Search as Learning: Comparing Knowledge Gain Estimation Methodologies

Jagjit Singh

San Jose State University, San Jose CA 95192, USA

Abstract. Understanding the learning goals and information needs of users during a web search session is a critical aspect of a Search as Learning system to allow it to gain awareness about the users' knowledge acquisition process. Through this understanding, the system can keep up with the evolving information needs of the user by adapting search results, maximizing the facilitation of learning in the process. Recent research has demonstrated that the dynamic approach of inferring learning goals of users through their web interactions is more effective than assigning predefined learning goal representations to search topics, as different people can have different learning goals for the same topic. And using these learning goals, estimations can be made about the user's knowledge gain during the course of a search session.

In this study, we test five different methods which can be used to make estimations about a user's knowledge gain. We compare the result of these approaches to the baseline. Our results show a slight improvement over previous works and possibly point to a more effective strategy in measuring knowledge gain of users during web search sessions.

Keywords: Search as Learning, Inferring Learning Goals, Knowledge Gain Estimation.

1 Introduction

Search as Learning (SAL) is a field of study that proposes that Information Retrieval (IR) systems should facilitate learning by personalizing the web search experience. Modern search engines, however, are instead mostly concerned with performing simple look-up tasks and maximizing sold advertisements [3]. And while LLMs can provide personalized information, Melumad and Yun [4] demonstrate that users develop a shallow knowledge using LLMs than when they learn through standard web search. Therefore, there's a need for learning-oriented search systems that can support complex learning tasks.

In order to build learning-oriented search systems, a deeper understanding needs to be developed regarding the way humans acquire knowledge and learn. Insights into the human learning process can be integrated into an IR system to allow it to gain awareness of the user's evolving information needs. As a result, the system will be able to develop an understanding of the user's learning goals, track their knowledge state throughout a search session, and adapt the search results in a way that maximizes learning.

In a recent study, Nasser et al [1] employed a clustering approach to dynamically infer learning goals of users based on their web-activity, taking into consideration things like user search queries and documents clicked. Based on these learning goals, they estimated knowledge of users and compared it to actual knowledge gains. They demonstrated considerable improvements over previous approaches which relied on pre-defined and fixed and learning goal representations for search topics [1].

This paper will build on the methodologies discussed in [1] and suggest potential improvements that can be used to better model the human knowledge acquisition process. More specifically, in this paper, five different methodologies will be tested and compared in their effectiveness at making estimations about the knowledge acquisition process of users during web search sessions. By doing this, this paper aims to make a meaningful contribution to the field of SAL.

2 Dataset

The dataset employed in this paper is the same as the one used by the baseline [1] and it comes from a crowdsourcing study done by Gadiraju et al [2]. The dataset consists of 500 unique users and data regarding their search sessions on 10 different topics. The information includes logs of user activity, consisting of things like submitted queries, visited web documents, page dwell time, etc. Pre-search and post-search test assessment scores with respect to search topics are also provided, which correspond to actual knowledge gains. These actual knowledge gains are compared in this paper with the estimated knowledge gains to measure the performance of the suggested approach.

Similar to the baseline [1[, we refine the dataset by excluding users who demonstrated a negative knowledge gain, meaning users who ended up having a lower post-search test assessment score than a pre-search score. Non-English submitted queries are ignored as well. For extracting the text content of the URLs visited by the users, the Beautiful Soup Python Library is utilized.

3 Approach

3.1 Inferring Learning Goals

After the user submits a query, the search results are clustered into *k*-clusters using *k*-Means clustering. Each cluster then represents a subtopic within the general search query topic. To actually do the clustering, the documents are first converted into Term Frequency-Inverse Document Frequency (TF-IDF) vectors. And then the *KMeans* utility function provided by the *sklearn* library is used to cluster the TF-IDF document vectors into *k* clusters.

As a next step, a goal text vector is calculated for each cluster in a query. The goal texts for a cluster consist of the top *h* terms with the highest cumulative TF-IDF values in the documents of the cluster. The goal text vector is seen to represent the information needs of the user for that particular cluster.

From here on, user activity begins to provide valuable insights into their learning goals and interests. If a user clicks on documents in the same cluster again and again, the system gains confidence that the information the user's looking for consists of documents inside that particular cluster. On the other hand, if the user switches clusters, that might indicate a shift in their information needs or it might be indicative of a more general interest rather than a specific one. For each query the user submits, new clusters are generated. In this way, the system continuously keeps up with the changing information needs and interests of the user.

3.2 Estimating Knowledge Gain

Since the goal text vector of a cluster represents the learning goal, when a user clicks on a specific document, the TF-IDF of that document is mapped onto the goal text vector of the cluster to which the document belongs. Through this mapping, a new vector is created that aligns with the goal texts of the cluster. This new mapped vector retains the TF-IDF values of the document terms that match the goal text vector and assigns zero to all other values. This mapped vector represents the knowledge that the user gained from the document.

