Building Better Search Engines by Measuring Search Quality

Ellen M. Voorhees, Paul Over, Ian Soboroff

National Institute of Standards and Technology

Gaithersburg, MD 20899-8940, USA

Abstract

Search engines help users locate particular information within large stores of content developed for human consumption. For example, users expect web search engines to direct searchers to web sites based on the content of the site rather than the site address, and video search engines someday to return video clips based on the actions recorded in the clip rather than file names and donor tags. Search engines are developed using standard sets of realistic test cases that allow developers to measure the relative effectiveness of alternative approaches. The NIST Text REtrieval Conference (TREC) project has been instrumental in creating the necessary infrastructure to measure the quality of search results for more than twenty years, and has thus helped fuel the recent explosive growth in search-related technologies.

Keywords: information retrieval; effectiveness measurement; multimedia search

1 Origins of TREC

Today we take search for text documents in our native language for granted, but web search engines such as Yahoo, Google, and Bing were not built in a day, nor is web content the only area where we need search. As data has become more ubiquitous, search needs have correspondingly expanded. People search for a variety of reasons(e.g., to re-locate known data items, to answer specific questions, to become informed on a particular issue, to monitor a data stream, to browse) across a variety of media (e.g., text, web pages, Tweets, speech recordings, still images, video). In many cases the technology to support these varied types of searches is still maturing. How is progress made in search technology? How do search engine developers know what works and why? Careful measurement of

Digital Object Indentifier 10.1109/MITP.2013.105 1520-9202/$26.00 2013 IEEE

search engine performance on standard, realistic tests with participation from a large, diverse research community has proved to be key, and through its Text REtrieval Conference (TREC) project NIST has been instrumental in assembling community evaluations to spur progress in search and search-related technologies over the last quarter century. Search algorithms are generally developed by comparing alternative approaches on benchmark tasks called test collections. The first test collection resulted from a series of experiments regarding indexing languages at the Cranfield College of Aeronautics in the1960s [1]. The Cranfield test collection consists of a set of abstracts from journal articles on aeronautics, a set of queries against those abstracts, and an "answer key" of correct responses for each query. Though minuscule by today's standards, the Cranfield collection broke new ground by creating the first shared measurement tool for information retrieval systems: researchers could write their own search engines to retrieve abstracts in response to the queries, and those responses could be measured by comparing against the answer key.

Other research groups began to follow the experimental methodology the Cranfield tests had introduced, producing several other test collections that were used in the 1970's and 1980's. 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 search systems to consolidate findings. Commercial search engine companies did not incorporate findings from the research systems into their products because they believed the test collections in use by the research community were too small to be of interest.

Amidst this discontent, 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 [2]. NIST agreed to undertake the construction of a large test collection using a workshop format that would also support examination of the larger issues surrounding test collection use. This workshop, the first TREC meeting, was held in 1992, and there has been a TREC meeting every year since. TREC accomplished the original goal of building a large test collection early on; indeed, it has now built dozens of test collections that are in wide-spread use throughout the international research community. TREC's greater accomplishment has been the establishment and validation of a research paradigm that continues to be extended to new tasks and application contexts every year.

2 Community Evaluations

The research paradigm is centered on community-based evaluations that have come to be called "coopetitions", borrowing the neologism that reflects cooperation among competitors that leads to a greater good. The main element of the paradigm is the evaluation task; this task is generally an abstraction of a set of user tasks that defines exactly what a system is expected to do. Associated with the evaluation task is one or more metrics that reflect the quality of a system's response and a means by which any infrastructure necessary to compute those metrics can be constructed. An evaluation methodology encompasses the task, the metrics, and a statement of the valid interpretations of the metrics' scores. A standard evaluation methodology allows results to be compared across system -- important not so there can be winners of retrieval competitions, but because it facilitates the consolidation of a wider variety of results than any one research group can tackle.

