Learning to Rank for Knowledge Gain

Markus Rokicki[1], Ran Yu[2] and Daniel Hienert[3]

[1]L3S Research Center, Leibniz University Hannover, Hannover, Germany
[2]Data Science & Intelligent Systems Group, University of Bonn, Germany
[3]GESIS - Leibniz Institute, for the Social Sciences, Cologne, Germany

Abstract

Web search has often been used as a starting point to learn. Search as Learning (SAL) research aims at supporting learning activities through techniques such as user interface optimization, retrieval, and ranking. In this work, we investigate the possibility of re-ranking search engine results towards learning to improve the overall knowledge gain of the learner. We make two contributions: (1) proposing a framework for re-ranking search results by attributing the overall knowledge gain to viewed documents in the session. (2) Applying this framework to a SAL evaluation dataset. We show that the ranking can be significantly improved with respect to knowledge gain by using ranking and content features.

1. Introduction

The research field 'Search as Learning' (SAL) is focused on supporting users with learning tasks on the Web. For that goal, it is essential to understand how users learn and behave in learning tasks [1] and what might be factors for predicting and improving knowledge gain [2]. In web-based learning, one of the main factors is the ranking of search engine results pages (SERPs) [3]. The ranking determines which resources are suggested to the user for a specific learning task. So far, the understanding of how to improve users' knowledge gain through re-ranking strategies is still very limited. In this research, we use an existing data set of learning-focused web search sessions and re-rank them towards knowledge gain using different machine learning models and feature groups. We show that the rankings can be improved significantly towards knowledge gain by considering ranking and content features. In particular, in-session signals can be beneficial for improving rankings toward better learning outcomes. Our contributions in this paper are two-fold:

1. We propose a general framework for re-ranking search results regarding knowledge gain attributed to individual documents in the search session to optimize rankings for the learning outcome.
2. Applying the framework to an existing search as learning dataset. Results show that the ranking can be improved for a specific learning task. Therefore, we encourage other researchers to apply our framework to different learning tasks and topics to identify factors that can be used to further improve rankings toward knowledge gain.

Joint Proceedings of the 10th International Workshop on News Recommendation and Analytics (INRA'22) and the Third International Workshop on Investigating Learning During Web Search (IWILDS'22) co-located with the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'22), July 15, 2022, Madrid, Spain
rokicki@l3s.de (M. Rokicki); ran.yu@uni-bonn.de (R. Yu); daniel.hienert@gesis.org (D. Hienert)

2. Related Work

Many previous works investigated the relation between the user knowledge state change, their search behavior, and the Web resources consumed. For instance, Gadiraju et al. [4] studied the impact of information needs on the search behavior and knowledge gain of search engine users. Collins-Thompson et al. [5] studied the influence of query types on knowledge gain, finding that intrinsically diverse queries lead to increased knowledge gain. Bhattacharya et al. [6] studied the relation between eye-tracking measures and users' knowledge change. Liu et al. [7] investigated the influence of three different types of learning resource on users' learning outcome in search sessions.

Effort has also been made in assessing user knowledge state/gain with automated approaches. User interaction features [2, 8], web resource content features [9, 10], and multimedia features [11] have been considered by previous works to build classification models to predict user knowledge state and knowledge gain in search sessions. Gwizdka et al. [12] proposed to assess learning outcomes in search environments by correlating individual search behaviors with corresponding eye-tracking measures.

With this extended understanding on human learning in web search, the next goal is to optimize search systems to better support user learning. Syed et al. [13] proposed a retrieval algorithm that focuses on diversification to help with the exploration of topics. In later works [14, 10], Syed et al. 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, and proposed a theoretical frameworks accordingly. However, the question of how to improve users' learning gains through re-ranking in a general search engine context has not been sufficiently explored. In this work, we explored the use of learning-to-rank techniques to improve users' knowledge gain.