At the conclusion of a session, all the mapped documents vectors (for all the clicked documents) are summed together. The final vector that results from this summation is taken to be the user's final knowledge state,

This brings us to the final step of the process: measuring the knowledge gain. This part of the process is where our approach differs from the baseline. In the baseline [1], the authors state that they measure the knowledge gain by calculating the cosine similarity between the final knowledge state vector and goal text vector. However, at the end of each session, there are multiple clusters across different queries with which the user will have interacted. Therefore, for each user, multiple knowledge gain values are computed aftering taking the cosine similarity between the user's final knowledge state and the various goal text vectors of the clusters with which they interacted. The baseline says nothing about how to reconcile these different knowledge gain values in a meaningful way to end up with a single knowledge gain value.

This paper looks at various different approaches to do this: (1) combining the various goal text vectors into a single combined goal text vector; (2) taking the weighted average of the cosine similarity of each cluster (more engaged clusters get more weight); (3) taking the average knowledge gain with regards

to the various clusters; (4) using only the dominant cluster (cluster that was engaged with the most); and (5) taking the median of the various knowledge gain values for the user.

Different values of h (length of the goal text vector) are also experimented with to see the effect it has on the performance of the algorithm.

4 Experiments & Results

To measure the performance of the system, like the baseline [1], we also calculated the *Pearson correlation* between the estimated knowledge gain values and the actual knowledge gain values of the users. However, unlike the baseline, which focused on 10 different topics, this paper only focuses on one topic: NASA. This is due to time and resource constraints, and the fact that it was a multiple hour process to fetch, read, filter, clean, and process the necessary amounts of documents for each cluster. With the chosen number of returned documents being 30, for a single topic like NASA, thousands of documents had to be fetched and processed. For the topic NASA, to be consistent with the baseline [1], the value of 4 was chosen for k in the k-Means clustering algorithm.

Furthermore, 20 independent runs for the k-Means clustering algorithm were performed with 20 different random seeds to account for the randomness that exists in the clustering process and measure the stability and reliability of the results.

4.1 Results & Discussion

As shown by Table I, taking approach (5), which involved taking the median of the cosine similarities of the final knowledge state vector with the various goal text vectors of the clusters that the user interacted with yielded the highest correlation coefficient with a median of 0.412 across 20 independent k-Means clustering runs. This value is higher compared to the baseline value of 0.397 for the topic NASA. This suggests that taking the median of the cosine similarities over the various engaged clusters might be a better indicator of knowledge gain compared to the method the baseline employed.

The remaining approaches (3), (1), (2), and (4) all performed worse than the baseline with values of 0.364, 0.256, 0.253, and 0.182 respectively.

Table 1. Comparison of Knowledge Gain Estimation Approaches.

Technique (ranked in order of performance)	**Correlation Coefficient (Average of 20 independent *k*-Means clustering runs)**
(5) Median of Cosine Similarities	0.412
(3) Average of Cosine Similarities	0.364
(1) Combined Goal Text Vector	0.256
(2) Weighted Average of Cluster	0.253
(4) Dominant Cluster	0.182

Additionally, the baseline [1] does not mention the chosen value of *h*, which determines the number of terms in the goal text vector for a cluster, which represents the learning goal for that cluster. Table II lists the different values of *h* that were experimented with in this research and the corresponding results for the best performing approach (5).

There's a clear pattern that emerges showing that the mean correlation increases as the value of *h* increases. However, an increasing correlation in relation to an increasing *h* does not necessarily mean that the resulting value truly represents a better knowledge gain estimation. Choosing too many goal text values, for example 50, can lead to a bloated goal text vector, causing it to overlap with vectors even if they aren't conceptually aligned. The results of this experiment were derived using an *h* value of 30.

Table 2. Comparison of different *h* values

***h* value**	Mean Correlation Coefficient of Approach (5)
10	0.300
20	0.365
30	**0.412**
40	0.416
50	0/437

5 Conclusions

In this paper, building off the work done by Nasser et al [1], we implemented an approach to measure the knowledge gain of users through an online search session. We experimented with five unique ways to measure the final knowledge gain of users and compared the results to the baseline [1]. One approach in particular, which involved taking the median of the cosine similarities between a user's final knowledge state and engaged clusters proved to be slightly more effective than the baseline. Future work includes expanding the scope of the project and including the 9 other search topics to see if the performance of the new approach holds for a more general case and not just a single topic.

References

1. Nasser, H., Da Costa Pereira, C., Escazut, C., Tettamanzi, A.G.: Estimating knowledge gain in search as learning: A clustering-based approach to inferring learning goals. 2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). 333–337 (2024).
2. Gadiraju, U., Yu, R., Dietze, S., Holtz, P.: Analyzing knowledge gain of users in informational search sessions on the web. Proceedings of the 2018 Conference on Human Information Interaction & Retrieval - CHIIR '18. 2–11 (2018).

3. von Hoyer, J., Hoppe, A., Kammerer, Y., Otto, C., Pardi, G., Rokicki, M., Yu, R., Dietze, S., Ewerth, R., Holtz, P.: The Search as Learning Spaceship: Toward a Comprehensive Model of Psychological and Technological Facets of Search as Learning. Frontiers in Psychology. 13, (2022).
4. Melumad, S., Yun, J.H.: Experimental Evidence of the Effects of Large Language Models versus Web Search on Depth of Learning. SSRN. (2025).