User knowledge and Search Goals in Information Retrieval: A benchmark and study on the evolution of users' knowledge gain

Dima El-Zein

Université Côte d'Azur, CNRS,Laboratoire I3S, UMR 7271, Sophia Antipolis, France

Abstract

This abstract presents an Information Retrieval framework that personalises results based on the user's knowledge and search goals. The framework utilises the content of the pages visited by the user to represent his/her knowledge, and a set of questions/statements the user wishes to answer to represent his/her search goals. In the absence of related datasets and benchmarks, we propose a methodology to evaluate the framework.

Keywords

Information Retrieval, User Knowledge, User Search Goals, Search Personalisation

1. FRAMEWORK AND EVALUATION

The consideration of the user's cognitive components in the domain of Information Retrieval IR was set as one of the "major challenges" by the IR community in 2018 [1]. To our knowledge, there is no research dealing with the content of the documents read by the user as his/her acquired knowledge. In general, such content has been used to construct the user profile from which the user preferences could be obtained, for example. Those profiles are usually either static or not frequently updated, therefore cannot help in representing the user's knowledge, which is constantly evolving. The constant evolution of the user's knowledge is an important aspect to be considered when proposing documents that are supposed to have novel content and/or help him/her achieve a goal not yet achieved.

The IR Framework: We propose a cognitive agent that is "aware" about its user's knowledge and goals; those information are set as the agent's *beliefs*. The user's knowledge is represented by the content of the documents he/she reads; the agent will update its beliefs about the user's knowledge after every document read. The goals are represented by the set of questions the user wishes to answer at the end of a search session. The proposed agent employs its beliefs to provide the user with documents that contain novel information in respect to what he/she already knows and that also help to reach his/her search goals. Therefore, in response to a user query, it is expected to receive a ranked list of documents that are the least similar to the user's knowledge and the most similar to his/her goals. The decision of returning a document or not is based on three elements: the knowledge, the goal, and the document to be proposed. All three elements are supposed to have a textual format; we propose 3 methods to represent them: (1) Keyword representation using RAKE - Rapid Automatic keyword extraction [2] (2) Vector representation using GloVe - Global Vectors for Word Representation [3] (3) Vector representation using BERT - Bidirectional Transformers for Language Understanding - embedding [4]. Finally, the similarity between those elements' representation is calculated and documents are returned accordingly.

Evaluation Challenges : The challenge to evaluate the framework is the lack of adequate datasets or related benchmarks. Numerous existing datasets logged search sessions' activities, however to the best of our knowledge, none did track the user's knowledge and its change after reading a document. Our idea is to obtain such information by adapting a public dataset [5] that measured the user's knowledge gain during a search session. That will allow us to evaluate the framework.

Dataset's Experiment : The dataset's experiment quantified the user's knowledge gain about a topic after a search session. The participants were provided an *information need* sentence for a specific topic, then were invited to search the web about it; their behaviour was getting logged meanwhile. They also had to respond to pre- and post-session tests that consisted of statements related to the topic. The tests assessed the participants knowledge regarding the topics and were scored based on the correctness of the answers. A user's knowledge gain was measured as the difference between the post- and pre- tests' scores.

Benchmark Creation : To estimate the *page knowl-*

DESIRES 2021 – 2nd International Conference on Design of Experimental Search & Information REtrieval Systems, September 15–18, 2021, Padua, Italy

✉ elzein@i3s.unice.fr (D. El-Zein)

iD 00000-0003-4156-1237 (D. El-Zein)

CEUR Workshop Proceedings (CEUR-WS.org)

edge gain g_i brought by each page p_i, we perform a linear regression analysis of the *user knowledge gain* G against visited pages which are binary values- visited or not visited. g_i would then be the regression coefficient. We could hence understand and predict a user's knowledge gain after visiting a set of pages P. As the user visits one page after the other, we track the cumulative evolution of the knowledge gain. We construct a benchmark containing for each user, the set of submitted queries, the related visited pages and the associated evolution of knowledge gain.

Framework Evaluation : The evaluation's idea is to submit to the framework, the set of queries submitted by every user and suppose the user read the document returned by the agent. We consider the study population to be the set of users who scored *zero* in the pre-session test, representing those having no previous knowledge about the searched topic; the agent's beliefs about the user's knowledge are then still empty. They will get updated as the user starts visiting pages. The user goals consisted of the *information need* and the test statements. For the first query submitted by a user, since the agent has no information yet about the user's knowledge, we return the same page visited in the benchmark. The agent builds its initial beliefs about its user's knowledge and starts its personalising task. For the following queries, the agent compares the content of the pages to be proposed to the agent's beliefs (both the user knowledge and goal) and decides which document to return. We track the evolution of the user knowledge gain and compare it to the benchmark.

References

[1] J. S. Culpepper, F. Diaz, M. D. Smucker, Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in lorne (SWIRL 2018), SIGIR Forum 52 (2018) 34–90.

[2] S. Rose, D. Engel, N. Cramer, W. Cowley, Automatic keyword extraction from individual documents, Text mining: applications and theory 1 (2010) 1–20.

[3] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

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

[5] U. Gadiraju, R. Yu, S. Dietze, P. Holtz, Analyzing knowledge gain of users in informational search sessions on the web, in: Proceedings of the 2018 Conference on Human Information Interaction & Retrieval, 2018, pp. 2–11.