3. SAL Ranking Framework

As a base for our experiments, we use the openly available SAL-Lightning Dataset [15]. It was created from a lab study with 114 participants searching freely on the Web to learn about the formation of lightning and thunder. This is a complex topic which needs the understanding of several interwoven concepts. Participants could use every Web resource they would like within 30 minutes. Their knowledge states were measured before and after the search session with multiple-choice questionnaires and self-written essays. User behavior and resource features were recorded by screen recordings, visited Web pages, browsing timelines, gaze data, browser interaction data, knowledge data, and questionnaires.

A first analysis [16] shows that participants use search engines in their learning tasks to identify useful resources by scanning through search engine result pages (SERPs). Then they browse resources such as textual Web pages or video pages, checking for topic-related content to read or watch. Afterward, they return to the SERP, inspect resources further down in the result list, or refine their queries and start the cycle again.

Experimental Dataset. For this paper, we use a selection of the logged resources listed above. Namely, we use the search interaction data for 74 of the participants, excluding participants

with partially missing data[1].

Based on the materials explained in the resource paper, we build two more resources: (1) from the SERPs we extracted the search number and URL, the search type (web, images, videos, news, books) and the query terms. For every linked resource such as text links, images, videos and knowledge graph on the SERP we extracted the position, URL, title, and snippet. (2) For web pages clicked by a participant on a SERP, the HTML with the actual text content was obtained during the lab experiments. However, for those links not clicked, we only have the URLs. These resources were subsequently crawled from the web archive[2] with a snapshot date as close as possible to the time of the original experiment. The final data for our analysis consists of 706 rankings, 465 of which contain clicks on search results. These rankings consist of 25,829 links and for 99.34% we have the HTML content.

In line with [2], participants were divided into three groups based on their initial knowledge state (KS) measured by the multiple-choice questionnaires before the experiment (pre−KS). The number of users per group and some general statistics on the users' interaction with the rankings are given in Table 1.

Table 1
User interaction statistics. Users are distinguished based on their pre-session knowledge state into low, mid and high pre−KS groups.

pre-KS	users	rankings per user (with clicks)	clicks per user	mean click rank (first click)
All	74	9.54 (6.28)	9.26	5.20 (4.17)
Low	23	9.34 (6.35)	9.44	4.95 (3.84)
Mid	28	10.14 (6.35)	9.44	5.69 (4.77)
High	23	9.0 (5.39)	9.17	4.87 (3.62)

Clickthrough data and ranking labels. To optimize rankings for learning outcomes, the dataset offers two signals of result usefulness as possible basis of ranking optimization goals. Firstly, the clickthrough data obtained in the learning setting offers the first relevance signal to learn and evaluate a reranking model. To this end, the relevance label is taken to be 1 in case the result was clicked by the user and 0 otherwise. Secondly, the knowledge state measurements based on questionnaires before and after the search sessions allow for estimates on the usefulness of search results for learning outcomes.

As indicated in Figure 1, we assume that the contribution of a given document to the users' learning outcome is proportional to the dwell time. Hence, we devise a relevance label that attributes the knowledge gain achieved in the session to individual documents based on the dwell times. We compute the KG(d) relevance label for a document d as follows:

$$KG(d) = \frac{(post\text{-}KS - pre\text{-}KS) + 1}{10 - pre\text{-}KS + 1} \cdot \frac{dt_d}{\sum dt_i} \tag{1}$$

[1]Among others, 20 participants were excluded due to a bug in the tracking scripts that resulted in erroneous SERP interaction data at the beginning of the study.
[2]http://archive.org

