Summary about *The Philosophy of Information Retrieval Evaluation*

Original article by: Ellen M. Voorhees

Summary by: Carlos Andrés Reyes Evangelista | Universidad de las Américas Puebla

Table of contents

Evaluation of information retrieval systems is performed to quantify and determine how well a system meets the information needs of the users. There are two classes of evaluation: **user based** and **system** evaluation. User based evaluation is more accurate for define at what extent a retrieval system meets the information needs of users, whereas system evaluation focuses on how well the system can rank documents.

User-based evaluation is mostly preferable, but is also more expensive and difficult to execute correctly. System evaluation works over the same abstraction that the retrieval process is using and allows experimenters to control some variables that affect retrieval performance leading to more trustworthy and less expensive experiments.

Laboratory testing of retrieval systems was first done in the Cranfield 2 experiment.e The experiment introduced a paradigm called *Cranfield paradigm*. In the rest of the paper the assumptions inherent to the latter paradigm are analyzed to prescribe when such system testing is appropriate.

The Cranfield paradigm

Originally, the experiment consisted in an investigation that would lead to which of several alternative indexing languages was best. A design goal for the Cranfield 2 experiment was to create *“a laboratory type situation where, freed as far as possible from the contamination of operational variables, the performance of index languages could be considered in isolation”*. The same sets of documents and information needs were used for all the languages. It was used **precision and recall** to evaluate effectiveness. Relevance was based on topical similarity.

A test collection consists of three different components: the components, the statements of information need and a set of relevance judgments.

Cranfield assumptions:

1. Relevance can be approximated by topical similarity
    - All relevant documents are equally desirable
    - The relevance of one document is independent of the relevance of any other document
    - Information need is static
2. A single set of judgments for a topic is representative of the user population.
3. Lists of relevant documents for each topic is complete
    - All relevant documents are known

Due that these assumptions are not generally true, the evaluation of retrieval systems become a noisy process, nonetheless, a standard experimental **design** to decrease the noise has been developed and formed part of the Cranfield paradigm. This design works by making each competing retrieval strategy to produce and order a ranking of documents for each topic (the order is based on the likelihood that each document has to be retrieved as relevant). The effectiveness of a strategy for each topic is computed as a function of the rankings it produced. A final effectiveness of the strategy is computed by obtaining the average among the effectiveness that strategy obtained for each topic.

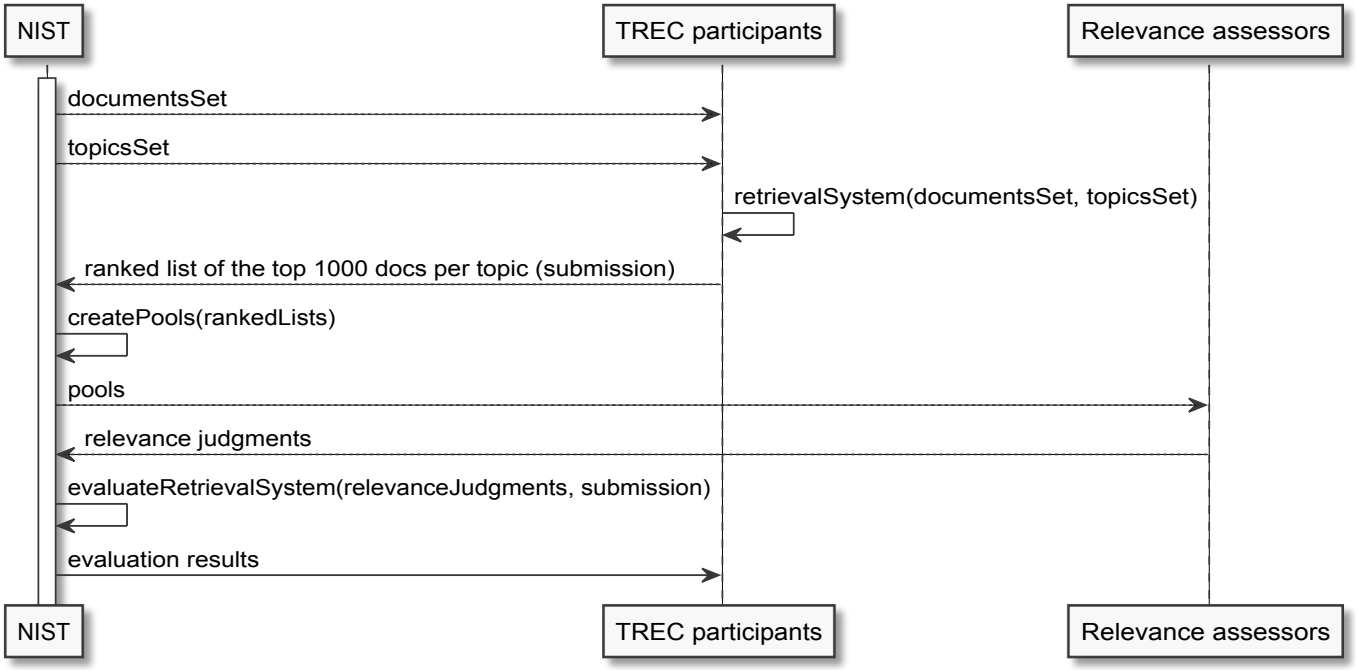
Since retrieval system effectiveness is known to vary widely across topics, the greater the number of topics, the more confident the experiment can be. Even though, the main focus in the abstraction of the Cranfield paradigm is comparative results, i.e., the absolute score of an evaluation is worthless in isolation: the only valid use of such score is to compare it to the score of other retrieval systems obtained in equal footing.