Digital Object Indentifier 10.1109/MITP.2013.105 1520-9202/$26.00 2013 IEEE

As a concrete example of the paradigm, consider the main "ad hoc" task in the first TRECs, which extended the Cranfield methodology that existed at the time. The ad hoc evaluation task is to retrieve relevant documents (or, more specifically, to create a list of documents such that relevant documents appear in the list before non-relevant documents) given a document set and a natural language statement of an information need, called a topic. Such retrieval output can be scored using precision (the fraction of retrieved documents that are relevant) and recall (the fraction of relevant documents that are retrieved) provided the set of relevant documents for each topic (i.e., the 'answer key') is known. TREC's innovation was to use pooling [3] to build the relevance sets for large document sets.

A pool is the union of the top X documents retrieved by each of the participating systems' searches for a given topic. Only the documents in a topic's pool are judged for relevance by a human assessor, with all other documents assumed to be not relevant when computing effectiveness scores. Subsequent testing verified that pooling as implemented in TREC finds a large majority of the relevant documents in a document set despite looking at only a tiny fraction of the whole collection, and further validated that retrieval systems that get higher scores on test collections built through pooling are generally more effective retrieval systems in practice than those that get lower scores [4]. This testing also revealed the limited valid uses of scores computed on test collections. Because the absolute value of scores depend on factors other than the retrieval system (for example, using different human judges will generally lead to somewhat different scores), it is only valid to compare scores computed from a test collection to scores computed for other systems on that exact same test collection. In particular, this means that it is not valid to compare the scores from different years in TREC, because each TREC built a new (different) test collection.

It is necessary to have a wide diversity of retrieval approaches contributing to the pools for pooling to be an effective strategy. Thus the community aspect of TREC -- many different retrieval approaches retrieving diverse document sets -- is critical to building good test collections. The community aspect is important to TREC's success in other respects, as well. TREC can benchmark the state-of-the-art only if all retrieval approaches are represented. The annual TREC meeting facilitates technology transfer among different research groups as well as between research and development organizations. The annual meeting also provides an efficient mechanism for methodological questions to be resolved. Finally, community members are frequently a source of data and use cases for new tasks.

When TREC began there was real doubt as to whether the statistical systems that had been developed in the research labs (as opposed to the operational systems that used Boolean searches on manually indexed collections) could effectively retrieve documents from "large" collections. The ad hoc task in TREC has shown not only that the retrieval engines of the early 1990's did scale to large collections, but that those engines have improved since then. This effectiveness has been demonstrated both in the laboratory on TREC test collections and by today's operational systems that incorporate the techniques. Further, the techniques are routinely used on collections far larger than what was considered large in 1992. Web search engines are a prime example of the power of the statistical techniques: the ability of search engines to point users to the information they seek has been fundamental to the success of the web.

As noted above, improvement in retrieval effectiveness cannot be determined simply by looking at TREC scores from year to year. However, 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 [5]. 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 that of the earlier scores. While this is evidence for only one system, the SMART

Digital Object Indentifier 10.1109/MITP.2013.105 1520-9202/$26.00 2013 IEEE

system results consistently tracked with the other systems' results in each TREC, and thus the SMART results can be considered representative of the state-of-the art.

3 Branching Out

While the initial intent for TREC was simply to build one or two large test collections for ad hoc retrieval and explore methodological questions related to pooling, it soon became obvious that the ad hoc task could be tweaked along several different dimensions. Each task that resulted from a tweak was related to the classic task, but was sufficiently different in some regard to require changes in the evaluation methodology. TREC therefore introduced a track structure whereby a given TREC contained several different retrieval subtasks that were each the focus of its own evaluation challenge. Figure 1 shows (most of) the tracks that were run in the different years of TREC, grouping the tracks by the dimension that differentiates them from one another. The dimensions, listed on the left of the figure show the breadth of the problems that TREC has addressed, while the individual tracks listed on the right show the progression of tasks within the given problem area. Today, each TREC contains seven or eight tracks that change frequently to keep TREC fresh and to support new communities. Several of the TREC tracks have been the first large-scale evaluations in that area. In these cases, the track has established a research community and has created the first specialized test collections to support the research area. A few times, the track has spun-off from TREC and a community of interest established its own evaluation conference. For example, CLEF (www.clef-initiative.eu/) spun-off from TREC in 2000 to expand the evaluation of cross language retrieval in Europe and has since broadened to encompass not only multilingual but multimodal (text, image, video) information. Other conferences such as NTCIR (research.nii.ac.jp/ntcir/) focusing especially on Chinese, Japanese, and Korean language texts; Initiative for the Evaluation of XML Retrieval (INEX, inex.mmci.uni-saarland.de/); and Forum for Information Retrieval Evaluation (FIRE, www.isical.ac.in/~clia/ focusing especially on languages of the Indian subcontinent were not direct spin-offs from TREC, but were inspired by TREC and extend the methodology to still other areas.