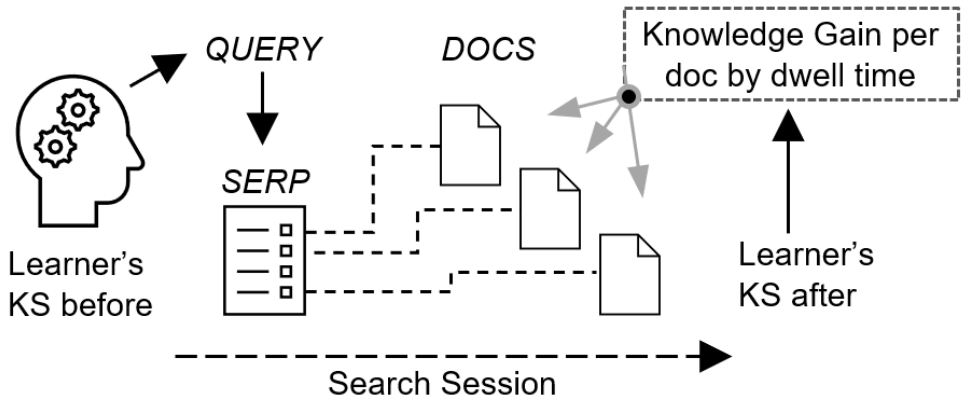

Figure 1: The Knowledge Gain after the learning session can be attributed to individual viewed documents in the session weighted by dwell time

where the user's knowledge gain after the session is computed as the difference between the knowledge state before ($pre\text{-}KS$) and after the session ($post\text{-}KS$), and dt_d is dwell time on document d. In the first term of the formula, we normalize the KG(d) by the maximal achievable value and add smoothing terms; in the second term we weight by dwell time, as explained above[3].

Ranking Features. We use the following 17 basic (re-)ranking features geared towards pointwise learning to rank. In addition to the position in the original ranking and the length of the query, we compute the following 4 features based on the title, the snippet, and the content of each search result, respectively:

- "sum_qterm": Sum of the number of query term occurrences within the search result.
- "jaccard_sim": the Jaccard similarity between search result and query.
- "bm25": BM25 measure based on search result field and query. Parameters were chosen according to [17].
- "bm25_ft": bm25 measure computed with alternative document frequency to account for the topically narrow dataset. Word counts from the fasttext[4] word vector model for the German language were used as a substitute.

We extracted the textual content from the crawled search results with the help of the Inscriptis library [18].

Content Features. We use the same 114 features as in [11] to represent the textual information in the web documents. These are computed from three different perspectives: 1) Complexity of the textual content in the document, including both the descriptive metrics (e.g. number of words, length of sentences) as well as scientifically defined complexity or readability measures (e.g. Gunning Fog Grade[5], Flesch-Kincaid Grade [19]); 2) HTML structure of the web

[3]Note that the sum of KG relevance labels for each user is less than 1 and the labels are therefore much lower when compared to the click based labels, with an average value of 0.058 for relevant documents.
[4]https://fasttext.cc/
[5]http://gunning-fog-index.com/

document, which indicates the type of content (e.g. existence of item list) and how a document is organized (e.g. length of paragraphs); 3) Linguistic features that reflect the psychological processes, sentiment, and the writing style of the content, which are computed based on the 2015 Linguistic Inquiry and Word Count (LIWC) dictionaries[6].

Experimental Setup. Based on the two kinds of relevance labels defined above, we compare basic pointwise learning-to-rank schemes against two baselines: a simple ranking based on the BM25 measure using the substitute document frequencies, and the original ranking given by the Google search. The learning models used in this work are *Lasso*, *Ridge*, and *Random Forest Regression*[7].

The models were trained and evaluated in a user-wise leave-one-out cross-validation scheme, whereby the models were iteratively evaluated on the rankings of one user and trained on the rest. Hyperparameters were not optimized. The performance is measured in terms of Normalized Discounted Cumulated Gain (NDCG) and Precision@10 (P@10) for the binary 'clicked' ranking goal. For the knowledge gain oriented rankings, we replace the precision measurement with a corresponding KG@10 measure, which is the average KG relevance among the 10 highest ranked results.

4. Results

We compare ranking performance to the original ranking in Tables 2 and 3 for each of the two target variables. We distinguish performances based on the pre−KS user groups; however, the models are always computed on the complete dataset.

Table 2
Ranking performance for the label "clicked". Best results in bold. Significance was tested in a pairwise manner against the original ranking.

