TREC: Continuing Information Retrieval's Tradition of Experimentation

Ellen M. Voorhees
National Institute of Standards and Technology
Gaithersburg, MD 20899

In contrast to most areas of computer science research, information retrieval research has a rich tradition of experimentation. In the 1960's, the librarian Cyril Cleverdon and his colleagues at the College of Aeronautics, Cranfield, England, UK ran a series of tests to determine appropriate indexing languages for information retrieval [Cle67]. The findings were highly controversial at the time, though the tests are better known today for the experimental methodology they introduced. This so-called Cranfield methodology was picked up by other research groups, most notably by Gerard Salton's SMART group at Cornell University [Sal71], and was sufficiently established by 1981 that Karen Spärck Jones edited the book *Information Retrieval Experiment* [Spä81]. The Text REtrieval Conference (TREC) [VH05], started in 1992, is a modern manifestation of the Cranfield methodology that attests to the power of appropriate experimentation. The state of the art in retrieval system effectiveness has doubled since TREC began and most commercial retrieval systems, including web search engines, contain technology originally developed in TREC.

The fundamental goal of a retrieval system is to help its users find information contained in large stores of free text. The problem is challenging because natural language is rich and complex: searchers and authors can easily express the same concept in widely different ways. Algorithms must be efficient due to the amount of text to be searched. The situation is further complicated by the fact that different information-seeking tasks are best supported in different ways, and different individual users have different opinions as to precisely what information should be retrieved. The core of the Cranfield methodology is to abstract away from the details of particular tasks and users to a benchmark task called a *test collection*.

A test collection consists of a set of *documents*; a set information need statements called *topics*; and *relevance judgments*, a mapping of which documents should be retrieved for which topics. The abstract retrieval task is to produce a ranking of the document set for each topic such that relevant documents are ranked above nonrelevant documents. The Cranfield methodology facilitates research by providing a convenient paradigm for comparing retrieval technologies in a laboratory setting. The methodology is useful since the ability to perform the abstract task well is necessary (though not sufficient) to support a wide range of information-seeking tasks.

The original Cranfield experiments created a test collection consisting of 1400 documents and a set of 225 requests. Many retrieval experiments were run in the twenty years following the Cranfield tests and several other test collections were built, but by 1990 there was growing dissatisfaction with the methodology. While some research groups did use the same test collections, there was no concerted effort to work with the same data, to use the same evaluation measures, or to compare results across systems to consolidate findings. The available test collections were so small that operators of commercial retrieval systems were unconvinced that the techniques developed using test collections would scale to their much larger document sets. Even some experimenters were questioning whether test collections had out-lived their usefulness.

At this time, the National Institute of Standards and Technology (NIST) was asked to build a large test collection for use in evaluating text retrieval technology developed as part of the Defense Advanced Research Projects Agency's (DARPA) TIPSTER project. NIST proposed that in addition to building a large test collection, it would also organize a workshop to investigate the larger issues surrounding test collection use. DARPA agreed, and TREC was born.

The first two TRECs had two tasks, the ad hoc task and the routing task. The ad hoc task is the prototypical retrieval task where the system knows the set of documents to be searched but cannot anticipate the particular topic that will be investigated. In contrast, the routing task assumes the topics are static but need to be matched to a stream of new documents. This technology is used by news clipping services, for example. Starting in TREC-3, additional tasks, called tracks, were added to TREC. The tracks serve several purposes. First, tracks act as incubators for new research areas: the first running of a track often defines what the problem *really* is, and a track creates the necessary infrastructure (test collections, evaluation methodology, etc.) to support research on its task. The tracks also demonstrate the robustness of core retrieval technology in that the same techniques are frequently appropriate for a variety of tasks. Finally, the tracks make TREC attractive to a broader community by providing tasks that match the research interests of more groups.

Tracks are organized by volunteer coordinators selected from proposals submitted to the TREC program committee. Recent TRECs have had six or seven separate tracks. Figure 1 shows a schematic view of the processing performed in a typical TREC track. Track organizers provide a document set and a set of topics whose information needs can be met from those documents. A "document" is loosely defined as an information-bearing unit; newswire articles, scientific abstracts, web pages, blog posts, email messages, recordings of speech, and video clips have each been used as documents in past TREC tracks. Information needs have been mined from logs of existing commercial search systems or created especially for the task. Participants use their system to rank the documents for each topic and return the ranked lists to NIST. Human judges at NIST look at (a subset of) the returned documents and decide which documents are relevant to which requests. Based on these judgments, NIST scores the submissions and returns the results to the participants. A TREC cycle ends with a conference held at NIST where participants gather to discuss their findings, debate methodological issues, and plan the next cycle.

