The Evolution of Cranfield

Ellen M. Voorhees

Abstract Evaluating search system effectiveness is a foundational hallmark of information retrieval research. Doing so requires infrastructure appropriate for the task at hand, which generally follows the Cranfield paradigm: test collections and associated evaluation measures. A primary purpose of *Information Retrieval (IR)* evaluation campaigns such as *Text REtrieval Conference (TREC)* and *Conference and Labs of the Evaluation Forum (CLEF)* is to build this infrastructure. The first TREC collections targeted the same task as the original Cranfield tests and used measures that were familiar to test collection users of the time. But as evaluation tasks have multiplied and diversified, test collection construction techniques and evaluation measure definitions have also been forced to evolve. This chapter examines how the Cranfield paradigm has been adapted to meet the changing requirements for search systems enabling it to continue to support a vibrant research community.

1 Introduction

Information retrieval research has a rich tradition of experimentation. In the 1960s, Cyril Cleverdon and his colleagues at the College of Aeronautics, Cranfield, ran a series of tests to determine appropriate indexing languages—schemes to represent document content that would enable trained search intermediaries to find appropriate references for library patrons (Cleverdon 1967). The conclusion reached in the experiments, that a document's own words are effective for indexing, was highly controversial at the time though generally accepted today. The experiments are best known, however, for being the first to use a test collection. By comparing the effectiveness of different languages on a common document set with a common set

E. M. Voorhees (✉)
National Institute of Standards and Technology, Gaithersburg, MD, USA
e-mail: ellen.voorhees@nist.gov

N. Ferro, C. Peters (eds.), *Information Retrieval Evaluation in a Changing World*, The Information Retrieval Series 41,
https://doi.org/10.1007/978-3-030-22948-1_2

of information needs, Cleverdon was able to control for much of the variability that had plagued earlier attempts to compare languages and in the process established what has become known as the Cranfield paradigm.

Test collections have been used in information retrieval research since then, though they have also had their detractors from the start (Taube 1965; Cuadra and Katter 1967). Early detractors feared the fluidity of 'relevance' made it unsuitable to be a fundamental component of an evaluation strategy. Later, concerns arose over the unrealistically small size of test collections compared to the data sets of operational systems, as well as the use of incompatible evaluation measures by different research groups that prevented their published retrieval results from being truly comparable. By the early 1990s even some Cranfield practitioners were questioning whether test collections had out-lived their usefulness (Robertson and Hancock-Beaulieu 1992).

In response to these concerns, in 1992 the U.S. *National Institute of Standards and Technology (NIST)* founded the TREC workshop with the goal of building a single realistically-large test collection to support IR research. TREC not only accomplished that goal, but along with its companion evaluation conferences such as CLEF, *NII Testbeds and Community for Information access Research (NTCIR)*, and *Forum for Information Retrieval Evaluation (FIRE)* that followed, went on to build dozens of large test collections for a variety of tasks. In addition, the conferences have standardized and validated best practices in the use of test collections.

This chapter examines how the Cranfield paradigm has been adapted to meet the changing requirements for information retrieval research in the era of community evaluation conferences. The next section gives a short recap of Cranfield before TREC to give context; see Robertson (2008) and Sanderson (2010) for more detailed accounts of the history of IR evaluation in this period. The following section describes research surrounding the early TREC collections that both enabled larger collections to be built and validated the existing experimental protocol. Section 4 then examines some of the ways the Cranfield paradigm has been extended in support of research in other areas and modern IR research problems.

2 Cranfield Pre-TREC

The Cranfield paradigm can be summarized as follows. A test collection consists of a set of documents, a set of information need statements (called "topics" in the remainder of this chapter), and a set of relevance judgments that list which documents should be returned for which topic. A researcher runs a retrieval system on a test collection to produce a ranked list of documents in response to each topic (this is a "run"). A ranking reflects the system's idea of which documents are likely to be relevant to the topic; documents it believes more likely to be relevant are ranked ahead of documents it believes are less likely to be relevant. Using the relevance judgments, some evaluation metric is computed for the ranked list for each topic and scores for individual topics are averaged over the set of topics in the

test collection. Different systems produce runs for the exact same test collection, and the average scores are compared. Retrieval systems producing runs with better average scores are considered more effective retrieval systems.

Cranfield is an intentionally stark abstraction of any real user search task, representing what Spärck Jones (2001) calls the core competency of search. The abstraction arises through three major simplifying assumptions:

- relevance can be approximated by topical similarity, which implies all relevant documents are equally desirable; relevance of one document is independent of the relevance of any other document; and the user information need is static.
- a single set of judgments for a topic is representative of the user population.
- (essentially) all relevant documents for a topic are known.

While these assumptions are not true, the abstraction represents a fundamental capability that any actual search system must possess. It is hard to imagine how a search system could be effective if it cannot at least distinguish relevant documents from not relevant ones.

Use of the Cranfield methodology also requires an evaluation measure. Early work in information retrieval, including the Cranfield tests, produced retrieved sets of documents (as opposed to ranked lists), and measured the effectiveness of retrieval in terms of the *precision* and *recall* of the set (Keen 1966). Precision is the fraction of retrieved documents that are relevant, and recall the fraction of relevant documents that are retrieved. In practice, the two measures vary inversely with one another, so both measures were needed to get an accurate view of the quality of the system. The advent of using ranked output required adapting the measurement methodology to define the retrieved set. This was generally done by defining a cut-off level, the rank such that everything in the list at or before that rank was considered retrieved and everything after it not retrieved. Alternatively, precision could be reported for a standard set of recall values, for example precision at 0.25, 0.5, and 0.75 recall. The use of standard recall values requires interpolation since the actual recall values that are obtainable for a given topic depends on the number of relevant documents the topic has, and there are various methods by which the interpolation could be performed. Averaging the results over a set of topics also has different options. Using precision as an example, one can compute the precision on a per-topic basis and then take the mean over the set of topics, or one can divide the total number of relevant documents retrieved for all topics by the total number of documents retrieved for all topics. Averaging schemes, and especially interpolation schemes, were the subject of much debate in these early years.

During the 25 years following the Cranfield experiments, other retrieval test collections were built and these collections were often shared among different research groups. But there was no agreement on which measures to use and measures proliferated. Research papers of the time generally reported only the authors' own favorite measures, and even when two papers reported values for what was called the same measure the actual implementations of that measure differed (interpolation differences), leading to incomparable results. Research groups could not build on one another's work because there was no common basis to do so.

The test collections in use by the research community were also small in comparison to the document set sizes used in commercial retrieval systems. Commercial retrieval systems were searching document sets that were orders of magnitude larger than the publicly available test sets of the time, and it was believed that operators of the commercial systems would continue to discount research results unless those results were demonstrated on comparably-sized collections (Ledwith 1992). Cleverdon had made the deliberate decision to use small document set sizes so that all documents could be judged for all topics, known as complete judgments (Cleverdon 1991). Unfortunately, even for document set sizes of several thousand documents, getting complete judgments is an arduous task; getting them for very much larger collections is out of the question.