Method	Low pre−KS		Medium pre−KS		High pre−KS		All	
	NDCG	P@10	NDCG	P@10	NDCG	P@10	NDCG	P@10
Lasso	0.2970	0.0752	0.2935	0.0620	0.2962	0.0734	0.2953	0.0694
Ridge	0.3109	0.0752	0.2873	0.0630	0.2894	0.0710	0.2951	0.0691
Random Forest	**0.3303**	**0.0799*****	**0.3268**	**0.0676*****	**0.3147**	**0.0826*****	**0.3243***	**0.0757*****
BM25	0.2926	0.0664	0.2793	0.0556	0.2957	0.0715	0.2882	0.0635
Original Ranking	0.3185	0.0584	0.3062	0.0535	0.2915	0.0570	0.3056	0.0560

Note: *p<0.1; ** p<0.05; *** p<0.01

Ranking performance. Overall, the re-ranking approaches outperform the original ranking baseline for both ranking tasks. Among the tested approaches, Random Forest performs best in most situations, with P@10 and KG@10 significantly improved when compared to the original ranking.

There are differences in performance between the two ranking tasks, which are in line with our expectations based on the different value ranges of the underlying relevance labels described

[6] http://liwc.wpengine.com/compare-dictionaries/
[7] The experiments were carried out using scikit-learn version 1.0.1.

Table 3
Ranking performance for the label "KG". Best results in bold. Significance was tested in a pairwise manner against the original ranking.

Method	Low pre−KS		Medium pre−KS		High pre−KS		All	
	NDCG	KG@10	NDCG	KG@10	NDCG	KG@10	NDCG	KG@10
Lasso	0.2949	0.0048	0.2846	0.0033	0.2348	0.0047***	0.2731	0.0042
Ridge	0.2895	0.0049	0.2748	0.0039	**0.2735**	0.0046	0.2789	0.0044
Random Forest	**0.3261**	**0.0056****	**0.3211**	**0.0041*****	**0.2735**	**0.0045**	**0.3087***	**0.0047*****
BM25	0.2859	0.0044	0.2731	0.0028	0.2692	0.0042	0.2758	0.0037
Original Ranking	0.3092	0.0039	0.2968	0.0031	0.2635	0.0031	0.2908	0.0033

Note: *p<0.1; ** p<0.05; *** p<0.01

in Section 3. However, there are also differences in performance between pre−KS groups. The ranking appears to be more challenging to optimize for high pre−KS users. Particularly for the "KG" model results shown in Table 3, both the original ranking, as well as the re-ranked results lists perform worse in terms of NDCG, when compared to low or medium pre−KS user rankings. This suggests that for the lower pre−KS users there is more potential to optimize for improved learning outcomes.

Feature importances. Figure 2 shows feature importance in terms of mean decrease in impurity, computed with the random forest model. Overall, the ranking related features are the most useful features for both ranking tasks – except for the content length, all of them appear among the 25 most useful features. The most important feature overall is the text query length. As is expected for a re-ranking task, the original ranking is the second most useful feature on average and even, barely, the most important feature for the 'clicked' relevance task. Content features on the other hand are overall lower ranked, although still useful apparently. One could assume that information needs evolve throughout a learning session and the query length might be an indicator of that, with longer, more specific, queries in the later stages of learning sessions. However, upon correlating query length to session progress, we find that only the high pre-KS users exhibit a relationship, with significantly shorter queries ($p < 0.01$) in the second halves of the sessions – queries in the first half were dominated by general queries on formation of thunderstorms, while in the second half of the sessions, short and specific queries into aspects and technical terms could be observed. This difference might be one of the ways in which the query length helps the models to optimize rankings for improved learning outcomes.

In terms of differences between the ranking tasks, among the most useful features we observe higher importance values for the 'ranking_query_length' and 'ranking_snippet_bm25_ft' features for the knowledge gain based relevance prediction, when compared to ranking based on relevance derived from clicks. This indicates that there may be a difference when optimizing for knowledge gain directly, over just optimizing the rankings for the click based relevance, which, in contrast, is an indicator of the usefulness of results perceived by the users.