Completeness of relevance judgments

Within the Cranfield paradigm, two elements especially important: collection size and completeness of relevance judgments. Cleverdon, of the Cranfield experiments, assures that having a complete set of relevance decisions was vital, even more than having a large collection size. Evidence, though, shows that collection size matters. The problem to issue is that the larger the collection is, the harder it becomes to have a complete set of relevance judgments.

To address this problem a technique called pooling is created, to pool means to create a subset of the documents to judge for a topic, only the elements in the topic are personally judged by the topic author. Any document not in the pool is assumed to be irrelevant to the topic.

Building large test collections



The TREC provide topics, among other reasons, to make a clear statement of what criteria make a document relevant. A topic statement consists of: an identifier, a title, a description and a narrative. The topic statements are created by the person who performs the relevance assessments. The NIST TREC then selects a set of topics from the candidate topics. **The relevance judgments are what turns a set of documents into a test collection.**

Effects of Incompleteness

Pooling breaks one of the basic assumptions of the Cranfield paradigm since it does not produce complete judgments. By not considering all the documents, it is possible to end with test collections that contain relevant documents that have not been judged.

Differences in relevance judgments

Whereas incompleteness is a relatively new criticism to the Cranfield paradigm, the inconsistency of relevance judgments, result of the subjective selection of relevant documents, is the main perceived problem with test collections. Critics question how objective conclusions can be drawn when the evaluation process is based on a subjective process.

To overcome this issue, NIST created a process where each topic is evaluated for its author ─primary assessor─ and at least two more assessors, for whom a new pool to evaluate was created. This latter pool consisted of a random sample of documents that the principal assessor had judged relevant and a equivalent amount of documents that were judged to be irrelevant.

Assessor agreement

Overlap of the relevant document sets can be used to quantify the level of agreement among different sets of relevance assessments.

Effect of Inconsistency

In order to understand how the inconsistency really affects the systems, an exhaustive experiment was performed by comparing how well the ranking systems behave on different sets of judgments ─the ones created by each assessor─ about the same topic. The experiment was performed by running the systems over each set of judgments (*qrels*), the union of them ─where at least one assessor considered the documents relevant─ and the intersection of them ─such that all the assessors agreed on the relevance of the document.

This entire experiment was repeated several times using different evaluation measures, different topic sets, different systems, and different groups of assessors. The correlation between system rankings was very high in each experiment, thus confirming that the comparative evaluation of ranked retrieval results is stable despite the idiosyncratic nature of relevance judgments

Cross-Language Test collections

Cranfield paradigm may create collections of cross-language documents, but is more difficult than monolingual, since it requires to have a set of assessors for each language. Pooling in multi-lingual collections is also way harder, because the number of runs submitted by participants and the documents retrieved within a run are usually skewed in favor of more popular languages. Furthermore, the pools for the minority of languages are smaller and less diverse than the others, this introduces an unknown bias into the judgments.

Conclusion

Test collections are research tools that provide a means for researchers to explore the relative benefits of different retrieval strategies in a laboratory setting. Test collections are useful because they allow researchers to control some of the variables that affect retrieval performance, increasing the power of comparative experiments while drastically decreasing the cost as compared to user-based evaluations. It was shown throughout the summary that comparative results are more stable than using a static set of binary, topical relevance judgments to represent correct retrieval behavior.

Because the assumptions upon which the Cranfield paradigm is based are not strictly true, the evaluation of retrieval systems is a noisy process. The primary consequence of the noise is the fact that evaluation scores computed from a test collection are valid only in comparison to scores computed for other runs using the exact same collection.

Important note: *This document contains an academic summary of an existing article. None of the ideas expressed here represent the author's opinion or work, but a paraphrased and summarized version of the original article: "**The Philosophy of Information Retrieval Evaluation**" published by Ellen M. Voorhees in September 2001. Available in their original profile in ResearchGates.*