In the latter half of the 1970s, Karen Spärck Jones and colleagues argued the need for and proposed how to build a large (30,000 documents!), general-purpose test collection that they called the 'IDEAL' collection (Spärck Jones and Van Rijsbergen 1975; Spärck Jones and Bates 1977; Gilbert and Spärck Jones 1979). The desire for a general-purpose collection acknowledged the problem that the collections being shared had each been developed to test a specific hypothesis and were not necessarily appropriate for the different research questions being investigated in subsequent use. The proposal was wide-ranging, touching on many different aspects of test collection methodology, but in particular suggested using pooling to obtain essentially complete judgments without actually having all documents in the collection judged. The essential idea of pooling is to obtain a human judgment for just the union of the retrieved sets of many different searches for the same topic and assume all unjudged documents are not relevant. The IDEAL collection itself was never constructed, but pooling was used as the methodology to build the first TREC collections.

3 TREC Ad Hoc Collections

TREC was conceived as a way of supporting the IR research community by developing the infrastructure necessary to do IR research. It began in 1992 with the goals of creating a single large test collection and standardizing the evaluation measures used to compare test results (Voorhees and Harman 2005). It has accomplished those goals and much more, building scores of test collections for a wide range of search tasks and inspiring other test-collection-building efforts, such as CLEF, that have built yet more collections. Equally as important, the repository of runs collected by the various community evaluations has provided the data needed to empirically examine the soundness of the test collection methodology (Buckley and Voorhees 2005).

3.1 Size

The main task in the first 8 years of TREC was the ad hoc task that built eight ad hoc test collections. The ad hoc task is the prototypical Cranfield evaluation task in which systems return a ranked list of documents from a known, static document set for each of a set of previously-unseen topics. What set the TREC ad hoc collections apart from those that had been created earlier was the size of the document set. The collections contain 2–3 gigabytes of text and 500,000–1,000,000 documents. The documents are mostly newswire or newspaper articles, but also include other document types such as government publications to have a heterogeneous mix of subject matter, literary styles and document formats.

As mentioned above, the relevance judgments for these collections were created using pooling. A subset of the runs submitted to TREC in a given year was selected to be the set of judged runs. The number of judged runs was determined such that the total number of documents to be judged fit within the available budget and each TREC participant had an equal number of runs judged (assuming they submitted at least that number). For each judged run, the top X documents per topic were added to the topics' pools. Most frequently, X was set to 100. Human assessors then judged each document in the pools where a single assessor judged all the documents for a given topic.

The critical factor in pooling is that unjudged documents are assumed to be not relevant when computing traditional evaluation scores. This treatment is a direct result of the original premise of pooling: that by taking top-ranked documents from sufficiently many, diverse retrieval runs, the pool will contain the vast majority of the relevant documents in the document set. If this is true, then the resulting relevance judgment sets will be "essentially complete", and the evaluation scores computed using the judgments will be very close to the scores that would have been computed had complete judgments been available.

Various studies have examined the validity of pooling's premise in practice. Harman (1996) and Zobel (1998) independently showed that early TREC collections in fact had unjudged documents that would have been judged relevant had they been in the pools. But, importantly, the distribution of those "missing" relevant documents was highly skewed by topic (a topic that had lots of known relevant documents had more missing relevant), and roughly uniform across runs. Zobel demonstrated that these "approximately complete" judgments produced by pooling were sufficient to fairly compare retrieval runs. Using the leave-out-uniques test, he evaluated each run that contributed to the pools using both the official set of relevant documents published for that collection and the set of relevant documents produced by removing the relevant documents uniquely retrieved by the run being evaluated. The results showed that mean evaluation scores for the runs were only marginally impacted.

Of course, pooling still requires relevance judgments for the documents in the pools, and the reliance on subjective human relevance judgments was the major criticism of the Cranfield methodology from its beginning (Taube 1965; Cuadra and

Katter 1967; Harter 1996). In response to the criticism, Cleverdon examined the effect of using different assessors' judgments when comparing nineteen indexing methods using four independent sets of judgments. When the indexing methods were ranked by score, he found a few differences in the ranks of some methods when varying the judgment sets, but the correlation between the methods' rankings was always very high and the absolute difference in performance of the indexing methods was quite small (Cleverdon 1970). Both Lesk and Salton (1969) and Burgin (1992) also examined the effect of varying judgments on different indexing methods using different collections and found no differences in the relative performance of their respective methods. However, each of these studies was performed on small collections (fewer than 1300 documents) so topics had a correspondingly small number of relevant documents and absolute scores had limited possibility to change. To ensure that stability of retrieval results held for collections with much larger relevant sets, similar tests were repeated on two TREC collections (Voorhees 2000). Those tests included a variety of different conditions including judgments made by query authors vs. judgments by non-authors; judgments made by different non-authors; judgments made by a single judge vs. group judgments; and judgments made by different people in the same environment vs. judgments made in very different environments. In each of these conditions the absolute value of the effectiveness measure was affected by different judgments, but the relative performance of the retrieval runs was almost always the same.

A major result of these ad hoc experiments was the demonstration that the size of a test collection's document set does in fact matter. IR is challenging because of the large number of different ways the same concept can be expressed in natural language, and larger collections are generally more diverse. Further, small collections can have at most a small number of relevant documents for a topic while larger collections can have a much more variable number of relevant documents across topics. Both retrieval system effectiveness and the ability to evaluate retrieval system effectiveness are more challenging when there is greater variability.

3.2 Evaluation Measures

The second goal for the initial TREC was to standardize evaluation practice, particularly the evaluation measures used to report results. The measures used to evaluate the runs in TREC-1 were measures in common usage right before TREC began. These included precision and recall at various document cut-off levels, interpolated precision at recall points from 0.0 to 1.0 in steps of 0.1 (used to plot a recall-precision graph), and the 11-point average or 3-point average of interpolated precision (averaged over the same recall levels as the recall-precision graph, or at {0.25, 0.5, 0.75} recall, respectively). Two problems became immediately obvious, both related to the number of relevant documents per topic in large collections. First, because different topics had very different numbers of relevant documents, measures based on a constant document cut-off level averaged very poorly. Precision at 30

documents retrieved represents very different retrieval performance for a topic that has only 5 relevant documents as compared to a topic that has 878 (as one of the TREC-1 topics did). Whatever cut-off level is chosen will be appropriate for some topics but wildly inappropriate for others. The second problem was caused by the fact that TREC-1 runs were evaluated over only the top 200 ranks. When topics have more relevant documents than the size of the evaluated set, high recall levels are not obtainable, causing all but the smallest recall level values to be unstable. As an example, consider a topic with 499 relevant documents. A system that retrieves 149 relevant documents in the top 200 ranks never reaches 0.3 recall, so interpolated precision at all recall levels greater than 0.2 is 0.0 (using the interpolation method of trec-eval). However, if the system retrieves 150 relevant documents in the top 200 ranks, its interpolated precision score at 0.3 recall is 0.75 instead of 0.0.

