Exploring Users' Learning Gains within Search Sessions

Nirmal Roy
TU Delft
Delft, Netherlands
n.roy@tudelft.nl

Felipe Moraes
TU Delft
Delft, Netherlands
f.moraes@tudelft.nl

Claudia Hauff
TU Delft
Delft, Netherlands
c.hauff@tudelft.nl

ABSTRACT

The area of search as learning is concerned with the optimization of search systems (that is, retrieval functions, user interface elements, etc.) for *human learning*—this is in contrast to the currently dominant paradigm of optimizing the search experience by optimizing for relevance. While prior work typically considers learning as something that happens at some point during the search session, we are interested in *when* during the search session learning occurs. In order to answer this question, we here present the results of a user study ($N = 64$) in which searchers were tasked with learning about a topic by searching the web for 20 minutes; they were prompted at *regular intervals* during the search session on their knowledge about the topic. We find that for study participants with little to no prior knowledge the learning gains are sublinear, while participants with some prior knowledge have the largest knowledge gains towards the end of the search session.

ACM Reference Format:
Nirmal Roy, Felipe Moraes, and Claudia Hauff. 2020. Exploring Users' Learning Gains within Search Sessions. In *2020 Conference on Human Information Interaction and Retrieval (CHIIR '20), March 14–18, 2020, Vancouver, BC, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3343413.3378012

1 INTRODUCTION

The seminal paper of Marchionini [19] defines "learning searches" as search activities whose ultimate goal is human learning. Those searches are typically iterative and (in contrast to some other types of searches) require the user to scan, read and process a large number of documents. With the—by now—omnipresent use of the web for learning purposes [2, 21, 24] the need for search systems that are designed for human learning is great. Despite several recent initiatives in the search as learning area [6, 13], we are still a long way from solving this issue, as evident in the many remaining research challenges [23, 28].

While different aspects of search as learning have been considered, the setup to *explicitly measure learning*—which requires lab studies and is in contrast to large-scale query log analyses [10] that have to make assumptions about the amount of learning taking place—typically looks as follows: a pre-test is administered to study participants to determine their knowledge levels, and after the search episode a post-test is similarly administered such that the difference in knowledge can be calculated.

While this approach to measure learning tells us that during the search session learning occurred, we still do not know at what point of the search session this was. We lack insights into *when* learning occurs and whether this differs among different types of users (e.g. those without prior knowledge and those with prior knowledge).

In this work we explore the question of *how the knowledge gain of users develops over the time of a search session* with a lab study conducted with 64 study participants. We administer regular knowledge tests during the search session and find that for study participants with little to no prior knowledge the learning gains are sub-linear, while participants with some prior knowledge have the largest knowledge gains towards the end of the search session. In terms of observable search behaviours and their use as proxies for learning, we find the number of submitted queries and the document dwell time to be the most predictive ones for learning—in line with prior works [7, 10, 31].

2 RELATED WORK

Past work in the area of search as learning has focused on the optimization of retrieval functions for human learning [26, 27], observational studies of how users employ search engines for learning-oriented searches [4, 10, 16, 18], different types of users (e.g., experts vs. novices) and their search behaviours [1, 10], different types of learning setups (e.g., search only vs. search plus designed learning materials [20]), learning in specific domains (e.g., health [5]) and the development of metrics and techniques to quantify open-ended learning [17, 30].

In terms of scalable behavioural metrics as proxies for measuring knowledge gain across a search session, the document dwell time has been found to be a good indicator for learning [7, 10] as well as the number of SERP clicks [7] and the number of unique domains present among the top-ranked search results [10]. Recently, Yu et al. [31] conducted a large-sale study of about 70 search-based features as predictors of learning and found them to be only weakly correlated with knowledge gain; however, since the search sessions were very short (5 minutes in total) it remains to be seen whether those findings hold in more common longer search sessions.

What is common in the prior studies we described above, e.g., [1, 12, 16, 31], is the use of a single pre-test and post-test setup to measure learning (and in turn to quantify the learning gain). This is in contrast to our study, where we are interested not just in the final learning outcome at the end of the search session, but also *when* the learning during the search session takes place. Lastly, we point to the recent work by Liu et al. [17] which has a similar goal to

This research has been supported by NWO projects LACrOSSE (612.001.605), SearchX (639.022.722) and NWO Aspasia (015.013.027).

