vice versa for Fig. 4d–f. Larger values for α lead to fewer features while larger values for β lead to higher numbers of features—i.e. x-axis from left to right the filter settings are increasingly restrictive and darker colors of lines show more restrictive filter settings as well. The threshold for *feature redundancy* is fixed at $\tau = 0.9$, the most conservative value observed in the best performing classifier configurations.

In the pre-KS prediction task, low *feature effectiveness* thresholds of $\alpha \leq 0.2$ result in the best classification performance. More restrictive filter settings result in performance decreases for this prediction task. Similarly, the best performances are achieved with non-restrictive *generalizability* filter settings, i.e. $\beta \geq 0.25$. On their own, either of these filters removes useful features and results in a decrease in performance (both in terms of F1-score and accuracy); pairing $\beta = 0.35$ (does not remove any feature) with any $\alpha > 0$, for instance, results in a drop in F1-score from 0.481 to 0.425 or below. Nevertheless, a combination of moderate settings of $\alpha = 0.1$ and $\beta = 0.25$ selects 79 features (out of 136) that result in the best overall classification performance for this task: an F1-score of 0.497 (compared to 0.481 without filters) and Accuracy of 0.532 (compared to 0.524 without filters).

In the most challenging prediction problem, post-KS prediction, we observe a slightly positive impact in prediction Accuracy when choosing a moderate *feature effectiveness* filter setting of $\alpha = 0.1$. A combination with the least restrictive *generalizability* filter setting that still removes features ($\beta = 0.25$) results in 57 features that allow the *nb* classifier to identify low and high knowledge classes better and improves its Accuracy from 0.373 to 0.391, while the average F1-score does not benefit due to a reduced recall for the medium class.

For the KG prediction task, there is overall a marked negative performance impact for introducing moderate *feature effectiveness* filter settings of $\alpha = 0.1$ and $\alpha = 0.2$, while the most restrictive setting of $\alpha = 0.3$ results in the highest performance, particularly when paired with the three least restrictive *generalizability* filter settings of $\beta \geq 0.3$. Within these settings, paired with $\alpha = 0.3$ there is no difference in the selected features, while more restrictive settings of $\beta < 0.3$ lead to a deterioration in performance. Applying the *feature effectiveness* filter in this prediction task improves F1-score from 0.401 to 0.490 and Accuracy from 0.408 to 0.489.

Overall, with regards to the *feature effectiveness* selection strategy (see Fig. 4a–c), the best classification performance for each of the prediction tasks is achieved with $\alpha > 0$, confirming previous results that this is an effective strategy for reducing the feature set in our scenario. A similar observation can be made with respect to our additional *generalizability* selection strategy (see Fig. 4d–f). Although for the filter settings with $\beta < 0.4$, the improvements are only minor and the effects of the filter vary across the different prediction tasks. In terms of the prediction tasks, the filters were least effective for the Post-KS prediction, which also showed the worst performance overall. In contrast, for KG prediction the *feature effectiveness* filter shows the largest effect, particularly for the logistic regression model.

8 Discussion

The experimental results underline the feasibility of predicting a user's knowledge state (change) without prior awareness of the specific learning intent of the user. Our approach outperforms State-of-the-Art baselines on unseen topics by considering additional features of Web resources that users interact with.

Limitations However, while providing important contributions towards improving knowledge gain of users during Web search, the experimental results indicate that the current performance of predictive models requires improvement for real-world applications. Potential reasons for this may include (1) the limited scale of training data, (2) the lack of diversity of search sessions, in particular with respect to session length, and (3) issues related to our stratification approach when building classes for knowledge state (gain). Regarding (1), especially given that the topics in our experimental dataset are spanning across several different domains and considering the large number of features (179 features in total), the training data may not be sufficient for capturing the signals carried by all meaningful features. A larger dataset with more diverse topics would certainly improve the robustness of our model. Methodology-wise, our approach pipeline, including all steps (feature extraction, feature selection, and model training), is directly applicable when new data is available. With respect to (2), the comparable short duration of all search sessions limits the signals provided for each feature. Certain features may provide more meaningful signals for longer search sessions only. Regarding (3), our stratification approach for separating knowledge state (change) classes using the *Standard Deviation Classification* approach may not be the an ideal solution for user knowledge assessment. More targeted and domain-specific knowledge assessment methods may provide more meaningful classes, where classification performance may yield better results. Despite the aforementioned limitations, our experiments provide crucial insights into the effectiveness of a wide range of features and their use as part of supervised models for predicting knowledge gain and knowledge state of users during Web search.