Digital Object Indentifier 10.1109/MITP.2013.105 1520-9202/$26.00 2013 IEEE

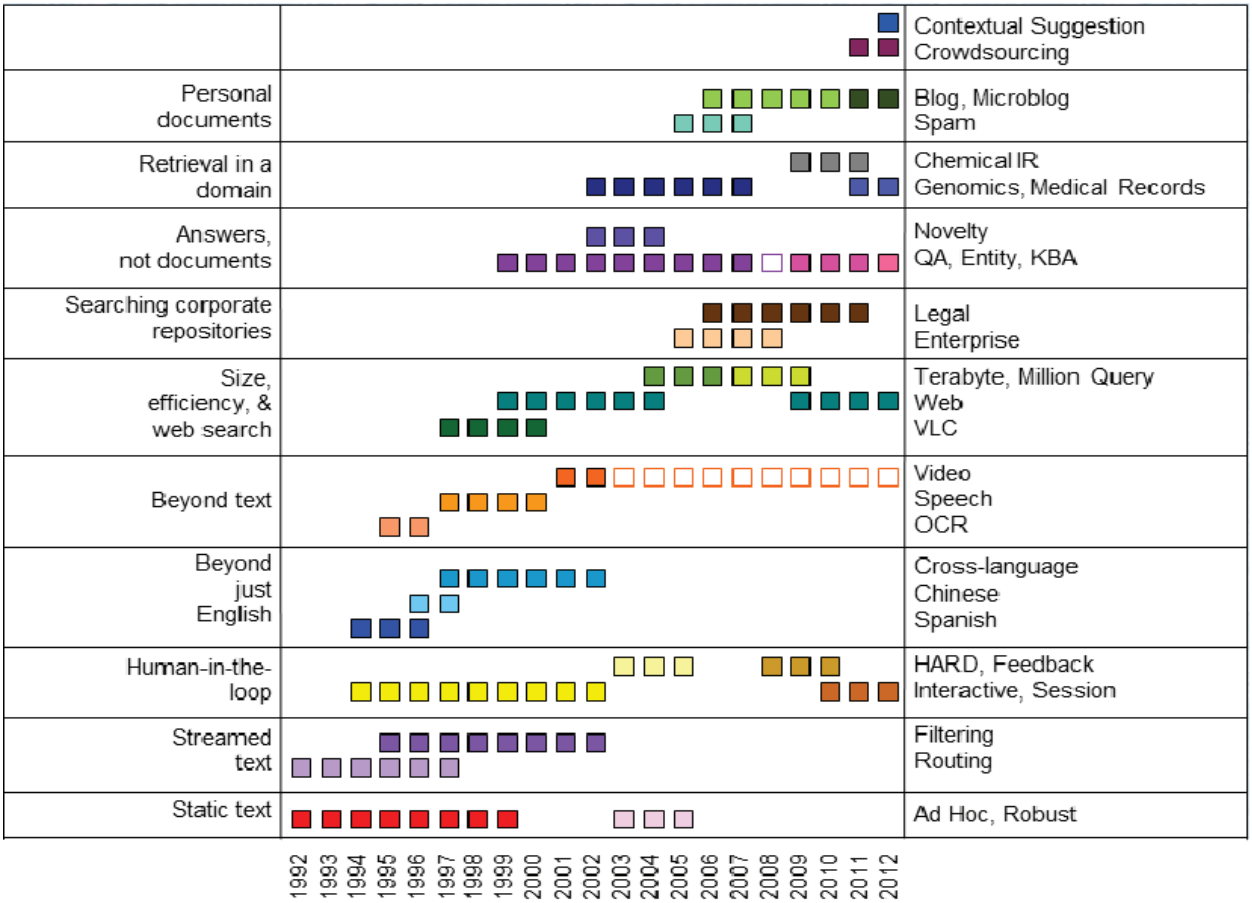

Figure 1: TREC tracks and the years in which they were held. The track name is given on the right while the distinguishing focus of the track is given on the left.

Space limitations prohibit even a cursory discussion of each TREC track, but a sampling of tracks are highlighted in this section. These include the filtering, question answering, and legal e-discovery tracks that have addressed particularly pressing search problems. Also included is the video retrieval track, which, driven by the growing availability of digital video, has grown into its own NIST workshop series: TRECVID.

3.1 Filtering

The ad hoc task and a routing task were the only two tasks in the first years of TREC. The routing task was designed to simulate a user monitoring a stream of documents, selecting relevant documents and ignoring unwanted ones. In TREC-4, routing evolved into filtering, a harder, but more realistic scenario. Just as an email filtering system processes an incoming stream of email in real time to remove spam and apply filing rules, an information filtering system processes an incoming stream of documents and decides whether to deliver them to the user according to a profile that models the user's interests based on feedback regarding documents that were already delivered. [6]

Digital Object Indentifier 10.1109/MITP.2013.105 1520-9202/$26.00 2013 IEEE

Whereas the routing evaluation task allowed systems to process all documents in the collection in a batch fashion, the filtering evaluation task requires a system to process the documents as they arrive in a stream, and to adapt its user model on-line. If the system chooses to show the document to the user, and there exists a relevance judgment for that document, then the system is given that judgment (simulating real-time user feedback), and can adapt itself immediately based on that information. If the system decides not to show the document, any relevance information is missed. The effectiveness of the filtering system is measured by its utility: utility increases by finding relevant documents and falls with non-relevant ones. A fundamental result of the filtering track was the understanding of just how hard the task actually is. Under