CHIIR '20, March 14–18, 2020, Vancouver, BC, Canada

ACM ISBN 978-1-4503-6892-6/20/03…$15.00
https://doi.org/10.1145/3343413.3378012

our work: here, mind-maps which searchers created during search sessions were analyzed and manually evaluated for their evolution over time to make statements about the change of knowledge.

3 STUDY

Overview. Our user study is inspired by the setup of Moraes et al. [20][1]: we make use of the same open-source search system SearchX [22] (which provides quality control features for crowdsourcing experiments and fine-grained search logs) and adapt it slightly to enable the use of *intermediate tests* during the search session. The flow of our study is shown in Figure 1: at the start of the study we conduct a pre-test to determine our participants knowledge levels prior to the search session. Participants have to answer ten vocabulary knowledge questions on two randomly picked topics (out of ten available); as a sanity check we include questions on a third topic (*sports*) that we expect reasonable participants to demonstrate high knowledge levels on; participants that do not are rejected from our study. The topic they know the least about is then chosen as the topic to learn more about during the search session. We require participants to search the web—facilitated by the Bing API, one of SearchX's backends—for at least twenty minutes (a timer on the interface helps participants to keep track of time) as this has been shown [10, 14, 20] to be a reasonable time for people to gain some knowledge. During the search session, participants can view, read and bookmark documents. At regular intervals—every five minutes—the participants are "interrupted" in their search session and asked the same vocabulary knowledge questions as in the pre-test. Participants can score their vocabulary knowledge on a scale from 1 to 4 (see below for details); during the intermediate tests we omit vocabulary items that have been self-assessed as "4" in either the pre-test or any of the preceding intermediate tests. This means that *at most* 10 questions have to be answered in each intermediate test. As we are interested in the participants' knowledge gain over time, we are constrained by the fact that we need to ask the participants about the same set of vocabulary terms repeatedly. In the post-test, we ask our participants one last time about their knowledge on the ten vocabulary items. In addition, we ask them to write a summary (100 words minimum) about the topic assigned to them. Overall, thus every study participant completes five knowledge tests. We note here that the knowledge tests require understanding, but no application or synthesis (i.e. higher-level cognitive processes of learning [15]) of the materials. This is in line with prior works in this area due to the limited amount of search time study participants have.

Topics. We employ the same ten *learning tasks* as Moraes et al. [20], which are based on introductory material of Bachelor-level Massive Open Online Courses on a variety of subject areas and ask each participant to use our search engine to learn more about a particular topic such as *qubit* or *radioactive decay* (cf. Table 1 for the full list of topics). The task descriptions[2] encourage participants to explore multiple aspects of the topic. For each topic, a list of 10 vocabulary items are available that were determined in [20] to be very relevant to the topic[3].

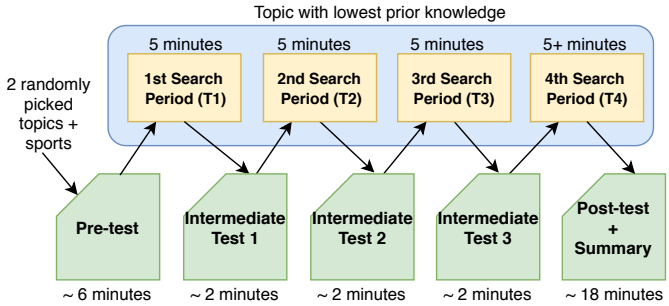

Figure 1: Overview of our user study setup.

Vocabulary Knowledge Scale. In order to evaluate the knowledge gain of our participants we ask them to self-assess their knowledge on those vocabulary items according to the vocabulary knowledge scale (VKS) [29] across four levels:

(1) *I don't remember having seen this term/phase before.*
(2) *I have seen this term/phrase before, but I don't think I know what it means.*
(3) *I have seen this term/phrase before and I think it means ...*
(4) *I know this term/phrase. It means ...*

A self-assessment of (3) or (4) requires participants to write down a definition of the term in their own words. In order to compute the learning gain, we *rescore* those self-assesments: we assign a score of 0 to knowledge levels (1) and (2). Since level (3) indicates uncertainty regarding the meaning of the term, we assign it a score of 1. Choosing level (4) indicates the participant is confident in the definition and we assign it a score of 2. This scoring scheme is equivalent to the fine-grained setup of [20].