Feature topic dependency We conduct topic-dependency analysis on both Web resource features and user behavior features. Table 3 shows the user behavior features that are discussed in this section together with their *SDoC* corresponding to each KI. More details about the complete list of user behavior features and their correlation to the KI can be found in (Yu et al. 2018).

The 5 features with highest $SDoC(f_i, pre\text{-}KS) \geq 0.292$ are *s_duration, s_duration_per_q, b_time_per_q, b_revisited_ratio, q_uniqT_ratio*. These user behavior features suggest that effort (e.g. session length) and browsing behavior are influenced by the topic itself and the knowledge gap of the user. Further, we observe that users are more likely to revisit pages during longer sessions on broad and complex topics.

The 5 features with highest $SDoC(f_i, post\text{-}KS) \geq 0.283$ are *q_uniq_term_ ratio, q_len_first, l_relig, l_relativ, q_uniq_term_first*. Unlike the result for pre-KS, more linguistic and query term related features are found to be topic dependent with respect to post-KS. A possible reason for this finding is that the different level of specificity of the topic might influence the observed words. Hence the assumption from previous work (e.g. Arguello 2014) that higher knowledge state leads to higher coverage of keywords in the query and resources may not hold for all topics.

The 5 features with highest $SDoC(f_i, KG) \geq 0.315$ are *b_revisited_ratio, m_scroll_dist, s_duration, l_percept, m_rank_max*. Overall, the feature performance seems to vary more strongly for KG than for pre- and post-KS. Topic-dependency appears intuitive in a number of cases. For instance, in case of *l_percept*, i.e. the number of perceptual process words (such as *see* or *hear*), may be specifically popular for certain topics. *l_bio*, i.e. the number of biological process words, is an example of a highly domain-specific feature which contributes strongly to the overall performance in the pre-KS predication task. Our observations suggest that this feature contributes very strongly in life sciences-related topics, such as *Carpenter Bees*, *Altitude Sickness* or *HIV*. These findings underline that highly domain-specific linguistic features may provide very effective signals for KI prediction on unseen topics, in particular in more domain-specific models.

Correlation analysis Whereas the correlation between user behavior features and KIs has been investigated by previous work (Gadiraju et al. 2018), here we focus on the Web resource features. We notice that the correlation between the Web resource features and different KIs varies strongly. For instance, the feature *c_verb* is moderately positively correlated with pre- and post-KS and negatively correlated with KG.

The top 5 Web resource features positively correlated with pre-KS ($Corr(f_i, pre\text{-}KS) \geq 0.469$, $p < 0.05$) are *l_bio, l_health, l_focuspresent, ls_you, l_body*. The top 5 features negatively correlated with pre-KS ($Corr(f_i, pre\text{-}KS) \leq -0.287$, $p < 0.05$) are *c_adj, ls_article, ls_Quote, ls_number, l_Analytic*.

Similarly, for post-KS, the top 5 positively correlated ($Corr(f_i, pre\text{-}KS) \geq 0.302$, $p < 0.05$) resource features are *l_bio, l_health, l_focuspresent, ls_you, l_body* and the top 5 negatively correlated ($Corr(f_i, pre\text{-}KS) \leq -0.294$, $p < 0.05$) resource features are *ls_article, l_relativ, h_oth_ul, ls_Quote, c_adj*.

For KG, the top 5 positively correlated ($Corr(f_i, pre\text{-}KS) \geq 0.169$, $p < 0.05$) resource features are *l_Analytic, ls_number, c_noun, l_anger, l_money* and the top 5 negatively correlated ($Corr(f_i, pre\text{-}KS) \leq -0.349$, $p < 0.05$) resource content features are *ls_Dic, ls_conj, l_focuspresent, l_bio, l_health*. In particular with respect to the positive correlation, we observe that the amount of analytical words (*l_Analytic*) correlates positively with KG. This is intuitively explained, assuming that analytical words may have higher occurrences in suitable learning material.

We observe that seemingly topic-dependent features such as the number of biological process words (*l_bio*) correlate more strongly with the corresponding KI. This may be due to the selection of topics in the dataset we considered, which include a larger proportion of life sciences-related topics. Given that these features also proved useful in cross-topic prediction of KIs, we argue that sufficient coverage of domains may be desirable, as it may allow for capturing topic-dependent usefulness of resources and thus improve domain-specific model performances even on unseen topics.