TREC test collections vary in size according to the needs of the track and the availability of the data, but the standard ad hoc collections generally contain between 800,000–1,000,000 documents and 50 topics. Having human judges review all documents for all topics is clearly infeasible, so a strategy for deciding which documents to examine is required. Judging a uniform random sample of the document set for a given topic is not a useful alternative since the number of relevant documents for a topic is such a small percentage of the total number of documents that the expected number of relevant documents in a reasonably-sized sample is close to zero. Instead, TREC uses a process known as pooling [Sv75] in which the judge reviews only the documents in a topic's pool. The pool for a topic is the union of the set of X top-retrieved documents for that topic by each participant (where X is usually set to 100). Since retrieval systems are designed to rank the documents most likely to be relevant first, pools created in this manner contain sufficiently many of the relevant documents that retrieval systems can be compared fairly by assuming all unjudged documents are not relevant.

An important feature of test collections is that they are reusable: once the relevance judgments are created, they can be used to score not only the original result sets that contributed to the pools but also subsequent result sets produced using the same topic and document sets. This facilitates research by allowing a tight development cycle. Given a test collection, a researcher can quickly and easily compare a variety of different alternative retrieval approaches. TREC makes both the "trec_eval" program that computes a variety of evaluation scores and the test collections it creates publicly available (subject to licensing restrictions to protect the intellectual property rights of the documents' owners) to support the broader retrieval research community.

Evaluating a retrieval system's effectiveness can be done in a variety of ways—trec_eval reports approximately 85 different numbers for a result set—but a relatively small set of measures has emerged as the de facto standard by which retrieval effectiveness is characterized. These measures are derived in some way from *precision* and *recall*, where precision is the proportion of retrieved documents that are relevant, and recall is the proportion of relevant documents that are retrieved. For ranked retrieval, a cut-off level is needed to define the retrieved set over which precision or recall is computed; for example, a cut-off level of ten defines the retrieved set as the top ten documents in the ranked list. Since precision and recall tend to be inversely related in practice, the most common way of reporting retrieval evaluation results is a plot of the average

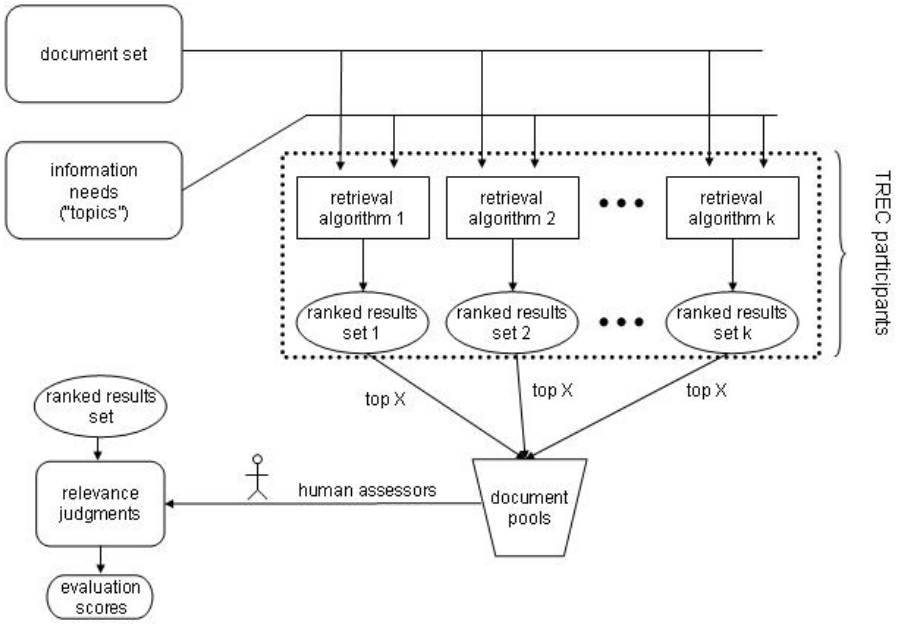

Figure 1: Processing performed in a typical TREC track. Organizers release document and topic sets to participants who use their retrieval systems to rank the documents for each topic. Ranked results are returned to NIST where pools are created for human assessors. The assessors judge each document in the pool producing the relevance judgments. The relevance judgments can then be used to score the output of both the participants' result sets and any subsequent results created using the same topic and document sets.

value of precision obtained at various standard recall levels where the average is computed over all the topics in the test collection.