Table 1: Overview of the topic distribution, average number of queries per topic (AQ) and average number of bookmarks per topic (AB) among the N=64 study participants (SP).

	Anesthesia	Depression	Glycolysis	Industrial biotechnology	Qubit	Radioactive decay	Religions	Sedimentary rocks	Urban water cycle	Water quality chemistry
SP	7	3	10	3	5	7	3	6	6	14
AQ	12	10	5.9	6.67	10.6	4.86	8.67	7.83	6.33	7.07
AB	6	5.33	3.8	6.0	7.8	5.86	4	3.5	6.67	7.14

Measuring the Learning Gain. Similar to [8, 20, 25–27], we use *realized potential learning* (RPL) as our learning gain metric. RPL normalizes the absolute learning gains (ALG) measured in terms of the number of new vocabulary terms learnt by the maximum possible learning potential (in our case this is 2.0 as we rescored the

[1] More specifically, our study setup follows closely the *search only* condition of Moraes et al. [20].

[2] As an example, the task description for *qubit* is: *Imagine you are taking an introductory Computer Science course this term. For your term paper, you have decided to write about a topic where computer science meets physics: quantum information. You also would like to learn about the difference between a classical bit and a quantum bit (a so-called qubit).*"

[3] For example, for the *qubit* topic we have terms such as *ket* and *quantum cryptography* in our vocabulary list

self-assessments) for each unknown vocabulary term. We compute RPL each test stage *with respect to the pre-test vocabulary assessment*:

$$ALG = \frac{1}{n}\sum_{i=1}^{n} max(0, vks^{X}(v_i) - vks^{pre}(v_i)) \quad (1)$$

$$MLG = \frac{1}{n}\sum_{i=1}^{n} 2 - vks^{pre}(v_i) \quad (2)$$

$$RPL = \frac{ALG}{MLG} \quad (3)$$

where $vks^{X}(v_i)$ is the rescored self-assessment of vocabulary item v_i (i.e. either 0, 1 or 2). X is either one of the intermediate tests or the post-test and n is the number of evaluated vocabulary items.

Search Behaviour Metrics. Based on prior research [7, 10, 31] we extract six different search features from our search logs for each search period (i.e., the period between two tests): (i) the **number of queries** a participant formulates; (ii) the **number of search result page clicks** a participant makes; (iii) the **number of bookmarks** a participant makes; (iv) the **number of documents a participant views**; (v) the **average document dwell time**, i.e., the average time in seconds a participant spends reading the viewed documents; and (vi) the average number of **unique domains** on each SERP.

Study Participants. We conducted our study on the Prolific Academic platform[4] across three days. In order to ensure responses of high quality, we required our participants to have at least 15 previous submissions, an approval rate of 90+% and be native English speakers. The study took about an hour to complete and we paid £6. Seventy three participants completed our study; we rejected nine participants because they did not comply with our rules, most importantly that at most three browser tab changes are allowed and participants need to be actively using our system. Not all topics had the same number of participants, while *water quality chemistry* and *glycolysis* were assigned to 10+ participants each, *religions* and *depression* were assigned to three participants each as seen in Table 1. The table also contains the average number of queries per topic (between between 5 and 12) and the average number of bookmarks (between 3 and 8), showing that our study participants actively engaged in the search session.

Self-assessment Quality. In order to determine the quality of the vocabulary knowledge self-assessments, we sampled 100 definitions written by our participants (50 for knowledge level (3) and (4) respectively). Two annotators labelled them as *correct*, *partially correct*[5] and *incorrect*[6]. We find that 74% of the definitions self-assessed as level (4) were correct and 16% partially correct. For self-assessed knowledge level (3), 68% of the definitions are correct and 24% partially correct. Based on these numbers, we consider the self-assessment to be largely reliable and we thus report RPL based on the self-assessed vocabulary knowledge levels.