Effect of topic and data diversity The topic HIV was used for our first study, for which we reused items in Carey et al. (1997). After a preliminary analysis of the first study result, we gathered more diverse topics to cross-validate our findings. Given that no equivalent dataset exists, we decided to use (Carey et al. 1997) as the guideline for building knowledge tests for the 10 topics that were randomly selected from TREC dataset. Topic familiarity could be one of the reasons that the HIV topic results in the highest number of sessions, as less users drop out during the session when they are more familiar with the topic.

Table 5 Best performing results of different approaches according to average F1 score. (10 topics)

KI	Approach	Feature cat.	clf	Low			Moderate			High			Average			Accu
				P	R	F1	P	R	F1	P	R	F1	P	R	F1	
Pre	New	WR&UB	rf	0.462	0.269	0.340	0.429	0.656	0.519	0.538	0.493	0.515	0.476	0.473	**0.458**	**0.470**
	baseline	–	rf	0.358	0.358	0.358	0.344	0.344	0.344	0.408	0.408	0.408	0.370	0.370	0.370	0.371
Post	New	WR&UB	nb	0.384	0.475	0.424	0.424	0.368	0.394	0.429	0.403	0.415	0.412	0.415	**0.411**	**0.411**
	baseline	–	rf	0.360	0.305	0.330	0.398	0.461	0.427	0.313	0.299	0.305	0.357	0.355	0.354	0.361
KG	New	WR&UB	lr	0.449	0.639	0.527	0.357	0.167	0.227	0.393	0.373	0.383	0.400	0.393	**0.379**	**0.421**
	baseline	–	rf	0.458	0.590	0.516	0.250	0.200	0.222	0.404	0.322	0.358	0.371	0.371	0.366	0.396

The F1-score and the accuracy for the best performing model of each task are marked in bold

Since the goal of this work is to build generalizable models, we decided not to leave out the data collected for the HIV topic to increase the size and topic diversity of our ground truth dataset and avoid overfitting in model training as much as possible.

We have run the same experiment using 202 sessions data from 10 topics (i.e. excluding HIV topic), the results (Table 5) show that the overall performance of models is worse than using data from all 11 topics. The reason could be that the smaller size of training data has led to overfitting. Meanwhile, as the number of features used by the baseline approach is much less than the proposed approach (i.e. only including user behavior features), models are less affected by reducing ground truth data. Compared to using data from 11 topics, on average across 3 tasks, performance of the proposed model dropped by 0.072 in macro average F1 and 0.065 in accuracy, while the performance of the baseline approach dropped by 0.027 in macro average F1 and 0.028 in accuracy. Overall, when less training data is provided, the advantage of the proposed model is less significant.

Future work As discussed in Sect. 2, some of the previous works analysed the relation between user knowledge state (change) and their precise resource consumption behavior based on behavioural data such as eye tracking logs. In this work, we restricted our feature sets for two reasons: (1) our goal is to build models that can be applied in real-world search environments, hence we computed features based on user behaviors and resources that are easily accessible by common search systems. (2) To ensure reasonably large amounts of data/sessions for training/testing, we decided to use a crowdsourcing platform instead of a controlled lab study environment, and we did not put any requirement on the hardware and software used by the participants. Under these experimental conditions, we are not able to capture the layout of the page on the user's end. It is noteworthy, however, that some of the features we present in this paper are reflecting users' resource consumption behavior to a certain extent. This includes, for instance, the time duration a user spent on browsing the resources. As part of ongoing work, we are exploring the relationship between users' consumption of learning resources and their knowledge gain using data collected from lab studies, where the details of the search environment (e.g. the device and browser the participants used) are controlled. We capture accurate user browsing behaviors by two means: (1) combining screen recordings with eye-tracking data to capture the content seen by participants, and (2) hooking into the browser's rendering engine to capture the page layout, and combine it with eye-tracking data to capture the content seen by participants. However, whereas the limited size of the lab study data with respect to the number of search sessions and number of topics provides interesting analytical insights, it is not well suited to build robust and generalizable models that are comparable to the models presented in this paper.