Several changes to the evaluation were therefore instituted for TREC-2 that remained for the rest of the ad hoc tasks. An easy change was to increase the number of documents submitted for a run for a topic from 200 to 1000. Topics were also "narrowed" somewhat such that the target number of relevant documents would be no more than about 350. The idea was that the evaluated set should be at least three times as large as the number of relevant documents to avoid erratic behavior when measuring high recall levels. Another change was to introduce two new evaluation measures, R-precision and noninterpolated average precision. R-precision for a topic with R relevant documents is precision at rank R. Noninterpolated average precision, now generally just called average precision, is the mean of the precision at each relevant document over all relevant documents, using 0.0 as the precision of a relevant document not retrieved. When this value is averaged over all topics in a topic set, the result is known as *Mean Average Precision (MAP)*, and it became the single measure most often used in IR research to represent the overall effectiveness of a run.

3.3 Reliability Tests

Empirical investigation of the reliability of test collection experiments with respect to two aspects of the methodology—the effect of differences in opinions of relevance and the effect of using essentially complete rather than truly complete judgments sets—was summarized in Sect. 3.1. Here, reliability means that a researcher can have confidence that if an experiment shows that system A is better than system B, then system A will be better than system B in other equivalent environments with high probability. The investigations demonstrated that absolute scores of effectiveness measures change as conditions change, but relative scores are highly consistent. These results underscore an important property of the Cranfield methodology, namely that the only valid use of evaluation scores computed on a test collection is to compare them to other scores computed on the exact same collection. This means, for example, that scores computed on CLEF collections

from two different years are *not* directly comparable, nor are scores computed on a collection and a subset of it.

The collection of runs submitted to various evaluation tasks enabled other empirical studies that help validate the reliability of the Cranfield methodology. One such study examined the size of the collection with respect to the number of topics it contains (Voorhees and Buckley 2002). Another examined the stability of different evaluation measures (Buckley and Voorhees 2000; Sakai 2006). The results of these studies are summarized here.

3.3.1 Effect of Topic Set Size

Retrieval system effectiveness has been reported as an average over topics since the first Cranfield experiments because retrieval system performance is known to vary widely depending on the topic. An analysis of variance model fitted to the TREC-3 results demonstrated that the topic and system effects, as well as the interaction between topic and system, were all highly significant, with the topic effect the largest (Banks et al. 1999). What this means is that retrieval effectiveness depends on both which question is asked and which retrieval mechanism is used, but on average which question is asked has a bigger effect on effectiveness than the retrieval mechanism used. Further, different mechanisms work relatively better on different question types.

The set of topics in a test collection is assumed to be a random sample of the universe of possible questions, so there is always some chance that a comparison of two systems using any given test set will lead to the wrong conclusion. The probability of an error can be made arbitrarily small by using arbitrarily many topics, but there are practical limits to the number of topics that can be included in a test collection. While experienced researchers knew that a sufficient number of topics was needed so average scores would be stable, there was little concrete evidence to suggest what was sufficient. The design study for the IDEAL collection posited that fewer than 75 topics would not be useful (Spärck Jones and Van Rijsbergen 1975). TREC organizers, who had to balance cost of topic development and relevance judgments against the quality of the collection, chose 50 topics as the default size for the TREC ad hoc collections.

Voorhees and Buckley (2002) used TREC results to empirically derive collection error rates. An error rate is defined as the likelihood of reaching a wrong conclusion from a single comparison as a function of the number of topics used in the comparison and the size of the difference of the evaluation scores (called Δ). Once established, the error rates were used to derive the minimum difference in scores required for a certain level of confidence in the results given the number of topics used in the comparison.

The core of the procedure used to estimate the error rates was comparing the effectiveness of a pair of runs on two disjoint topic sets of equal size to see if the two sets disagreed as to which of the runs is better. The comparisons were repeated for many different pairs of runs and many different topic sets. The error rate is defined

as the percentage of times that the two topic sets disagreed as to which is the better system. Since TREC runs contain 50 topics, this procedure was used to directly compute error rates for topic set sizes up to 25. Curves of the form ErrorRate $= A_1 e^{-A_2 S}$ where S is the size of the topic set were fit to the observed error rates, and then those curves were used to extrapolate error rates for larger topic sets. A different curve was fit for each of a set of binned Δ values. As expected, error rates are larger for smaller Δ's and decrease as the number of topics increases.

Spärck Jones (1974) suggested the rule-of-thumb that differences in scores of 0.05 were noticeable and differences of 0.1 were material (for small collections and using measures other than MAP). For MAP and topic set sizes of 25, the error rate computed over the TREC collections for a difference of 0.05 is approximately 13% on the TREC ad hoc collections. This means that if we knew nothing about systems A and B except their MAP scores which differed by 0.05, and if we repeated the experiment on 100 different sets of 25 topics, then on average we would expect 13 out of those 100 sets to favor one system while the remaining 87 would favor the other. The error rate for a difference of 0.1 with 25 topics is much smaller at approximately 2.5%. The error rates are also much smaller for sets of 50 topics, 3.7% and 0.15% respectively. For topic sets of 50 topics, a difference of 0.05 was the smallest Δ with an error rate less than 5%.

These differences in MAP scores used to compute the error rates are *absolute* differences, while much of the IR literature reports *percentage* differences. An absolute difference of 0.1 is a very substantial difference, especially given that the best MAP scores on the TREC ad hoc collections are approximately 0.3. The percentage difference between a run with a 0.3 MAP score and a run with a 0.10 absolute difference is approximately 33% and for a 0.05 absolute difference is approximately 15%. However, the computed error rates are also for a single comparison of two arbitrary runs. In practice, researchers will use multiple test collections to compare different techniques, and the techniques being compared will likely be variants of some common system. Comparisons of different instances of a common system will have less variability overall, so error rates will be smaller in this case. Using multiple test collections is sound experimental practice and will again increase the confidence in conclusions reached.

3.3.2 Effect of Evaluation Measure Used

The study of the effect of the topic set size summarized above showed that the reliability of experimental findings depends on (at least) three interrelated components of the Cranfield paradigm: the number of topics used, the evaluation measure used, and the difference in scores required to consider one method better than the other. The evaluation measure used makes a difference because measures have different inherent reliabilities. Buckley and Voorhees (2000) focused on quantifying these differences among measures using the TREC Query Track (Buckley 2001) data.

The Query Track data provides different expressions of the same underlying information need. That is, in TREC parlance, the track gathered different queries for

the same topic and ran different retrieval systems on each of the different queries. Each query provides a separate evaluation score for the corresponding topic, thus producing a set of scores for the exact same topic. While using different queries does affect retrieval behavior—some queries are clearly better expressions of the topic than others—the effect of the number of relevant documents on system behavior is controlled because it remains constant. Controlling this topic effect allows the error inherent in the evaluation measure itself to be isolated.