While the original catalyst for TREC was the request to create a single large test collection for the classical ad hoc retrieval task, TREC has accomplished much more in its 15 year history. A variety of different collections have been constructed, including collections for languages other than English, media other than text, and tasks that range from answer finding to text categorization. In each case the test collections have been integral to progress on the task. Additional collections have been constructed in other evaluation projects based on the TREC model such as the NII Test Collection for IR Systems project (NTCIR, `http://research.nii.ac.jp/ntcir/`), the Cross Language Evaluation Forum (CLEF, `http://www.clef-campaign.org/`), and the Initiative for the Evaluation of XML Retrieval (INEX, `http://inex.is.informatik.uni-duisburg.de`).

In addition, TREC has succeeded in validating the use of test collections as a research tool for ad hoc retrieval, and has extended the use of test collections to other tasks. Using the large repository of retrieval results submitted to TREC over the years, researchers have empirically demonstrated the soundness of the conclusions reached in test collection experiments. For example, studies have examined the sensitivity and stability of different evaluation measures, the impact of experimental design decisions such as number of topics used and the size of the observed difference in retrieval scores, and the effect of changes in the

documents considered relevant to a topic [BV05]. Nonetheless, studies on the very latest collections built from millions of web pages suggest that pooling has a size dependency that prevents it from producing reusable test collections for arbitrarily large document sets. Devising new techniques for building massive test collections is thus an area of active research.

Improvement in retrieval effectiveness cannot be determined simply be looking at TREC scores from year to year: it is invalid to compare the results from one year of TREC to the results of another year since any differences are likely to be caused by the different test collections in use. An experiment conducted by the SMART retrieval group demonstrates that retrieval effectiveness has indeed improved over the course of TREC. Developers of the SMART retrieval system kept a frozen copy of the system they used to participate in each of the eight TREC ad hoc tasks. After every TREC, they ran each system on each test collection. For every test collection, the later versions of the SMART system were much more effective than the earlier versions of the SMART system, with the later scores approximately twice those of the earliest scores. While this is evidence for only one system, the SMART system results consistently tracked with the other systems' results in each TREC. SMART results can therefore be considered representative of the field as a whole.

The cumulative effort represented by TREC is significant: almost 300 distinct groups representing more than 20 different countries on six continents have participated in at least one TREC, thousands of individual retrieval experiments have been performed, and hundreds of papers have been published in the TREC proceedings. TREC's impact on information retrieval research has been equally significant. A variety of large test collections have been built and made publicly available. TREC has standardized the evaluation methodology used to assess the quality of retrieval results, and demonstrated both the validity and efficacy of the methodology. The meetings themselves have provided a forum in which researchers can efficiently learn from one another, facilitating technology transfer and improving retrieval research methodology. By evaluating competing technologies on a common task, TREC has built on information retrieval's tradition of experimentation to significantly improve retrieval effectiveness and to extend the experimentation to new subproblems.

More information regarding TREC can be found on the TREC web site, `http://trec.nist.gov`.

References

[BV05] Chris Buckley and Ellen M. Voorhees. Retrieval system evaluation. In Ellen M. Voorhees and Donna K. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, chapter 3, pages 53–75. MIT Press, 2005.

[Cle67] C. W. Cleverdon. The Cranfield tests on index language devices. In *Aslib Proceedings*, volume 19, pages 173–192, 1967. (Reprinted in *Readings in Information Retrieval*, K. Spärck-Jones and P. Willett, editors, Morgan Kaufmann, 1997).

[Sal71] G. Salton, editor. *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice-Hall, Inc. Englewood Cliffs, New Jersey, 1971.

[Spä81] Karen Spärck Jones, editor. *Information Retrieval Experiment*. Butterworths, London, 1981.

[Sv75] K. Spärck Jones and C. van Rijsbergen. Report on the need for and provision of an "ideal" information retrieval test collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge, 1975.

[VH05] Ellen M. Voorhees and Donna K. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.