Deep learning models, such as Recurrent Neural Networks (RNNs), have been used for user behavior modeling in recent years. For instance, Donkers et al. (2017) adapted RNNs for recommender systems. Smirnova and Vasile (2017) proposed a new class of Contextual Recurrent Neural Networks for recommendation which takes into account both the sequence of items that the user has interacted with and other contextual information such as the time gaps between interactions. Xu et al. (2019) proposed a novel Recurrent Convolutional Neural Network model for recommendation based on sequential user behaviors. Wang et al. (2020) proposed a graph neural networks based model that incorporates items, sessions, locations and time of user interactions for user representation. Given the nature of the learning process, sequential models are suitable for this task and are one direction to go for optimizing the prediction models. If applying sequential models (e.g. LSTM based RNN) for knowledge prediction, the sequence of user interactions and the content consumed by users could be used as input to sequential models. However, the nature of the knowledge prediction task also brings limitations to the application of such models, as it requires knowledge assessment data as ground truth. To acquire such ground truth, researchers usually need to rely on user studies. The diversity and sparsity of data collected from user studies are not well-suited to deep learning methods. We will continue to explore the possibility of using a sequential model for knowledge prediction, and hope that the positive result in this paper, which demonstrates that it is possible to predict the knowledge state and gain of users in Web search, can encourage researchers to advance the exploration on this task and/or to accumulate larger amounts of suitable datasets for training/testing of models.

9 Conclusions

We propose to improve the performance and generalizability of knowledge prediction models of informational Web search sessions. We extracted a feature set, which extends prior work with Web resource features, and combine them with user behavior features introduced by Yu et al. (2018). We also conducted a preliminary analysis with respect to the correlation and dependency of features to the KIs. For the knowledge modeling, we applied and evaluated several feature selection strategies that focus on different aspects of feature effectiveness, showing that reducing the feature set and accounting for topic-dependency of features improves generalization performance. For each of the studied knowledge indicators, our approach outperforms the State-of-the-Art baseline in the cross-topic experimental evaluation.

There are limitations in the experimental dataset, most notably caused by the limited availability of search session data together with corresponding knowledge indicators. A range of the tested features maybe effective on the task, but we require longer and more varied sessions to discover their relation to the KIs. As part of future work, we aim to extend the investigation of this paper and make use of more varied informational search sessions, for instance, including additional domains and covering longer sessions.

Funding Open Access funding enabled and organized by Projekt DEAL..

References

Anderson, L. W., Krathwohl, D. R., Airasian, P., Cruikshank, K., Mayer, R., Pintrich, P., et al. (2001). *A taxonomy for learning, teaching and assessing: A revision of bloom's taxonomy*. New York: Longman Publishing.

Arguello, J. (2014). Predicting search task difficulty. *ECIR*, *14*, 88–99.

Bhattacharya, N., & Gwizdka, J. (2018). Relating eye-tracking measures with changes in knowledge on search tasks. In *Proceedings of the 2018 ACM symposium on eye tracking research & applications* (pp. 1–5).

Bhattacharya, N., & Gwizdka, J. (2019). Measuring learning during search: Differences in interactions, eye-gaze, and semantic similarity to expert knowledge. In *Proceedings of the 2019 conference on human information interaction and retrieval* (pp. 63–71). ACM.

Broder, A. (2002). A taxonomy of web search. In *ACM Sigir forum* (vol. 36, pp. 3–10). ACM.

Carey, M. P., Morrison-Beedy, D., & Johnson, B. T. (1997). The HIV-knowledge questionnaire: Development and evaluation of a reliable, valid, and practical self-administered questionnaire. *AIDS and Behavior*, *1*(1), 61–74.

Cole, M. J., Gwizdka, J., Liu, C., Belkin, N. J., & Zhang, X. (2013). Inferring user knowledge level from eye movement patterns. *Information Processing & Management*, *49*(5), 1075–1091.

Collins-Thompson, K., Rieh, S. Y., Haynes, C. C., & Syed, R. (2016). Assessing learning outcomes in web search: A comparison of tasks and query strategies. In *Proceedings of the 2016 ACM on conference on human information interaction and retrieval* (pp. 163–172). ACM.

DeStefano, D., & LeFevre, J. A. (2007). Cognitive load in hypertext reading: A review. *Computers in Human Behavior*, *23*(3), 1616–1641.

Donkers, T., Loepp, B., & Ziegler, J. (2017). Sequential user-based recurrent neural network recommendations. In *Proceedings of the eleventh ACM conference on recommender systems* (pp. 152–160).

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). ACM.

