TREC-COVID: building a pandemic retrieval test collection

Ellen Voorhees (a) and Evangelos Kanoulas (b)
(a) National Institute of Standards and Technology, Gaithersburg, Maryland, USA
(b) University of Amsterdam, Amsterdam, The Netherlands

Abstract
Assessing how good is a search engine has been an active area of development for more than three decades. During the COVID-19 pandemic however the rate of change in what people are interested in, and the available information online has introduced further challenges for search. TREC-COVID introduces a benchmark collection to evaluate search engines and provide the means to improve them under the special circumstances of a pandemic.

Key words: *COVID-19; search engine; benchmarking; systematic review; automation.*

Introduction

One of the first steps to conduct a systematic review is to search and find all the articles published on the question of interest. Currently researchers in the field of evidence-based medicine depend on medical keywords and Boolean logic for their searches and manually inspect all the resulting articles. This manual process has become unsustainable due to the vast, increasing amount of medical scientific literature. Automation calls for the use of modern search technology that goes beyond keywords, allowing searches to identify and prioritize only the most promising fraction of articles for researchers to examine (1).

But how good is modern search technology at finding articles in biomedical repositories? Can it be trusted for the important task of evidence synthesis? Does it lead to unbiased reviews? Is it better than the current methodology used in the field? Can it speed up the synthesis of evidence? Could it be better? These are important measurement questions because we cannot build better search systems if we do not know how good current systems are.

The Text REtrieval Conference (TREC)

The US National Institute of Standards and Technology (NIST) develops the infrastructure necessary to evaluate the quality of search engines. The work is done through a project called the Text REtrieval Conference (TREC) (2). The first TREC was held in 1992, which means TREC started before web search engines even existed. In fact, the first search engines were library systems that dated back to the 1960s. The researchers of that era were the first to grapple with basic questions of search engine performance: what it means for a search result to be "good" or for one result to be better than another, and whether people agree on the relative quality of different search results. Evaluating search engine effectiveness is hard in part because people don't agree surprisingly often, and while it is easy to tell when returned information is not on-topic, it is very difficult to know if a system has not returned something you would want to see. Think about it: If you as a user of a search system knew all of the information that should have been returned, you wouldn't have searched!

As a way of investigating these questions, a British librarian named Cyril Cleverdon developed a measurement device called a "test collection" (3). A test collection contains a set of documents, a sample set of questions that can be answered by information in the documents, and an answer key that says which documents have information for which questions. For example, the initial test collection that Cleverdon built contained a set of 1,400 abstracts of scientific journal articles and 225 questions that library patrons had asked in the past. Cleverdon enlisted graduate students to go through the abstracts and indicate which articles

Address for correspondence: Evangelos Kanoulas, P.O. Box 94323, 1090 GH Amsterdam, The Netherlands.
E-mail: e.kanoulas@uva.nl.

should have been given to the researcher who had asked that particular question. Once you have a test collection, you can score the quality of a search engine result by comparing how closely the search result matches the ideal result of returning all relevant documents and no nonrelevant documents.

In the ’70s and ’80s, several more test collections were created and shared among research groups. But there was a problem. To create the answer key for each question, some human had to look at all the documents to determine the relevant set. This necessarily limited the size of the test collections that could be built. To build a large test collection, you need to avoid having a human look at every document in the collection for a question while still finding the set of relevant documents for that question. It turns out that if you assemble a broad cross-section of different types of search engines and look at only the top-ranking documents from each system, you find the vast majority of relevant documents and look at a very tiny percentage of the total number of documents. TREC was the first to implement this so-called pooling strategy, and by doing so it built a sound test collection that was 100 times bigger than the other test collections that existed at the time. No single organization could produce a collection of comparable quality because it would lack the diversity of search results that are necessary.

TREC-COVID

TREC has gone on to standardize evaluation methodology and to build dozens of collections for a variety of different types of search problems. Then in March 2020 TREC launched TREC-COVID, an effort to build a test collection for search during a pandemic.

Why was a pandemic test collection needed? While test collections based on scientific articles already existed, the information needs during a pandemic are different. The biggest difference is the rate of change: Over the course of a pandemic, the scientific questions of interest change and the literature explodes. The variability in the quality of the literature increases, too, since time pressures mean a much smaller percentage of the articles are subject to full peer review. By capturing snapshots of this progression during the early part of the COVID pandemic, TREC-COVID created data that search systems can use to train for future biomedical crises.

TREC-COVID was structured as a series of rounds, with each round using a later version of the coronavirus scientific literature dataset called CORD-19 and an expanding set of queries (4, 5). The queries are based on biomedical researchers’ real questions from harvested logs of medical library search systems. TREC-COVID participants used their own systems to search CORD-19 for each query to create search results they submitted to NIST. Once all the results were in, NIST used the submissions to select a set of articles that were judged for relevance by humans with medical expertise. Those judgments were then used to score the participants’ submissions, while the set of relevant articles is a human-curated answer for the original question.

TREC-COVID resulted in a collection of 50 queries, and a total of 69,381 judgments. The test collection was used to evaluate hundreds of participating search engines and many different technologies, some of which have been deployed as online open access tools. Quality control tests of the collection itself demonstrate that having a set of diverse, high-quality search engines did indeed enable an effective collection to be built (6). TREC-COVID also confirmed the research hypothesis that hybrid search approaches in which systems incorporate users’ feedback regarding the quality of previous search results retrieve relevant articles more quickly than fully automatic approaches. Whether the quality of the developed search technology is sufficient for automating systematic reviews remains an open question; however, TREC-COVID provides the means to study and further improve search under the special circumstances of a pandemic.

Acknowledgements

This paper originates from a presentation at the International Collaboration for the Automation of Systematic Reviews (ICASR) meeting held in April 2021.

Submitted on invitation.
Accepted on 7 June 2021.

REFERENCES

1. Kanoulas E, Li D, Azzopardi L, Spijker R. CLEF 2019 technology assisted reviews in empirical medicine overview. In: CEUR Workshop Proceedings 2019 Sep 12 (Vol. 2380).

2. Voorhees EM, Harman DK, editors. TREC: Experiment and evaluation in information retrieval. Cambridge, MA: MIT press; 2005 Sep.
3. Cleverdon C. The Cranfield tests on index language devices. In: Aslib Proceedings 1967 Jun 1. MCB UP Ltd.
4. Roberts K, Alam T, Bedrick S, Demner-Fushman D, Lo K, Soboroff I, Voorhees E, Wang LL, Hersh WR. TREC-COVID: rationale and structure of an information retrieval shared task for COVID-19. Journal of the American Medical Informatics Association. 2020 Sep;27(9):1431-6.
5. Voorhees E, Alam T, Bedrick S, Demner-Fushman D, Hersh WR, Lo K, Roberts K, Soboroff I, Wang LL. TREC-COVID: constructing a pandemic information retrieval test collection. In: ACM SIGIR Forum 2021 Feb 19 (Vol. 54, No. 1, pp. 1-12). New York, NY, USA: ACM.
6. Voorhees E, and Roberts K. On the quality of the TREC-COVID IR test collections. In: International ACM SIGIR Conference on research and development in information retrieval. New York, NY, USA: ACM.