Call a *query set* a collection of 50 queries, one for each topic. Each of 21 query sets was run using nine different retrieval methods, producing a data set consisting of nine sets of the top 1000 documents retrieved for each of 1050 queries (21 versions of 50 topics).

As in the topic set size experiment, error rates for an evaluation measure are computed by comparing the scores obtained by different retrieval methods, but the particulars of how the error rate is defined differ. Buckley and Voorhees (2000) used the error rate calculation described here, while Sakai (2006) used a separate, more mathematically-principled definition, with both definitions leading to the same conclusions. The first approach counts the number of times each retrieval method was better than, worse than, and equal to each other retrieval method when compared over a given query sets, using many different permutations of queries assigned to query sets and considering scores within a given percentage difference (say, 5%) of one another to be equivalent. Assuming that the correct answer is given by the greater of the better-than and worse-than values, the lesser of those two values is the number of times a test result is in error. Hence the error rate is defined as the total number of errors across all method pairs divided by the total number of decisions. With this definition, the error rate can never be more than 50%, and random effects start dominating the calculation of the error rate if it exceeds approximately 25%. The number of times methods are deemed to be equivalent is also of interest because it reflects on the power of a measure to discriminate among systems. It is possible for a measure to have a low error rate simply because it rarely concludes that two methods are different. The proportion of ties, defined as the total number of equal-to counts across all method pairs divided by the total number of decisions, quantifies this effect.

The error rates for different measures were found to be markedly different. Measures that depend on a relatively few highly ranked documents, such as precision at small cut-off levels, have higher error rates than measures that incorporate more documents. For example, when using a fuzziness factor of 5%, Prec(10) and Prec(30) had error rates of 3.6% and 2.9% respectively, while MAP had an error rate of 1.5%. The proportion of ties for the various measures also differed substantially. Precision failed to distinguish between two systems almost a quarter of the time (24%) while MAP failed to distinguish about 13% of the time.

3.3.3 Significance Testing

The error rates computed in the two investigations described earlier in this section are different from statistical significance tests, but all acknowledge the same underlying truth of test-collection-based experiments: that there is a fair amount of noise in the process. Statistical significance tests are run on the results of a retrieval experiment to determine whether the observed variation in topic scores is consistent with chance fluctuations.

Statistical significance testing has been used in IR experiments for almost as long as test collections have existed (Lesk 1967), though their application in retrieval experiments has not been without controversy. Early critics were concerned that retrieval system output does not meet the distributional assumptions of parametric tests (Van Rijsbergen 1979). Proponents demonstrated that the test were robust to the types of violations seen in practice (Hull 1993; Smucker et al. 2007) or suggested non-parametric schemes such as the bootstrap method (Savoy 1997). More recent concerns have arisen because of the wide availability of test collections, especially collections with very many topics. The wide availability of test collections means that it is easy to run experiments: a wide variety of different techniques can all be compared to one another, but corrections for multiple comparisons are seldom used (Carterette 2012). Further, given that the field (re)uses the same collections, there are also sequential testing effects (Carterette 2015). Sakai (2016) provides a survey of current practices in significance testing in IR.

The final test of the validity of the Cranfield paradigm is whether the conclusions reached from the laboratory experiments transfer to operational settings. Hersh and his colleagues suggest that the results may not transfer since they were unable to verify the conclusions from a laboratory experiment in either of two user studies (Hersh et al. 2000; Turpin and Hersh 2001). However, their tests were small and the user studies did not show that the conclusions from the laboratory test were wrong, simply that the user studies could not detect any differences. Furthermore, using a different approach Al-Maskari et al. (2008) demonstrated that users were indeed able to discern and act on the differences found in systems whose test-collection-based scores were only slightly different. Even a cursory examination of retrieval technology actually in use today makes it clear that the results do transfer. Basic components of current web search engines and other commercial retrieval systems—including full text indexing, term weighting, and relevance feedback—were first developed on test collections.

4 Moving On

TREC was founded on the belief that the Cranfield paradigm of using test collections as laboratory tools to compare the effectiveness of different retrieval methods was fundamentally sound though in need of updating with regard to collection size and standardization of evaluation metrics. Research using subsequently constructed

collections and retrieval results confirmed this belief, as summarized above. Yet those findings apply to a fairly narrowly proscribed protocol that is not strictly applicable to much of IR research in the ensuing years.

This section looks at ways in which the Cranfield paradigm has been extended or modified to continue to support the IR research community. In keeping with the scope of the chapter, the section only focuses on evaluation protocols connected to some form of a test collection. Protocols for controlled experiments involving users of operational systems, including traditional interactive IR studies and newer online evaluation (e.g., A/B testing and reuse of data gleaned from query logs) experiments, have also evolved in the ensuing years, but are not discussed here. Kelly (2009) provides a comprehensive review of interactive IR and Hofmann et al. (2016) provides the same for online evaluation.

4.1 Cross-Language Test Collections

A cross-language ad hoc retrieval task was the inaugural task in CLEF and is also featured prominently in NTCIR and FIRE. As an ad hoc retrieval task, Cranfield is clearly an appropriate evaluation tool for it, but building a good cross-language test collection is much more difficult than building a monolingual collection.

When creating a cross-language collection, a topic will be created in an initial language, and then usually translated into some of the other languages of the document set (to facilitate multiple monolingual experiments or cross-language experiments with differing source languages, for example). The quality of this translation is very important: a too literal translation depresses retrieval results because the language use in the translated topic does not match how the concept is natively expressed in the documents (Mandl and Womser-Hacker 2003).

Even with good translations, a given topic is much more likely to pertain to some parts of the collection than others since cultural differences make some topics more apt to be discussed in some subset of languages. This complicates pooling for cross-language collections. The quality of a test collection depends on having diverse pools, yet it is very difficult to get equally large, diverse pools for all languages contained within a multilingual collection. Both the number of runs submitted by participants and the documents retrieved within a run are usually skewed in favor of some languages at the expense of others. As a result, the pools for the minority languages are smaller and less diverse than the pools for the majority languages, which introduces an unknown bias into the judgments. Ensuring an equal number of documents is judged per language is not a solution to this problem because of the inherent differences in the true number of relevant documents per language. One way that does help enhance the quality of the pools is for the collection builders to supplement pools built from participant runs with documents discovered through the builders' own manual searches, a technique used to good advantage for the early NTCIR collections (Kando et al. 1999).

Obtaining a consistent set of relevance judgments is also more difficult for cross-language collections. In monolingual collections, the judgments for a topic are produced by one assessor. While this assessor's judgments may differ from another assessor's judgments, the judgment set represents an internally consistent sample of judgments. Using a single individual to judge documents across the multiple different languages represented in a cross-language collection is generally infeasible, however. Instead, cross-language collections are typically produced using a separate set of assessors for each language, and thus multiple assessors judge the same topic across the entire collection. This necessitates close coordination among assessors so that different cultural understandings of the topic can be resolved and the typical "gray areas" of relevance can be judged consistently across languages.