Gadiraju, U., Yang, J., & Bozzon, A. (2017). Clarity is a worthwhile quality—on the role of task clarity in microtask crowdsourcing. In: *Proceedings of the 28th ACM conference on hypertext and social media* (pp. 5–14). ACM.

Gadiraju, U., Yu, R., Dietze, S., & Holtz, P. (2018). Analyzing knowledge gain of users in informational search sessions on the web. In *2018 ACM on conference on human information interaction and retrieval (CHIIR)*. ACM.

Gwizdka, J., & Chen, X. (2016). Towards observable indicators of learning on search. In *SAL@ SIGIR*

Gwizdka, J., & Spence, I. (2006). What can searching behavior tell us about the difficulty of information tasks? A study of web navigation. *Proceedings of the Association for Information Science and Technology*, *43*(1), 1–22.

Hagen, M., Potthast, M., Völske, M., Gomoll, J., & Stein, B. (2016). How writers search: Analyzing the search and writing logs of non-fictional essays. In *Proceedings of the 2016 ACM on conference on human information interaction and retrieval* (pp. 193–202). ACM.

Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human language technologies 2007: the conference of the North American chapter of the association for computational linguistics; proceedings of the main conference* (pp. 460–467).

Horne, B. D., & Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. arXiv preprint arXiv:170309398.

Jansen, B. J., Booth, D., & Smith, B. (2009). Using the taxonomy of cognitive learning to model online searching. *Information Processing & Management*, *45*(6), 643–663.

Kalyani, R., & Gadiraju, U. (2019). Understanding user search behavior across varying cognitive levels. In *Proceedings of the 30th ACM conference on hypertext and social media* (pp. 123–132).

Kincaid, JP., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990.

Li, Y., & Belkin, N. J. (2008). A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management*, *44*(6), 1822–1837.

Liu, C., & Song, X. (2018). How do information source selection strategies influence users' learning outcomes'. In *Proceedings of the 2018 conference on human information interaction & retrieval* (pp. 257–260).

Liu, H., Liu, C., & Belkin, N. J. (2019). Investigation of users' knowledge change process in learning-related search tasks. *Proceedings of the Association for Information Science and Technology*, *56*(1), 166–175.

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

Roy, N., Moraes, F., & Hauff, C. (2020). Exploring users' learning gains within search sessions. In *Proceedings of the 2020 conference on human information interaction and retrieval* (pp. 432–436).

Smirnova, E., & Vasile, F. (2017). Contextual sequence modeling for recommendation with recurrent neural networks. In *Proceedings of the 2nd workshop on deep learning for recommender systems* (pp. 2–9).

Syed, R., & Collins-Thompson, K. (2017). Retrieval algorithms optimized for human learning. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval* (pp. 555–564). ACM

Syed, R., & Collins-Thompson, K. (2018). Exploring document retrieval features associated with improved short-and long-term vocabulary learning outcomes. In *Proceedings of the 2018 conference on human information interaction & retrieval* (pp. 191–200). ACM.

Vakkari, P. (2016). Searching as learning: A systematization based on literature. *Journal of Information Science*, *42*(1), 7–18.

Wang, D., Jiang, M., Syed, M., Conway, O., Juneja, V., Subramanian, S., & Chawla, N. V. (2020). Calendar graph neural networks for modeling time structures in spatiotemporal user behaviors. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2581–2589).

White, R. W., Dumais, S. T., Teevan, J. (2009). Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the second ACM international conference on web search and data mining* (pp. 132–141). ACM.

Wu, W. C., Kelly, D., Edwards, A., & Arguello, J. (2012). Grannies, tanning beds, tattoos and NASCAR: Evaluation of search tasks with varying levels of cognitive complexity. In *Proceedings of the 4th information interaction in context symposium* (pp. 254–257). ACM.

Xu, C., Zhao, P., Liu, Y., Xu, J., Sheng, V. S. S. S., Cui, Z., Zhou, X., & Xiong, H. (2019). Recurrent convolutional neural network for sequential recommendation. In *The world wide web conference* (pp. 3398–3404).

Yu, R., Gadiraju, U., Holtz, P., Rokicki, M., Kemkes, P., & Dietze, S. (2018). Predicting user knowledge gain in informational search sessions. In *Proceedings of the 41st international ACM SIGIR conference on research and development in information retrieval*. ACM.

Zhang, X., Cole, M., & Belkin, N. (2011). Predicting users' domain knowledge from search behaviors. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in information retrieval* (pp. 1225–1226). ACM.

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