[4]https://www.prolific.co/

[5]Partially correct: *citric acid cycle* (Glycolsis topic) describes as 'chemical reaction'

[6]Incorrect: *conjugate transpose (Qubit topic) described as 'The smallest possible unit of quantum information'.*

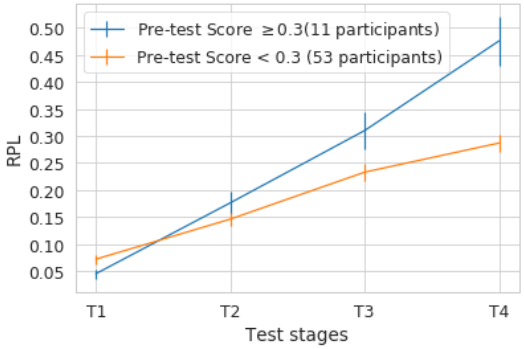

Figure 2: Overview of the knowledge gain (measured in RPL) at different stages of the search. The plot shows the mean RPL at each stage together with the standard error for the two types of participants. RPL is always measured in relation to the pre-test. T1 measures the RPL between the first intermediate test and the pre-test, T4 measures the RPL between the post-test and the fourth search period.

4 RESULTS

Learning Gains and Search Behaviour over Time. To answer our research question, we investigate how our participants' learning gains change over time. To do so, we compute RPL for each participant in the four test stages; note that RPL is always computed *with respect to the pre-test*. We find that, on average across all participants, the increase in RPL slows down as we move along the test stages: 55.5%, 38.5%, and 23.0%, of increase in test stage 1 to 2, 2 to 3, and 3 to 4, respectively—the increase is sub-linear. However, the picture becomes more differentiated as we distinguish two types of participants: those without prior topical knowledge (defined as participants who answered less than three questions correctly in the pre-test, $N = 53$) and those with some prior topical knowledge (defined as those with at least three correctly answered questions in the pre-test, $N = 11$). Plotting the development of RPL over the test stages, as done in Figure 2 shows that participants with some prior knowledge have higher learning gains towards the end of the search session (and lower relative gains at the start), in contrast to participants with no/little prior knowledge whose learning gains become less as the search session progresses.

In Figure 3 we plot the participants' search behaviours along the six dimensions listed in §3; we compute those behaviours individually for each search period preceding a test stage. We find participants with higher pre-test scores issue more queries at the beginning of the search session, spend more time reading documents and look at documents from more different domains than our participants with little prior knowledge, in line with the findings reported in [10]. In contrast, participants with little prior knowledge spend more time on the SERP. Among the search behaviours, we only observe a clear downward trend for the average document dwell time: as we move through the tests, participants spend less and less time on individual documents. Based on these findings we argue that we should not only consider the knowledge before and after a search period, but also incorporate the fact in our models that the knowledge state changes are not constant over time.

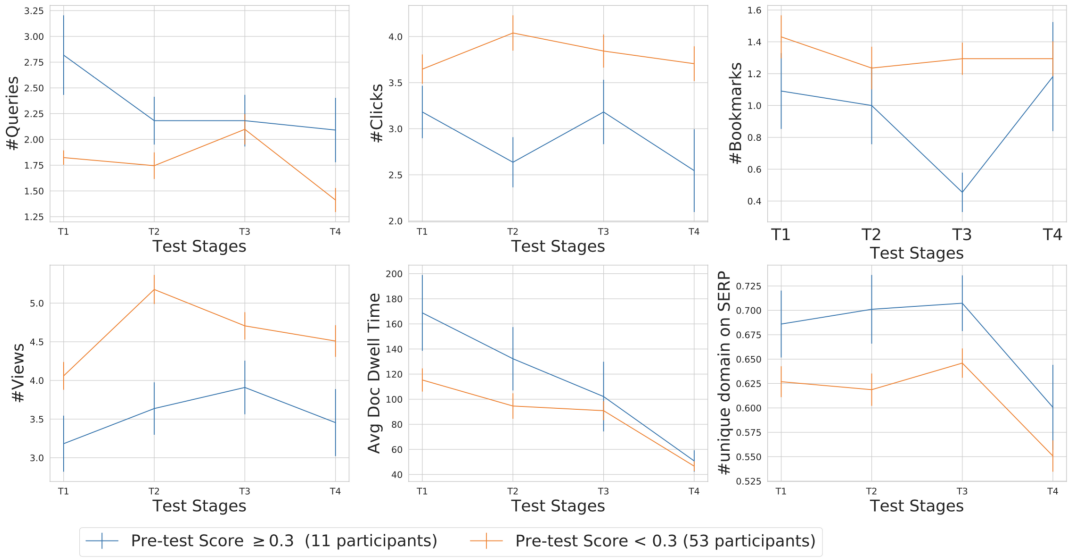

Figure 3: Overview of different search behaviours observed during the search period preceding each test stage.