the utility model, a system is penalized for returning non-relevant information. In the filtering track collections, as in real life, there tend to be only a small number of relevant documents in a stream of millions of documents. Therefore, a prudent system can actually score quite well by never returning any documents, in a sense deciding not to run the risk of wasting the user's time. Because the system is only given a little training data at the beginning of the stream, its initial performance tends to be poor -- as it performs initial refinement of its user model, it shows many promising documents to the user but many of these documents are probably not relevant! To perform well, the system has to be able to recover the initial expense of that feedback by performing extremely well very quickly.

3.2 Question Answering

While a list of on-topic documents is undoubtedly useful, even that can be more information than a user wants to examine. The TREC question answering track was introduced in 1999 to focus attention on the problem of returning exactly the answer in response to a question. The initial question answering tracks focused on factoid questions, questions with short, fact-based answers such as "Where is the Taj Mahal?". Later tracks incorporated more difficult question types such as list questions (a question whose answer is a distinct set of instances of the type requested such as "What actors have played Tevye in 'Fiddler on the Roof?'") and definitional or biographical questions (such as "What is a golden parachute?" or "Who is Vlad the Impaler?"). The question answering track was the first large-scale evaluation of open-domain question answering systems, and it has brought the benefits of test collection evaluation observed in other parts of TREC to bear on the question answering task. The track established a common task for the retrieval and natural language processing research communities, creating a renaissance in question answering research. This wave of research has created significant progress in automatic natural language understanding as researchers incorporated sophisticated language processing into their question answering systems. For example, the Jeopardy!-playing computer system 'Watson' had its origins in IBM's participation in the TREC question answering track [7].