5. Conclusions and Future Work

In this work we proposed a general framework for re-ranking search results regarding knowledge gain to optimize rankings for the learning outcome. We applied the framework to an existing search as learning dataset, showing that the ranking can be improved towards higher knowledge gain. In addition, our results indicate that for users with lower amounts of knowledge going into the session, there appears to be more potential to optimize for improved learning outcomes. In terms of features, the query length was particularly helpful in optimizing rankings for improved learning outcomes, with some users issuing shorter queries for more specific technical terms in the later part of the sessions. Our results also indicate that there might be differences when optimizing for KG directly, when compared to click based relevance indicators.

This work also has some limitations. We applied our SAL re-ranking framework to only one learning task and topic and to a limited number of participants and rankings. This could have influenced the results. Additionally, more specialized content extraction and representation approaches geared towards images and videos might be more appropriate for these types of search results. Finally, attributing knowledge gain to documents by dwell time is only a rough

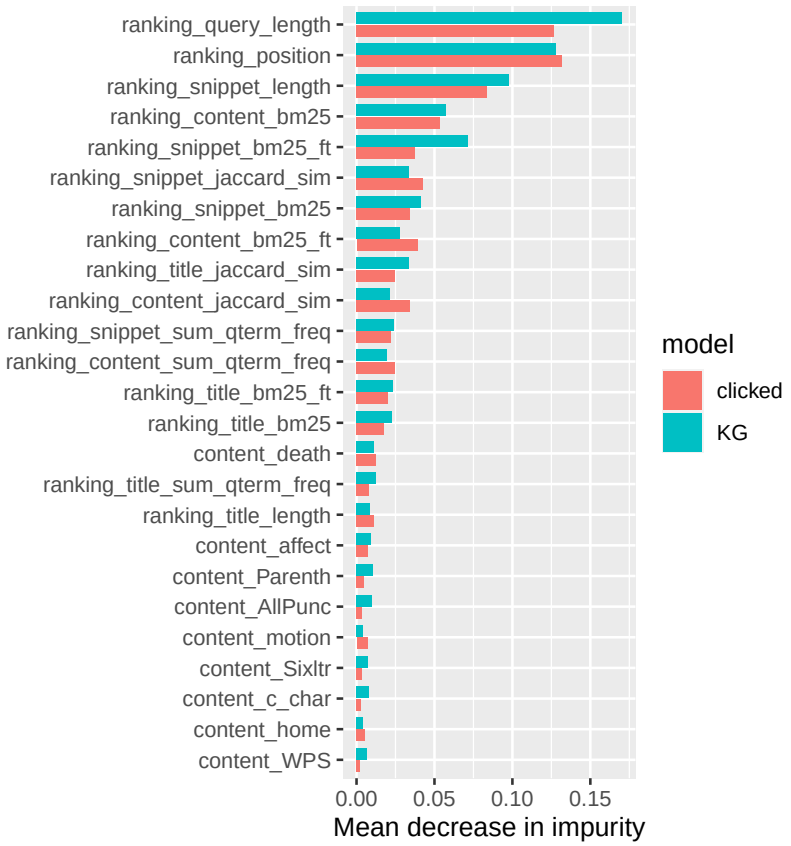

Figure 2: Feature importances based on mean decrease in impurity for the "clicked" and "KG" labels. The 25 most important features are shown, ordered according to the sum of importance values.

hypothesis. Higher dwell times could also indicate difficulties with the legibility of the document depending on the task and topic. However, we think that attributing the overall knowledge gain to the consumed resources within the session is a natural way to find the most helpful resources for learning – and in a second instance, these resources should be ranked higher. Therefore, we encourage other researchers to use our framework for other learning tasks and topics to understand the effects better.

Previous works indicated the possibility of identifying search sessions with a learning intent using in-session data. The preliminary findings in this work show that search result ranking can be optimized towards knowledge gain. Combining these insights, this work serves as a starting point for search engine optimization for human learning from the retrieval and ranking perspective. In future work, we will investigate the impact of knowledge gain oriented re-ranking strategies in real-world search sessions through field studies and continue improving the re-ranking algorithms.

Acknowledgments

This work is partially funded by the Leibniz Association, Germany (Leibniz Competition 2018, funding line "Collaborative Excellence", project SALIENT [K68/2017]).