Learning Gain Proxies. Lastly, as a validation of our study with respect to previous works, we use our six behavioural metrics as input to a predictive model that outputs the predicted knowledge gain using cross-validation. Due to our four test stages, we have 256 data points in total. Here, we consider the test stage itself to be a feature as well since it indicates how far a participant has reached in the search journey and consequently topic expertise. We train a random forest regression model [3][7] to predict knowledge gain in terms of RPL. This type of model has been shown to be highly effective for a wide range of tasks [11]. We use nested 5-fold cross-validation on the 256 data points for finding the optimal hyperparameters and evaluate the performance using mean squared error. The optimal hyperparameters are 500 trees with a maximum depth of 5. The prediction error is 0.013, which is reasonably small compared with RPL that ranges from 0 to 1 (mean of 0.196 across all 256 data points). Using the same hyperparameters, we then retrain the model using the entire dataset to estimate the contribution of each feature to the prediction.

The results are shown in Figure 4, where feature importance is defined as the average total decrease in node impurity as contributed by each feature towards the classification result with each feature value ranging from 0 to 1 and summing up to 1. The results corroborate our findings regarding study participants—the number of queries and document dwell time are the two most important features of the model.

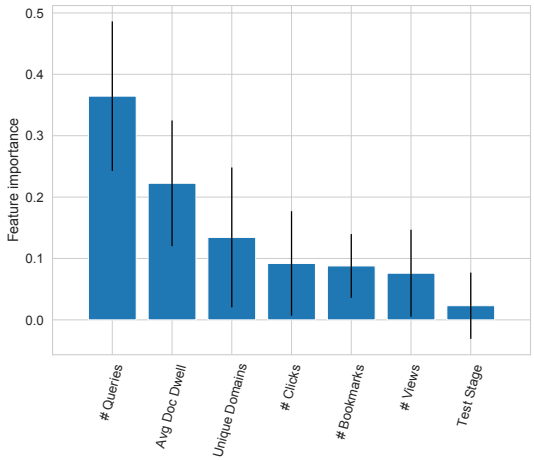

Figure 4: Overview of the feature importance of random forest regressor model. Higher values mean higher importance for a feature computed using node impurity.

5 CONCLUSIONS

In this work we investigated *when* learning occurs during search sessions with a learning-oriented information need based on a user study with 64 participants that regularly received prompts during a 20 minute search session in order to test their knowledge. We found that participants with little to no prior knowledge on a topic experienced sub-linear learning gains over the course of their search session, while participants with some prior knowledge experienced the largest learning gain towards the end of their search session.

In future work we will investigate whether a similar result can be observed for cognitively more demanding tasks (i.e. analysis and synthesis instead of recalling and understanding) and to what extent we can promote earlier learning by adaptively changing the search system interface (inspired by Diriye et al. [9]), depending on the amount of learning that is taking place.

[7] We used the implementation provided by scikit-learn v0.21.3.

REFERENCES

[1] Nilavra Bhattacharya and Jacek Gwizdka. 2018. Relating eye-tracking measures with changes in knowledge on search tasks. In *ETRA 2018*. ACM, 62.

[2] J Patrick Biddix, Chung Joo Chung, and Han Woo Park. 2011. Convenience or credibility? A study of college student online research behaviors. *The Internet and Higher Education* 14, 3 (2011), 175–182.

[3] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[4] Marc Bron, Jasmijn Van Gorp, Frank Nack, Lotte Belice Baltussen, and Maarten de Rijke. 2013. Aggregated search interface preferences in multi-session search tasks. In *SIGIR 2013*. ACM, 123–132.

[5] Yu Chi. 2019. Examining and Supporting Laypeople's Learning in Online Health Information Seeking. In *CHIIR 2019*. ACM, 425–428.

[6] Kevyn Collins-Thompson, Preben Hansen, and Claudia Hauff. 2017. Search as Learning (Dagstuhl Seminar 17092). *Dagstuhl Reports* 7, 2 (2017), 135–162.

[7] Kevyn Collins-Thompson, Soo Young Rieh, Carl C Haynes, and Rohail Syed. 2016. Assessing learning outcomes in web search: A comparison of tasks and query strategies. In *CHIIR 2016*. ACM, 163–172.

[8] Henri G Colt, Mohsen Davoudi, Septimiu Murgu, and Nazanin Zamanian Rohani. 2011. Measuring learning gain during a one-day introductory bronchoscopy course. *Surgical endoscopy* 25, 1 (2011), 207–216.