3.3 E-discovery

The legal track was started in 2006 to focus specifically on the problem of e-discovery, the effective production of electronically stored information as evidence in litigation and regulatory settings. Today's organizations depend on electronic records rather than paper records, but the volume of data and

Digital Object Indentifier 10.1109/MITP.2013.105 1520-9202/$26.00 2013 IEEE

potential ephemeral nature of it has overwhelmed traditional legal discovery procedures and practices. New discovery practices targeted for electronic data are required.

When the track began, it was common for the two sides involved in litigation to negotiate a Boolean expression that defined the discovery result set and have humans examine each document so retrieved to determine its responsiveness to the discovery request. The goal of the track was to evaluate the effectiveness of this baseline approach and other search technologies for discovery. The track used hypothetical complaints and corresponding requests to produce documents developed by practicing lawyers as topics. A designated "topic authority" played the role of the lead attorney in a case who set forth general strategy and guidelines for what made a document responsive to the request. Relevance determinations for specific documents were made by legal professionals who followed their typical work practices in reviewing the documents.

The track had a major impact in the legal community, including being cited in judicial opinions (see en.wikipedia.org/wiki/Paul_W._Grimm). Its main result was engendering conversation on the process by which e-discovery should be done by showing that an iterative process that included a human in the search loop almost always performed much better than one-off searches. On the information retrieval side, the track was important because it demonstrated deficiencies in the standard test collection evaluation methodology. To facilitate stable evaluations, especially when using test collections built from pooling, the standard methodology relies on average effectiveness over a set of topics where each topic has relatively few relevant documents. But the real use case in e-discovery is the need for gauging the effectiveness of a single response sets where the number of responsive documents can be very large.

