By Ellen M. Voorhees

TREC: CONTINUING INFORMATION RETRIEVAL'S TRADITION OF EXPERIMENTATION

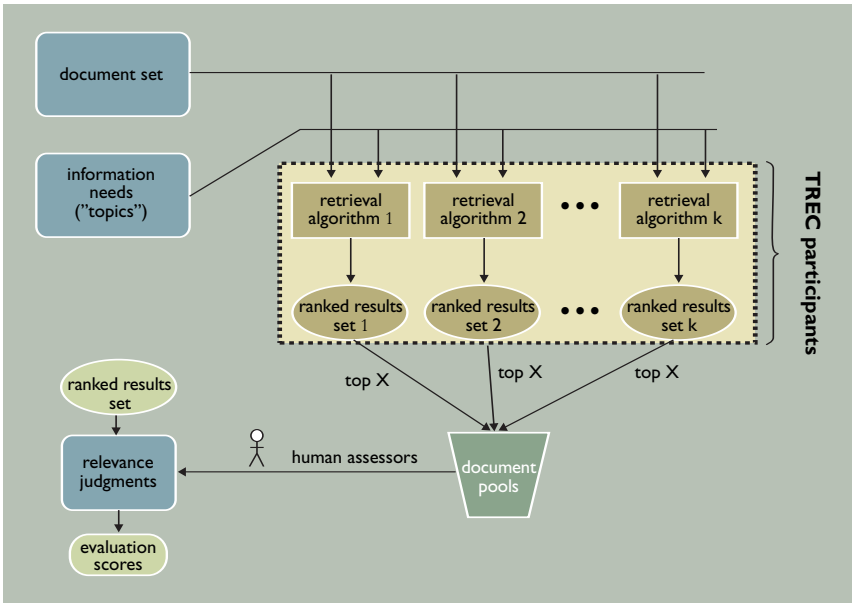
Large-scale test collections drive improvement in search technology to help users find information in free text.

Unlike most aspects of computer science research, information retrieval has a rich tradition of experimentation. In the 1960s, the librarian Cyril Cleverdon and his colleagues at the College of Aeronautics, Cranfield, England, ran a series of tests to identify appropriate indexing languages for information retrieval [2]. Their findings were highly controversial at the time, though the tests are better known today for the experimental methodology they introduced. The so-called Cranfield methodology was picked up by other research groups, most notably Gerard Salton's SMART group at Cornell University [3] and was sufficiently established by 1981 that it

A VARIETY OF COLLECTIONS HAS BEEN CONSTRUCTED, including for languages other than English, media other than text, and tasks that range from answer finding to text categorization.

was the subject of an entire book *Information Retrieval Experiment*, edited by Karen Spärck Jones of Cambridge University [4]. Beginning in 1992, the Text REtrieval Conference (TREC, trec.nist.gov/) [6] has represented a modern manifestation of the Cranfield methodology, attesting to the power of experimentation. The state of the art in retrieval system effectiveness has doubled since TREC began, and most commercial retrieval systems, including many Web search engines, feature technology originally developed through TREC.

The fundamental goal of a retrieval system is to help its users find information contained in large stores of free text. Natural language is rich and complex, but researchers and authors easily express the same concept in widely different ways. Algorithms must be efficient in light of how much text must 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 what information must be retrieved.



Processing 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 a pool to produce relevance judgments, which can then be used to score the output of both the participant result sets and any subsequent results created through the same topic and document sets.

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 three components: a set of documents; a set of information need statements called "topics"; and relevance judgments, a mapping of which documents should be retrieved for which topics. The abstracted retrieval task is to rank 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 range of information-seeking tasks.

The original Cranfield experiments created a test collection of 1,400 documents and a set of 225 requests. Many retrieval experiments have been run in the years following the Cranfield tests (several other test collections were also 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, use the same evaluation measures, or compare results across systems to consolidate findings. The available test collections contained so few documents that operators of commercial retrieval systems were unconvinced that the techniques developed through test collections would scale to their much larger and growing document sets. Some experimenters even questioned whether test collections had outlived their usefulness.

In 1991, the National Institute of Standards and Technology (NIST, www.nist.gov) was asked by the Defense Advanced Research Projects Agency

(DARPA) to build a large test collection for use in evaluating text-retrieval technology developed as part of its 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 the use of test collections; DARPA agreed, and TREC was born (see the figure here).