[9] Abdigani Diriye, Ann Blandford, Anastasios Tombros, and Pertti Vakkari. 2013. The role of search interface features during information seeking. In *TPDL 2013*. Springer, 235–240.

[10] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. 2014. Lessons from the journey: a query log analysis of within-session learning. In *WSDM 2014*. ACM, 223–232.

[11] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research* 15, 1 (2014), 3133–3181.

[12] Ujwal Gadiraju, Ran Yu, Stefan Dietze, and Peter Holtz. 2018. Analyzing knowledge gain of users in informational search sessions on the web. In *CHIIR 2018*. ACM, 2–11.

[13] Jacek Gwizdka, Preben Hansen, Claudia Hauff, Jiyin He, and Noriko Kando. 2016. Search as learning workshop. In *SIGIR 2016*. ACM, 1249–1250.

[14] Rishita Kalyani and Ujwal Gadiraju. 2019. Understanding User Search Behavior Across Varying Cognitive Levels. In *HT 2019*. ACM, 123–132.

[15] David R Krathwohl and Lorin W Anderson. 2009. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

[16] Chang Liu and Xiaoxuan Song. 2018. How do Information Source Selection Strategies Influence Users' Learning Outcomes'. In *CHIIR 2018*. ACM, 257–260.

[17] Hanrui Liu, Chang Liu, and Nicholas J Belkin. 2019. Investigation of users' knowledge change process in learning-related search tasks. *ASIS&T 2019* 56, 1 (2019), 166–175.

[18] Jingjing Liu, Nicholas J Belkin, Xiangmin Zhang, and Xiaojun Yuan. 2013. Examining users' knowledge change in the task completion process. *Information Processing & Management* 49, 5 (2013), 1058–1074.

[19] Gary Marchionini. 2006. Marchionini, G.: Exploratory search: from finding to understanding. *Commun. ACM* 49 (04 2006), 41–46.

[20] Felipe Moraes, Sindunuraga Rikarno Putra, and Claudia Hauff. 2018. Contrasting Search as a Learning Activity with Instructor-designed Learning. In *CIKM 2018*. ACM, 167–176.

[21] David Nicholas, Ian Rowlands, Dj Clark, and Peter Williams. 2011. Google Generation II: Web behaviour experiments with the BBC. *Aslib Proceedings* 63 (01 2011), 28–45.

[22] Sindunuraga Rikarno Putra, Kilian Grashoff, Felipe Moraes, and Claudia Hauff. 2018. On the Development of a Collaborative Search System.. In *DESIRES*. 76–82.

[23] Soo Young Rieh, Kevyn Collins-Thompson, Preben Hansen, and Hye-Jung Lee. 2016. Towards searching as a learning process: A review of current perspectives and future directions. *Journal of Information Science* 42, 1 (2016), 19–34.

[24] Ian Rowlands, David Nicholas, Peter Williams, Paul Huntington, Maggie Fieldhouse, Barrie Gunter, Richard Withey, Hamid R. Jamali, Tom Dobrowolski, and Carol Tenopir. 2008. The Google generation: The information behaviour of the researcher of the future. *Aslib Proceedings* 60 (06 2008), 290–310.

[25] John L Shefelbine. 1990. Student factors related to variability in learning word meanings from context. *Journal of Reading Behavior* 22, 1 (1990), 71–97.

[26] Rohail Syed and Kevyn Collins-Thompson. 2017. Retrieval algorithms optimized for human learning. In *SIGIR 2017*. ACM, 555–564.

[27] Rohail Syed and Kevyn Collins-Thompson. 2018. Exploring Document Retrieval Features Associated with Improved Short-and Long-term Vocabulary Learning Outcomes. In *CHIIR 2018*. ACM, 191–200.

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

[29] Marjorie Wesche and T Sima Paribakht. 1996. Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review* 53, 1 (1996), 13–40.

[30] Mathew J Wilson and Max L Wilson. 2013. A comparison of techniques for measuring sensemaking and learning within participant-generated summaries. *Journal of the American Society for Information Science and Technology* 64, 2 (2013), 291–306.

[31] Ran Yu, Ujwal Gadiraju, Peter Holtz, Markus Rokicki, Philipp Kemkes, and Stefan Dietze. 2018. Predicting user knowledge gain in informational search sessions. In *SIGIR 2018*. ACM, 75–84.