4.2 Other Tasks

The document ranking abstraction that is the basis of the standard Cranfield paradigm is applicable to many information access tasks. The abstraction (though not the technical solutions) is independent of document type, including not only various textual genres but other media types such as recordings of speech or images or videos. The abstraction is also independent of the expression of the information need, such as using a natural language statement, a structured query, or a sample relevant document. The abstraction applies whenever the actual user task involves a searcher interacting with a set of distinct, uniquely identified information units (the documents) returned in response to the searcher's request. Nonetheless, there are a number of realistic information access tasks that do not fit this precise abstraction. This section describes how standard Cranfield has been modified to support research for three other families of abstract tasks: filtering, focused retrieval, and web-based search.

4.2.1 Filtering Tasks

If ad hoc searching is thought of as "pull" technology where the user pulls documents from the system by querying, filtering is "push" technology where the system periodically informs the user of a new document. In the abstract filtering task, the topics of interest are relatively stable and are known in advance; the system task is to find relevant documents for each topic from a document stream (such as a newswire or social media feed). The main distinguishing feature of a filtering task is that the system must make a binary decision for each document in the stream as to whether that document will be returned to the user for the current topic, and that decision must be made relatively shortly after the document appears in the stream. Making a binary decision is a strictly more difficult task than ranking (Robertson and Callan 2005).

In a typical filtering task, systems receive feedback in the form of a relevant judgment for documents they retrieve, and adapt their processing based on the judgments. This makes set-based precision and recall measures inappropriate to evaluate system performance just as ranked-retrieval evaluation measures are clearly inappropriate. Filtering tasks are generally evaluated using some sort of utility measure, where systems are rewarded a gain for retrieving a relevant document and penalized a loss for retrieving a non-relevant document. Latency, the amount of time between when the first document appears in the stream and the decision to retrieve it can also be incorporated into the measure (Aslam et al. 2014).

Building test collections for filtering tasks requires having a document set that has a well-defined order to the document stream; generally such an order is related to time. To support adaptive filtering, the relevant set must be known prior to system execution, meaning traditional pooling based on participant runs is not an option. The TREC 2002 Filtering track compared two ways of building such a collection: using several rounds of relevance feedback searches during topic development, and using category descriptors from the document source as a kind of judgment (Soboroff and Robertson 2003). The results from the track demonstrated that these two types of collections were quite different in how they ranked systems.

4.2.2 Focused Retrieval Tasks

Focused retrieval is a general category of tasks in which documents are no longer treated as atomic entities (Trotman et al. 2010). This broad category of tasks includes passage retrieval such as in the TREC HARD track (Allan 2003); question answering such as in the TREC QA track (Voorhees 2005); and XML element retrieval as studied in *INitiative for the Evaluation of XML Retrieval (INEX)* (Bellot et al. 2014). The task abstraction for focused retrieval tasks generally maintains the ad hoc nature and ranking aspects of Cranfield, but systems do not retrieve distinct, uniquely identified information units. This means evaluation schemes can no longer use simple matching between a gold standard answer (e.g., judgment files) and system results.

Passage retrieval evaluation generally requires matching system-returned document extracts to a set of standard relevant passages. Since it is unlikely that systems will return extracts that match exactly, strict pooling to find a set of gold-standard relevant passages is not possible. Further, since systems are also unlikely to return extracts that match the gold standard passages exactly, evaluation measures must account for redundancy and omissions in the returned passages. The TREC HARD used a function of character-level recall and precision with respect to a set of gold-standard relevant passage extracts.

Question answering system evaluation shares the same problem as passage retrieval in that the answers returned by different systems are seldom exactly the same. Further, automatically determining whether a system response contains a correct answer is generally as difficult as the question answering task itself. To create a form of reusable test collection for the short-answer, factoid question in the initial

TREC QA track, organizers (manually) created regular expression patterns from the set of pooled system responses, and treated a new answer string as correct if and only if the string matched a pattern.

XML elements have unique ids (the path from the document root to the element is a unique id), but granularity is an issue for both defining the set of gold-standard relevant elements as well as matching system output to the relevant elements set. Intuitively, a system should receive some credit for retrieving an element that contains a relevant element, but only if the containing element is not too large (e.g., no credit for retrieving an entire book if the relevant element is one small element in a single sub-sub-section of a single chapter). Similarly, a system should receive credit if it retrieves a too narrow element, assuming the narrow element is comprehensible on its own. Balancing such considerations while also accounting for redundancy (e.g., not giving double credit for retrieving two elements each of which contains the same relevant sub-element) to accurately model when one system response is better than another is quite challenging. INEX has judged relevance on two scales, exhaustivity and specificity, and combined those judgments using a form of cumulated gain; see Lalmas and Tombros (2007) for details.

4.2.3 Web Tasks

While web search can be construed as an ad hoc search task over documents that happen to be web pages, Cranfield is not a good abstraction of web search for several reasons. Cranfield is an abstraction of informational search as befits its library heritage while people use the web in other ways including using search for navigation and for transactions (Broder 2002). Further, it is not clear what "the document set" is for a web test collection. Consider a web page that contains code that dynamically generates the content seen by a visitor to it. Is the document the static code? the entire data environment that determines the content seen at any given time? the particular content presented to some one visitor? The lack of editorial control gives rise to spam documents making the web the rare corpus in which the words of the documents themselves are *not* necessarily a good indicator of document content. The enormity of the web and its transitory nature precludes a classic static test collection that meaningfully represents general web search.

Particular aspects of the overall web search problem have been studied using test collections, however (Hawking and Craswell 2005). These efforts have been supported by specific crawls that gathered a coherent subset of the web at a given point of time[1] and which were then used with queries drawn from contemporaneous internet search engine logs.

Early editions of the TREC Web Track studied home page and named page finding, navigational search tasks. One outcome of this work was demonstrating

[1]See http://ir.dcs.gla.ac.uk/test_collections/access_to_data.html and https://lemurproject.org/clueweb12/.

the effectiveness of using anchor text to support navigational search (Hawking and Craswell 2005). Navigational search is generally evaluated using either reciprocal rank (the reciprocal of the rank at which the correct URL was retrieved) or "success at n", a binary measure that signifies whether the correct URL was retrieved in the top n ranks.

Given the size of the web and the brevity of the typical web search query, there is often a spectrum of information needs that the query might represent. Sometimes the query is inherently ambiguous; other times it may refer to a single broad area of interest that has multiple distinct aspects. For example, the query *avp* is ambiguous in that it might refer to the Wilkes-Barre Scranton International Airport (airport code AVP), the Avon Products Company (stock symbol AVP), the "Alien vs. Predator" movie franchise, AVP antivirus software, or the Association of Volleyball Professionals. The query *moths* likely refers to the winged insects, but the actual information need could be a desire to see pictures of moths in general; identification of a specific instance of a moth; controlling a moth infestation; distinguishing between moths and butterflies; understanding moth habitats; etc. Web pages that are excellent documents for one aspect may be completely irrelevant for another.

