It's messy out there: DBC's journey towards its first test collection

Serena Canu [1]

[1] DBC Digital A/S, Tempovej 7-11, Ballerup, Denmark

Abstract

The use of test collections to evaluate the effectiveness of Information Retrieval (IR) systems is wide-spread, and the literature covers many examples of improvements and ways to solve specific problems. In this abstract, we explore the initial difficulties of building and implementing a test collection outside of academic walls and without having easy or immediate access to many of the main tools discussed in academic papers.

This is an example of how a test collection can be a way to engage companies in finding creative solutions to make a first step towards the evaluation of IR systems.

Keywords

Test collections, IR systems evaluation, Information retrieval

1. Introduction. Back to the basics

Test collections have been used for decades as fundamental tools to evaluate IR systems [1], and they might even look like an easy-to-implement tool to many experts and scholars. The reality is that adopting a test collection impacts almost every aspect of an organization, especially if the organization was not structured to perform constant evaluation to begin with.

Nonetheless, there are situations in which using a test collection is a crucial step to evolve and improve the company's products. This has been the case at DBC Digital, a Danish company whose main task is to develop and maintain the bibliographic and IT infrastructure of the Danish public libraries. Among other things, DBC develops and deploys the search engine used by the public website bibliotek.dk, which gives access to the common catalogue of the Danish public libraries.

After years focusing mainly on the efficiency of the system, the need for a new way to evaluate the search engine was a necessary step forward. Bringing a perspective focused on effectiveness represented some sort of small revolution, and it brought back at the center the question "what do the users need?".

Unfortunately, good intentions per se were not sufficient, and we faced several challenges to understand how to create and use a test collection without having any prior experience. We settled on the work done by Sanderson and coll. [2][3][4] to guide our work, since they presented a useful and practical summary to understand the basic steps to make a representative test collection: a set of real queries, a set of real - or at least realistic – narratives, and a way to make relevance judgments.

Query logs, and user data analysis are often considered as a ground base to understand users' needs and define the list of queries [5]. When DBC decided to include a test collection among its tools, there were no comprehensive query logs to be used for an analysis, no recent data about loans or other users' behavior, nor the possibility to directly involve real users. And this is where we had to find creative ways to overcome the uncertainty, using as a sole starting point a dataset

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

EMAIL: seca@dbc.dk (A. 1)
ORCID: 0000-0001-9873-3218

with a list of the most searched queries in 2018 through the DDB CMS, the CMS used by the Danish public libraries. A work of classification and interpretation was then necessary, to try to give some sense to this dataset.

We started our test collection with 77 queries. This included both some of the most-searched queries and queries which we deemed to be challenging to our search engine. The corresponding narratives - initially written by a sole expert and assessor - were collectively reviewed with the help of colleagues with different expertise and backgrounds. The descriptions were then narrowed down for convenience.

This form of brainstorming, even though unconventional, proved to be particularly useful during the process, since it helped to manage the scarcity of resources, and started a discussion about - and a broader understanding of - the concept of relevance [6]. In fact, what might look like trivial questions to evaluation experts were in fact crucial steps for the definition of DBC's test collection.

The documents were initially chosen following an alternative method to the traditional pooling, and the judgments were made using a graded scale. With this first attempt, we were able to get an idea on how good our current search engine was, using the traditional metrics of Precision, Recall, F-measure, and nDCG.

Equally important, this was only the first step towards a new perspective that is consistently taking its place within the company, underlying the necessity of gathering more data, involving more experts, and finding ways to diminish some initially unavoidable biases.

2. More than just a test collection

Apart from the evaluation of the current and new IR systems, the test collection has also been used for other cross-departmental projects. For instance, as a baseline to observe possible differences between indexing strategies using curated metadata, full text indexing, and ML.

Also, it is a concrete way to communicate with our customers and QA, to explain which behavior they can expect from the search engine, and to define additional functionalities.

Research about IR has evolved significantly, so much that it might be hard to be aware of the real challenges that an average company has to face to implement evaluation tools in its practices. We believe that, sometimes, going back to the basics and dealing with a messy development can still be an option. Especially if the main result is to ignite a conversation, modifying entirely the company's understanding of what IR systems evaluation can mean. And despite the compromises made along the way.

References

[1] S. Robertson, On the history of evaluation in IR. Journal of Information Science *34*(4) (2008), 439-456.

[2] P. Clough, M. Sanderson. Evaluating the performance of information retrieval systems using test collections. Information research, (2013), 18(2), 18-2.

[3] M. Sanderson, M. Braschler, Best practices for test collection creation and information retrieval system evaluation, TrebleCLEF Project, 2009.

[4] M. Sanderson, Test collection based evaluation of information retrieval systems. Foundations and Trends in Information Retrieval, (2010) 4-4, 247-375.

[5] W. B. Croft, D. Metzler, T. Strohman, Search engines: Information retrieval in practice, 1st ed., Addison-Wesley, Boston, MA, 2010.

[6] T. Saracevic, Relevance reconsidered, in: Proceedings of the second conference on conceptions of library and information science (CoLIS 2), ACM Press, New York, NY, 1996, pp. 201-218.