Digital Object Indentifier 10.1109/MITP.2013.105 1520-9202/$26.00 2013 IEEE

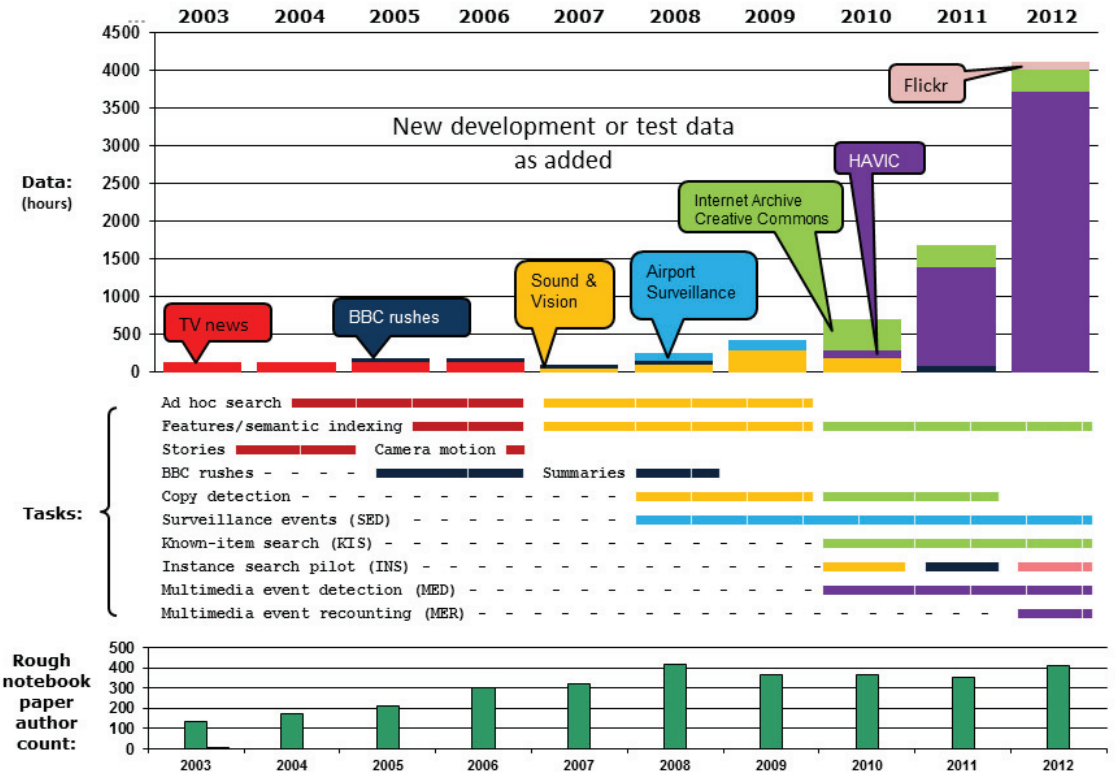

Figure 2: TRECVID evolution

3.4 Video

Beyond the TREC workshops but still at NIST, TRECVID has evolved (Figure 2) in many ways since its inception as a TREC 2001 track created to extend the TREC/Cranfield philosophy to content- based video analysis and retrieval. After two years, it became an independent workshop series and began a 4-year cycle using TV broadcast news (in English, Chinese, and Arabic), tripling the test data from 50 to 150 hours. System tasks included search using multimedia topics, high-level feature extraction, shot and story boundary determination, and camera motion detection. In 2007 a 3-year cycle began, using educational and cultural programming from the Netherlands Institute for Sound and Vision. Test data increased to 280 hours by 2009. A summarization task was added against BBC rushes (unedited program material) as well as an event detection task against airport surveillance video provided by the UK Home Office. Since 2010, TRECVID has focused on diverse, often non-professional Internet video from various community-donated sources in quantities from several hundred up to several thousand hours, extending the search and feature/event detection tasks while adding known-item and instance search to the evaluations.

Digital Object Indentifier 10.1109/MITP.2013.105 1520-9202/$26.00 2013 IEEE

TRECVID researchers have made significant contributions to the state of the art in the judgment of their scientific peers worldwide. A 2009 bibliometric study by library scientists at Dublin City University found 310 (unrefereed) workshop papers had been produced by TRECVID participants between 2003 and 2009 as well as 2,073 peer-reviewed journal articles and conference papers based on the TRECVID data [8].

Although measuring system improvement is difficult when test data changes, experiments by the University of Amsterdam's MediaMill team in 2010 demonstrated a threefold improvement in feature detection over 3 years -- this for a system usually ranked among the top performers in TRECVID [9]. The copy detection test data was the same in 2010 and2011 while the test queries (11256) were randomly created. This allows comparison of systems. Top teams' average scores for both detection and localization were better in 2011 than in 2010.

The TRECVID workshop series has brought together a diverse community of self-funded researchers attracted by tasks that can motivate interesting work in variety of fields, by lower entry thresholds, and by an open forum for scientific comparison. The number of groups worldwide that can complete one of the tasks has grown and new top-performers continue to emerge. Increasing the intellectual attention to enduring problems such as extracting meaning from video can only increase the likelihood of progress in the long run.

In its first 3 years as an independent workshop series the TRECVID community grew rapidly, tripling applications from 20 to 60 groups, of which 40 completed at least one task. From 2007 to 2009 applications rose to around 100 with 60 teams finishing and this level of community involvement has continued into the present. A rough count of the workshop paper co-authors indicates about 400 researchers are engaged in each year's TRECVID experiments. Although academic teams have predominated, commercial research laboratories have always been part of the mix. Europe and Asia vie for the region with the most participants, with North America close behind.

The TRECVID community has contributed more than just research. They have donated essential parts of the evaluation infrastructure, including: ground truth annotation systems and judgments, shot segmentation, automatic speech recognition, evaluation software, data hosting, trained detectors, etc. TRECVID would not be possible without this collaboration.