Diversity web tasks look to develop search systems that are able to cover the different aspects of a query within the top results. The evaluation methodology that supports diversity tasks requires a delineation of the aspects to be covered by a query statement, relevance judgments for each such aspect for each judged page, and evaluation measures that appropriately reward ranked lists for coverage as well as accuracy. Two such measures are α-NDCG (Clarke et al. 2008) and ERR-IA (Chapelle et al. 2011). Clarke et al. (2011) show that these measures behave as intended, rewarding systems that achieve a balance between novelty and overall precision.

4.3 Size Revisited

While TREC ad hoc collections contained much bigger document sets than the collections generally available at the start of TREC, the ad hoc collections are once again quite small compared to many document sets that are searched by operational systems—including, but not limited to, the web. Unfortunately, pooling has its own size dependency that prevents its successful application for arbitrarily large document sets. Pooling's fundamental assumption that the pools contain an unbiased sample of the relevant documents becomes untenable unless the size of the pools grows in concert with the size of the document set. Otherwise, the sheer number of documents of a certain type (for example, relevant documents that contain many query words) fill up the pools to the exclusion of other types of documents (for example, relevant documents that contain few query words) (Buckley et al. 2007). Systems that are able to retrieve the minority type of relevant documents are unfairly penalized when evaluated by the relevance judgments produced by shallow pools.

One way of adapting Cranfield to accommodate these larger document collection sizes is to explicitly acknowledge unjudged documents and account for them in the evaluation. Most frequently, this accommodation has been through the use of evaluation measures specifically designed for partial judgment sets.

4.3.1 Special Measures

Buckley and Voorhees (2004) introduced the bpref ("binary preference") measure as a means of evaluating retrieval systems when the relevance judgment sets are known to be far from complete. Sakai and Kando (2008) investigated the fidelity of evaluation results for incomplete judgments by comparing bpref and standard evaluation measures computed over only judged documents. That is, they computed evaluation scores by removing unjudged documents from the ranking rather than assuming those documents were not relevant and called these compressed-list versions *measure′*. Among other findings, they concluded that bpref was inferior to MAP′ in terms of both defining the set of statistically different run pairs and the overall similarity of runs ranked by effectiveness as measured by Kendall's τ correlation. To test the measures, they produced increasingly incomplete judgment sets by taking random subsets of the original judgment sets for existing test collections. Sakai subsequently showed that realistic judgment set building is subject to both system and pool depth biases (Sakai 2008a,b) that are not modeled well by random subsets. In more realistic scenarios, the compressed list versions of the measures had no clear advantage over the traditional versions, though they were superior to bpref.

Bpref and compressed list versions of standard measures can be computed using any existing test collection. A family of measures known as inferred measures are available if the test collection is constructed to support them. Inferred measures are defined as statistical estimates of the true value of the corresponding traditional measures (Yilmaz and Aslam 2008).

As an example, inferred AP computes an estimate of the expectation of the following random experiment when assuming the known (incomplete) set of relevance judgments is a uniform random sample of the complete judgment set. Given a retrieval result of a ranked list for a single topic:

1. Select a relevant document at random from the collection. Call the rank of this relevant document in the retrieved list k.
2. Select a rank i at random from among the set $\{1,\ldots k\}$.
3. Output the binary relevance of the document at rank i.

Under the assumption that a uniform random sample of the relevant documents is known, mean inferred AP is a good estimate of the actual value of MAP. However, in practice, incomplete judgment sets are seldom uniform random samples of the complete set—relevant documents retrieved higher in ranked lists are more likely to be included in the known set, for example. Inferred measures were thus extended to a collection-building method that samples from the runs in such a way as to

maintain accurate estimates of the measures' values, producing extended inferred measures (Yilmaz et al. 2008). The extended inferred measures technique builds judgment sets using stratified random sampling across the run set. That is, judgment sets to be used in computing extended inferred measures are created by taking uniform random samples of different regions of the ranked document lists where the different regions are sampled at different rates. The particular sampling strategy used affects the quality of the resulting estimates. Effective strategies do not include large, sparsely-sampled strata and do include a small top stratum that is exhaustively judged (e.g, depth-10 pools) (Voorhees 2014).

Much of the difficulty of getting fair evaluation results for large collections lies in getting good estimates of the number of relevant documents, R. The TREC Legal Track, which focused on the problem of legal discovery where *relevant* document sets can be very large, used stratified sampling (which differed from inferred measure sampling) to estimate R (Tomlinson and Hedin 2011). In this case, the strata were defined using the number of runs that retrieved a document. Others contend that R (and thus recall) is not a good basis for retrieval system evaluation since it is unimportant (relevant set sizes are too large to be meaningfully processed by a user) and unknowable. Moffat and Zobel (2008) introduced *Rank-Biased Precision (RBP)* as a measure that does not rely on knowledge of R and whose true value can be bounded in the presence of incomplete judgments.

RBP is based on a user model that assumes a user starts at the top of a ranked list and proceeds to the next document with probability p. The measure is defined as the expected rate at which relevant documents are found conditioned on p:

$$\text{RBP} = \sum_{i=1}^{n} \text{rel}_i (1 - p)^{i-1} p$$

where n is the number of ranks over which the score is computed. The RBP score for a ranking that is a prefix of another ranking is by definition a lower bound for the RBP score of the extended ranking. This property means that upper and lower bounds for RBP for a given ranking in the presence of unjudged documents are easy to compute: the lower bound is the RBP score when all unjudged documents are treated as not relevant, and the upper bound is the score when all unjudged documents are treated as relevant. Larger differences between the two bounds, caused by encountering more unjudged documents, are an indicator of greater uncertainty in the evaluation.

4.3.2 Constructing Large Collections

The stratified sampling used to support different evaluation measures is one example of modified collection construction techniques in support of building fair larger collections. Other construction techniques not tied to particular measures have also been tried.

Since the number of human judgments that can be obtained is usually the limiting factor on the size of a test collection, several approaches look at different ways of allocating assessor resources. Cormack et al. (1998) introduced the *Move-to-Front (MTF)* method of selecting the next document to judge from a set of runs based on the relevance judgments already received. The method favors selecting additional documents from runs that have recently retrieved relatively many relevant documents. Losada et al. (2016) showed that MTF is one instance of a family of multi-arm bandit document selection techniques. Each bandit method tries to maximize the number of relevant documents found while staying within a given budget of judgments, but differ in the details of precisely how the next run to contribute a document is selected. The most effective bandit methods are dynamic methods like MTF that require relevance judgments on previous selections before selecting the next document. The exploration of MTF and other bandit methods has used simulation on existing collections to show that the vast majority of relevant documents can be recovered with many fewer judgments than pooling required. Implementing dynamic methods to build a new collection from scratch is logistically more difficult than pooling, and also potentially adds assessor bias to the assessments since assessors know they are seeing documents in quality order.