References

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

[2] R. Yu, U. Gadiraju, P. Holtz, M. Rokicki, P. Kemkes, S. Dietze, Predicting user knowledge gain in informational search sessions, in: K. Collins-Thompson, Q. Mei, B. D. Davison, Y. Liu, E. Yilmaz (Eds.), The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018, ACM, 2018, pp. 75–84.

[3] A. Hoppe, P. Holtz, Y. Kammerer, R. Yu, S. Dietze, R. Ewerth, Current challenges for studying search as learning processes, in: Linked Learning Workshop – Learning and Education with Web Data (LILE), in conjunction with ACM Conference on Web Science, 2018.

[4] U. Gadiraju, R. Yu, S. Dietze, P. Holtz, Analyzing knowledge gain of users in informational search sessions on the web, in: C. Shah, N. J. Belkin, K. Byström, J. Huang, F. Scholer (Eds.), Proceedings of the 2018 Conference on Human Information Interaction and Retrieval, CHIIR 2018, New Brunswick, NJ, USA, March 11-15, 2018, ACM, 2018, pp. 2–11.

[5] K. Collins-Thompson, S. Y. Rieh, C. C. Haynes, R. Syed, 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, 2016, pp. 163–172.

[6] N. Bhattacharya, J. Gwizdka, Relating eye-tracking measures with changes in knowledge on search tasks, in: Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, 2018, pp. 1–5.

[7] C. Liu, X. Song, How do information source selection strategies influence users' learning outcomes', in: Proceedings of the 2018 Conference on Human Information Interaction & Retrieval, 2018, pp. 257–260.
[8] X. Zhang, M. Cole, N. Belkin, Predicting users' domain knowledge from search behaviors, in: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, ACM, 2011, pp. 1225–1226.
[9] R. Yu, R. Tang, M. Rokicki, U. Gadiraju, S. Dietze, Topic-independent modeling of user knowledge in informational search sessions, Information Retrieval Journal 24 (2021) 240–268.
[10] R. Syed, K. Collins-Thompson, 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, 2018, pp. 191–200.
[11] C. Otto, R. Yu, G. Pardi, J. v. Hoyer, M. Rokicki, A. Hoppe, P. Holtz, Y. Kammerer, S. Dietze, R. Ewerth, Predicting knowledge gain during web search based on multimedia resource consumption, in: International Conference on Artificial Intelligence in Education, Springer, 2021, pp. 318–330.
[12] J. Gwizdka, X. Chen, Towards observable indicators of learning on search., in: SAL@ SIGIR, 2016.
[13] R. Syed, K. Collins-Thompson, Optimizing search results for human learning goals, Information Retrieval Journal 20 (2017) 506–523.
[14] R. Syed, K. Collins-Thompson, Retrieval algorithms optimized for human learning, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2017, pp. 555–564.
[15] C. Otto, M. Rokicki, G. Pardi, W. Gritz, D. Hienert, R. Yu, J. von Hoyer, A. Hoppe, S. Dietze, P. Holtz, Y. Kammerer, R. Ewerth, Sal-lightning dataset: Search and eye gaze behavior, resource interactions and knowledge gain during web search, in: ACM SIGIR Conference on Human Information Interaction and Retrieval, CHIIR '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 347–352.
[16] G. Pardi, J. von Hoyer, P. Holtz, Y. Kammerer, The role of cognitive abilities and time spent on texts and videos in a multimodal searching as learning task, in: Proceedings of the 2020 Conference on Human Information Interaction and Retrieval, 2020, pp. 378–382.
[17] T. Qin, T.-Y. Liu, J. Xu, H. Li, Letor: A benchmark collection for research on learning to rank for information retrieval, Information Retrieval 13 (2010) 346–374.
[18] A. Weichselbraun, Inscriptis–a python-based html to text conversion library optimized for knowledge extraction from the web, arXiv preprint arXiv:2108.01454 (2021).
[19] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, B. S. Chissom, Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel, Research Branch Report 8-75, Naval Technical Training Command Millington TN Research Branch, 1975.