A 2009 review article in Foundations and Trends in Information Retrieval found the following:

> *Due to its widespread acceptance in the field, resulting in large participation of international teams from universities, research institutes, and corporate research labs, the TRECVID benchmark can be regarded as the de facto standard to evaluate performance of concept-based video retrieval research. Already the benchmark has made a huge impact on the video retrieval community, resulting in a large number of video retrieval systems and publications that report on the experiments performed within TRECVID,... p.272, [10].*

Innovations include the use of multimedia search topics, automatically determined shots as the basic unit of retrieval (allowing for efficient judging of system output), application of average precision as an effectiveness measure in video search and concept detection, adoption of cost-based measures for copy detection, a practical method for evaluating rushes summarization, etc.

Technology transfer occurs across research teams within TRECVID and the wider video analytics community. Approaches that work for one system in one year's task are commonly adopted with

Digital Object Indentifier 10.1109/MITP.2013.105 1520-9202/$26.00 2013 IEEE

variations by other systems in the next year's work. As a laboratory exercise with prototype systems, TRECVID results tend to be indicative rather than conclusive. Credible evidence for particular approaches grows gradually as algorithms prove themselves repeatedly as part of various systems and against changing test data. Significant amounts of engineering and in some cases usability testing are required to make laboratories successes available in real world applications.

The Netherlands Institute for Sound and Vision, a major data and use case donor to TRECVID, has documented the role TRECVID has played in allowing them engage a wide community of researchers at low cost to explore tasks of interest to them on their own data. Promising techniques have then been further explored in closer collaboration with a nearby TRECVID participant (University of Amsterdam) to do the engineering and user testing needed to move from prototype to operational system [11].

One specific case of the transition to real world use is the development and licensing of feature/concept detectors to a company in the Netherlands, which will integrate them into software tools for police searching confiscated video for illicit material [12]

4 Moving Forward

TREC's approach of evaluating competing technologies on a common problem set has proved to be a powerful way to improve the state of the art and hasten technology transfer. Hal Varian, Chief Economist for Google, described TREC's impact in a 2008 post on the Google blog [13]:

> *The TREC data revitalized research on information retrieval. Having a standard, widely available, and carefully constructed set of data laid the groundwork for further innovation in this field. The yearly TREC conference fostered collaboration, innovation, and a measured dose of competition (and bragging rights) that led to better information retrieval.*

A more detailed study of the impact of TREC was undertaken by RTI International on commission from NIST [14]. In quantitative terms, the study estimated that the return on investment for every dollar spent on TREC was three to five dollars of benefits that accrued to information retrieval (IR) researchers. The study also enumerated a variety of qualitative benefits, concluding in part:

> *TREC's activities also had other benefits that were not quantified in economic terms. TREC helped educate graduate and undergraduate students, some who went on to lead IR companies and others who stayed in academia to teach and conduct research. TREC benefited IR product quality and availability -- our research suggests that TREC motivated a large expansion in IR research that has enabled high quality applications such as web search, enterprise search, and domain-specific search products and services (e.g., for genomic analysis). More specifically, this study estimates that TREC's existence was responsible for approximately one-third of an improvement of more than 200% in web search products that was observed between 1999 and 2009.*

Digital Object Indentifier 10.1109/MITP.2013.105 1520-9202/$26.00 2013 IEEE

Despite this success, much remains to be done. Computers are still unable to truly comprehend content generated for human consumption even while content stores continue to grow ever larger. The TREC and TRECVID workshops will continue for the foreseeable future, focusing retrieval research on problems that have significant impact for both the retrieval research community and the broader user community.

The TREC web sites, http://trec.nist.gov and http://trecvid.nist.gov, contain a wealth of information about TREC including the full proceedings of each workshop and details regarding how to obtain the test collections. Organizations can participate in TREC by responding to the Call for Participation that is issued each winter.

Disclaimer: Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

5 References

[1] 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).

[2] Donna Harman. The DARPA TIPSTER project. ACM SIGIR Forum, 26(2):26-28, 1992.

[3] 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.

[4] 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.

[5] Chris Buckley and Janet Walz. SMART at TREC-8. In Proceedings of the Eighth Text REtrieval Conference (TREC-8), pages 577-582, 1999. NIST Special Publication 500-246 .