The first two TREC conferences included two tasks, or "tracks": ad hoc and routing. An 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 to be investigated. A routing task assumes the topics are static but need to be matched to a stream of new documents. This retrieval technology is used by, for example, news clipping services and information analysts monitoring a data source. Starting with TREC 3 in 1994, additional tracks were included in the conference. They serve several purposes. First, they act as incubators for new research areas; the first running of a track often defines what the problem really is and creates the necessary infrastructure (such as test collections and evaluation methodology) to support research on its task. They also demonstrate that 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. TREC workshops now include six or seven separate tracks. The figure is a schematic of the processing performed in a typical TREC track. Track organizers provide a set of "documents" and a set of topics whose information needs can be met through the 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 all been used as documents in TREC tracks. Information needs have been mined from logs of existing commercial search systems or created specially for the task. Participants use their systems 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 ones 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 at NIST where participants 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 availability of the data, but the standard ad hoc collections generally contain from 800,000 to 1 million documents and 50 topics. Having human judges review all documents for all topics is 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 particular 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. TREC uses a process called "pooling" [5] 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 at 100). Since retrieval systems are designed to rank the documents most likely to be relevant first, pools created in this manner contain a sufficient number of the relevant documents that retrieval systems can be compared fairly by assuming that 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. Reusability facilitates research by allowing a tight development cycle. Given a test collection, a researcher can quickly compare a variety of 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 document owners) to support the broader retrieval research community.

Evaluating a retrieval system's effectiveness can be done in a variety of ways; for example, trec_eval reports approximately 85 different numbers for a result set, but a relatively small set of measures has emerged as the standard by which retrieval effectiveness is characterized. These measures are derived in some way from precision and recall, where precision is the proportion of relevant documents that are retrieved, and recall is the relevant proportion of retrieved documents. For ranked retrieval, a cut-off level is needed to define the retrieved set over which precision or recall is computed; for example, a cutoff level of 10 defines the retrieved set as the top 10 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 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 motivation for TREC was a request to create a single large test collection for a classic ad hoc retrieval task, TREC has accomplished much more in its 15-year history. A variety of collections has been constructed, including 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, including the Japanese National Institute of Informatics Test Collection for IR Systems project (NTCIR, research.nii.ac.jp/ntcir/), the Cross Language Evaluation Forum (CLEF, www.clefcampaign.org/), and the Initiative for the Evaluation of XML Retrieval (INEX, inex.is.informatik.uniduisburg.de).

TREC has also validated the use of test collections as a research tool for ad hoc retrieval and 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 effect of experimental design decisions (such as number of topics used, size of observed difference in retrieval scores, and effect of changes in the documents considered relevant to a topic) [1]. Nonetheless, studies on the very latest collections built from millions of Web pages suggest that the pooling process depends on the documents set size so it cannot produce 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 by looking at TREC scores from year to year, since any difference is likely caused by the different test collections being used. An experiment conducted by the SMART retrieval group in TREC 1–8 demonstrated that retrieval effectiveness did indeed improve over that time. 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. They ran each system on each test collection. For each collection, the later versions of the SMART system were much more effective than the earlier versions, with the later scores approximately twice those of the earliest scores. While this experiment involved only the SMART system, SMART results consistently tracked with the other systems' results in each TREC. SMART results can therefore be considered representative of the field as a whole.

Conclusion

Almost 300 distinct groups representing more than 20 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 contribution to information retrieval research has been equally significant. A variety of large test collections has 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 learn from one another, promoting 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 extend the experimentation to new problems. C

References

1. Buckley, C. and Voorhees, E. Retrieval system evaluation. In *TREC: Experiment and Evaluation in Information Retrieval,* E. Voorhees and D. Harman, Eds. MIT Press, Cambridge, MA, 2005, 53–75.
2. Cleverdon, C. The Cranfield tests on index language devices. In *Readings in Information Retrieval,* K. Spärck Jones and P. Willett, Eds., Morgan Kaufmann, San Francisco, 1997.
3. Salton, G., Ed. *The SMART Retrieval System: Experiments in Automatic Document Processing.* Prentice-Hall, Inc., Englewood Cliffs, NJ, 1971.
4. Spärck Jones, K., Ed. *Information Retrieval Experiment.* Butterworths, London, 1981.
5. Spärck Jones, K. and van Rijsbergen, C. 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.
6. Voorhees, E. and Harman, D., Eds. *TREC: Experiment and Evaluation in Information Retrieval.* MIT Press, Cambridge, MA, 2005.

Ellen M. Voorhees (Ellen.Voorhees@nist.gov) is the manager of the Retrieval Group in the Information Access Division of the Information Technology Laboratory at the National Institute of Standards and Technology, Gaithersburg, MD.