Another choice when allocating assessor resources is balancing the number of topics in the test set against the exhaustiveness of the judgments for those topics. Conventional wisdom such as in the IDEAL collection report is that more topics are always better, but that is assuming essentially complete judgments for each topic. Sanderson and Zobel (2005) found that many shallowly-judged topics (whose runs were thus evaluated using precision-based measures) resulted in collections that ranked systems more similarly to an existing high-quality collection than fewer topics with deeper judgments. However, several studies have also shown that you can find small subsets of topics from a larger collection that rank systems the same as the full collection (Guiver et al. 2009; Hosseini et al. 2012). Kutlu et al. (2018) reconciles these differing findings by showing that many shallowly-judged topics are better when topics are chosen at random, but for smaller budgets and when topic development costs are comparatively high, using a selectively chosen, smaller set of more thoroughly judged topics produces a more reliable collection that is also more reusable.

The *Minimal Test Collection (MTC)* protocol is a dynamic method that can be used to create a collection being built to compare a specific set of runs known at collection build time (Carterette et al. 2006). MTC identifies those documents that will best distinguish the set of runs under some measure. Empirically, using MAP as the focus measure creates a collection that also fairly compares the run set using other measures such as Prec(10) or R-precision.

4.4 User-Based Measures

Research on evaluation measures used with test collections has not focused solely on accommodating incomplete relevance judgments. Another focus of measure work has been incorporating different models of search behavior into the measures. This section summarizes some of work in this area; also see Sakai (2014) for an alternative summary.

RBP was introduced above as a measure that accommodates partial judgments, but it also codifies a specific user model of a searcher traversing a ranked list from the top who proceeds to the next rank in the list with probability p (Moffat and Zobel 2008). ERR-IA and *alpha*-NDCG, also mentioned earlier, similarly assume what Clarke et al. (2011) call a cascade model of user behavior: considering the relationship between successive elements of a result list. Indeed, *alpha*-NDCG is an extension of the *Discounted Cumulated Gain (DCG)* family of measures that explicitly encodes the cascade user model and also accommodates different grades of relevance (Järvelin and Kekäläinen 2002).

The user model for DCG is again a searcher traversing a ranked list from the top, this time proceeding to a fixed rank k. At each rank, the searcher accumulates a gain proportional to the relevance grade of the document at the rank, with the base amount of gain for a given relevance grade reduced in proportion to the depth of the rank. While there are several different formulations of the measure, a frequently used definition (Burges et al. 2005) is

$$\mathrm{DCG}_k = \sum_{i=1}^{k} \frac{2^{\mathrm{rel}_i} - 1}{\log_2(i+1)}$$

where k is the rank to which the searcher traverses and rel_i is the gain value for the relevance grade of the document at rank i. The rank-related penalty arises from the logarithm in the denominator. The base of the logarithm is a parameter of the measure and models the persistence of the searcher reviewing the ranked list.

The maximum DCG score for a topic depends on the number of relevant documents of each grade and the gains associated with the grades. This means the maximum score across topics varies widely and hence DCG does not average well, so must be normalized producing *normalized Discounted Cumulative Gain (nDCG)*. Each topic's DCG score is normalized by dividing it by the maximum score possible for the topic. The maximum is obtained by scoring a ranked list that contains all of the documents with maximum relevant grade first, followed by all relevants with the next highest relevance grade, and so forth until all relevant documents are ranked.

Carterette (2011) shows that both nDCG and RBP, as well as more traditional measures such as average precision and reciprocal rank, can all be modeled using

sums over the product of a discount function of ranks and a gain function that maps relevance judgments to numeric utility values:

$$M = \sum_{k=1}^{K} \text{gain}(\text{rel}_k) \times \text{discount}(k).$$

He then shows that any such measure is actually a composition of three independent component models, the model that describes how the user interacts with the results called the browsing model; the model that describes the utility the user obtains from an individual document called the document utility model; and the model that describes how utility is accumulated while browsing, the utility accumulation model. By relating measures that share a common component model, the framework can unify previously disparate measures into a small set of measure families, as well as suggest new measures that would fill previously unoccupied areas of the measure space. One outcome of the initial investigation into the measure space was a demonstration that DCG is a robust measure that does in fact model user-centered behavior.

5 Conclusion

The Cranfield paradigm has proved to be remarkably resilient. Despite fierce criticism from the start and periodic pronouncements of its impending demise, the paradigm has enabled research that has greatly improved retrieval performance in practice. This success has largely resulted *because* of the paradigm's limitations rather than despite them. The document ranking task is a carefully calibrated level of abstraction that has sufficient fidelity to real user tasks to be informative, but is sufficiently abstract to be broadly applicable, feasible to implement, and comparatively inexpensive. By eliminating anything that does not directly contribute to the core competency, Cranfield loses realism but gains substantial experimental power.

Maintaining a proper tension between realism and abstraction is key to extending the paradigm to new tasks. It obviously does no good to abstract an evaluation task to the point where test results do not reflect performance on the real task of interest; it is equally as unhelpful to include any operational variable that might possibly influence outcomes since generalization then becomes impossible and nothing is learned.

References

Al-Maskari A, Sanderson M, Clough P, Airio E (2008) The good and the bad system: does the test collection predict users' effectiveness? In: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '08. ACM, New York, pp 59–66

Allan J (2003) HARD Track overview in TREC 2003: high accuracy retrieval from documents. In: Proceedings of the twelfth Text REtrieval Conference (TREC 2003)

Aslam J, Ekstrand-Abueg M, Pavlu V, Diaz F, McCreadie R, Sakai T (2014) TREC 2014 temporal summarization track overview. In: Proceedings of the twenty-third Text REtrieval Conference (TREC 2014)

Banks D, Over P, Zhang NF (1999) Blind men and elephants: six approaches to TREC data. Inf Retr 1:7–34

Bellot P, Bogers T, Geva S, Hall MA, Huurdeman HC, Kamps J, Kazai G, Koolen M, Moriceau V, Mothe J, Preminger M, SanJuan E, Schenkel R, Skov M, Tannier X, Walsh D (2014) Overview of INEX 2014. In: Kanoulas E, Lupu M, Clough P, Sanderson M, Hall M, Hanbury A, Toms E (eds) Information access evaluation – multilinguality, multimodality, and interaction. Proceedings of the fifth international conference of the CLEF initiative (CLEF 2014). Lecture notes in computer science (LNCS), vol 8685. Springer, Heidelberg, pp 212–228

Broder A (2002) A taxonomy of web search. SIGIR Forum 36(2):3–10

Buckley C (2001) The TREC-9 query track. In: Voorhees E, Harman D (eds) Proceedings of the ninth Text REtreival Conference (TREC-9), pp 81–85