[6] Stephen Robertson and Jamie Callan. Routing and filtering. In Ellen M. Voorhees and Donna K. Harman, editors, TREC: Experiment and Evaluation in Information Retrieval, chapter 5, pages 99-122. MIT Press, 2005.

[7] David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty. Building Watson: An overview of the DeepQA project. AI Magazine, pages 59-79, Fall 2010.

[8] Clare V. Thornley, Andrea C. Johnson, Alan F. Smeaton, and Hyowon Lee. The scholarly impact of TRECVID (2003-2009). Journal of the American Society of Information Science and Technology, 62(4):613-627, April 2011.

Digital Object Indentifier 10.1109/MITP.2013.105 1520-9202/$26.00 2013 IEEE

[9] Cees G. M. Snoek, Koen E. A. van de Sande, Dennis C. Koelma, and Arnold W. M. Smeulders. Any Hope for Cross-Domain Concept Detection in Internet Video - http: //www-nlpir.nist.gov/projects/tvpubs/tv10.slides/mediamill.tv10.slides.pdf, 2010.

[10] Cees G. M. Snoek and Marcel Worring. Concept-based video retrieval. Found. Trends Inf. Retr., 2(4):215-322, April 2009.

[11] Johan Oomen, Paul Over, Wessel Kraaij, and Alan F. Smeaton. Symbiosis between the TRECVID benchmark and video libraries at the Netherlands Institute for Sound and Vision. International Journal on Digital Libraries, 13(2):91-104, 2013.

[12] Paul Over. Instance Search, Copy Detection, Semantic Indexing @ TRECVID. http://www.nist.gov/oles/upload/8-Over_ Paul-TRECVID.pdf, November 2012.

[13] Hal Varian. Why data matters. Google Official Blog, March 4, 2008.

[14] RTI International. Economic impact assessment of NIST's Text REtrieval Conference (TREC) program. http://www.nist.gov/director/ planning/impact_assessment.cfm, 2010.

6 Short author bios

Ellen Voorhees is a computer scientist at the National Institute of Standards and Technology where her primary responsibility is to manage the TREC project. Her research focuses on developing and validating appropriate evaluation schemes to measure system effectiveness for diverse user search tasks and for natural language processing tasks.

Paul Over is a computer scientist at the US National Institute of Standards and Technology and founding project leader for the TREC Video Retrieval Evaluations (TRECVID). Paul has also been responsible at NIST for evaluation of interactive text retrieval systems within the TExt REtrieval Evaluations (TREC) and has supported natural language processing researchers in evaluation of text summarization technology. Before joining NIST he worked for a dozen years in software development at IBM, including the application of his interests and training in linguistics and computer science to advanced natural language product development.

Ian Soboroff is a computer scientist and manager of the Retrieval Group at the National Institute of Standards and Technology (NIST). He has co-authored many publications in information retrieval evaluation, test collection building, text filtering, collaborative filtering, and intelligent software agents. His current research interests include building test collections for social media environments and nontraditional retrieval tasks.

7 Author contact information

Ellen Voorhees
NIST

Digital Object Indentifier 10.1109/MITP.2013.105 1520-9202/$26.00 2013 IEEE

100 Bureau Drive, STOP 8940
Gaithersburg, MD 20899-8940, USA
ellen.voorhees@nist.gov
V: 301 975 6731
F: 301 975 5287

Paul Over
NIST
100 Bureau Drive, STOP 8940
Gaithersburg, MD 20899-8940, USA
over@nist.gov
V: 301 975 6784
F: 301 975 5287

Ian Soboroff
NIST
100 Bureau Drive, STOP 8940
Gaithersburg, MD 20899-8940, USA
Ian.soboroff@nist.gov
V: 301 975 3987
F: 301 975 5287

Digital Object Indentifier 10.1109/MITP.2013.105 1520-9202/$26.00 2013 IEEE