Buckley C, Voorhees EM (2000) Evaluating evaluation measure stability. In: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval, SIGIR 2000, pp 33–40

Buckley C, Voorhees EM (2004) Retrieval evaluation with incomplete information. In: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, pp 25–32

Buckley C, Voorhees EM (2005) Retrieval system evaluation. In: Voorhees EM, Harman DK (eds) TREC: experiment and evaluation in information retrieval. MIT Press, Boston, chap 3, pp 53–75

Buckley C, Dimmick D, Soboroff I, Voorhees E (2007) Bias and the limits of pooling for large collections. Inf Retr 10:491–508

Burges C, Shaked T, Renshaw E, Lazier A, Deeds M, Hamilton N, Hullender G (2005) Learning to rank using gradient descent. In: Proceedings of the 22nd international conference on machine learning, ICML '05. ACM, New York, pp 89–96

Burgin R (1992) Variations in relevance judgments and the evaluation of retrieval performance. Inf Process Manag 28(5):619–627

Carterette B (2011) System effectiveness, user models, and user utility: a conceptual framework for investigation. In: Proceedings of the 34th International ACM SIGIR conference on research and development in information retrieval (SIGIR'11). ACM, New York, pp 903–912

Carterette BA (2012) Multiple testing in statistical analysis of systems-based information retrieval experiments. ACM Trans Inf Syst 30(1):4:1–4:34

Carterette B (2015) The best published result is random: Sequential testing and its effect on reported effectiveness. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, SIGIR '15, pp 747–750

Carterette B, Allan J, Sitaraman R (2006) Minimal test collection for retrieval evaluation. In: Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, pp 268–275

Chapelle O, Ji S, Liao C, Velipasaoglu E, Lai L, Wu SL (2011) Intent-based diversification of web search results: metrics and algorithms. Inf Retr 14(6):572–592

Clarke CL, Kolla M, Cormack GV, Vechtomova O, Ashkan A, Büttcher S, MacKinnon I (2008) Novelty and diversity in information retrieval evaluation. In: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '08. ACM, New York, pp 659–666

Clarke CL, Craswell N, Soboroff I, Ashkan A (2011) A comparative analysis of cascade measures for novelty and diversity. In: Proceedings of the fourth ACM international conference on web search and data mining, WSDM '11. ACM, New York, pp 75–84

Cleverdon CW (1967) The Cranfield tests on index language devices. In: Aslib proceedings, vol 19, pp 173–192, (Reprinted in *Readings in Information Retrieval*, K. Spärck-Jones and P. Willett, editors, Morgan Kaufmann, 1997)

Cleverdon CW (1970) The effect of variations in relevance assessments in comparative experimental tests of index languages. Tech. Rep. Cranfield Library Report No. 3, Cranfield Institute of Technology, Cranfield, UK

Cleverdon CW (1991) The significance of the Cranfield tests on index languages. In: Proceedings of the fourteenth annual international ACM/SIGIR conference on research and development in information retrieval, pp 3–12

Cormack GV, Palmer CR, Clarke CLA (1998) Efficient construction of large test collections. In: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '98. ACM, New York, pp 282–289

Cuadra CA, Katter RV (1967) Opening the black box of relevance. J Doc 23(4):291–303

Gilbert H, Spärck Jones K (1979) Statistical bases of relevance assessment for the 'IDEAL' information retrieval test collection. Tech. rep., Computer Laboratory, University of Cambridge. Available at http://sigir.org/resources/museum/

Guiver J, Mizzaro S, Robertson S (2009) A few good topics: experiments in topic set reduction for retrieval evaluation. ACM Trans Inf Syst 27(4):21:1–21:26

Harman D (1996) Overview of the fourth Text REtrieval Conference (TREC-4). In: Harman DK (ed) Proceedings of the fourth Text REtrieval Conference (TREC-4), pp 1–23, nIST Special Publication 500-236

Harter SP (1996) Variations in relevance assessments and the measurement of retrieval effectiveness. J Am Soc Inf Sci 47(1):37–49

Hawking D, Craswell N (2005) The very large collection and web tracks. In: Voorhees EM, Harman DK (eds) TREC: experiment and evaluation in information retrieval. MIT Press, Boston, chap 9, pp 199–231

Hersh W, Turpin A, Price S, Chan B, Kraemer D, Sacherek L, Olson D (2000) Do batch and user evaluations give the same results? In: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval, SIGIR 2000, pp 17–24

Hofmann K, Li L, Radlinski F (2016) Online evaluation for information retrieval. Found Trends Inf Retr 10(1):1–117

Hosseini M, Cox IJ, Milic-Frayling N, Shokouhi M, Yilmaz E (2012) An uncertainty-aware query selection model for evaluation of IR systems. In: Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval, SIGIR '12. ACM, New York, pp 901–910

Hull D (1993) Using statistical testing in the evaluation of retrieval experiments. In: Proceedings of the 16th annual international ACM/SIGIR conference on research and development in information retrieval, SIGIR '93. ACM, New York, pp 329–338

Järvelin K, Kekäläinen J (2002) Cumulated gain-based evaluation of IR techniques. ACM Trans Inf Syst 20(4):422–446

Kando N, Kuriyama K, Nozue T, Eguchi K, Kato H, Hidaka S (1999) Overview of IR tasks at the first NTCIR workshop. In: Proceedings of the first NTCIR workshop on research in Japanese text retrieval and term recognition, pp 11–44

Keen EM (1966) Measures and averaging methods used in performance testing of indexing systems. Tech. rep., The College of Aeronautics, Cranfield, England. Available at http://sigir.org/resources/museum/

Kelly D (2009) Methods for evaluating interactive information retrieval systems with users. Found Trends Inf Retr 3(1–2):1–224

Kutlu M, Elsayed T, Lease M (2018) Learning to effectively select topics for information retrieval test collections. Inf Process Manag 54(1):37–59

Lalmas M, Tombros A (2007) Evaluating XML retrieval effectiveness at INEX. SIGIR Forum 41(1):40–57

Ledwith R (1992) On the difficulties of applying the results of information retrieval research to aid in the searching of large scientific databases. Inf Process Manag 28(4):451–455

Lesk ME (1967) SIG – the significance programs for testing the evaluation output. In: Information storage and retrieval, Scientific Report No. ISR-12, National Science Foundation, chap II

Lesk M, Salton G (1969) Relevance assessments and retrieval system evaluation. Inf Storage Retr 4:343–359

Losada DE, Parapar J, Barreiro A (2016) Feeling lucky?: multi-armed bandits for ordering judgements in pooling-based evaluation. In: Proceedings of the 31st annual ACM symposium on applied computing, SAC '16. ACM, New York, pp 1027–1034

Mandl T, Womser-Hacker C (2003) Linguistic and statistical analysis of the clef topics. In: Peters C, Braschler M, Gonzalo J, Kluck M (eds) Advances in cross-language information retrieval: third workshop of the cross-language evaluation forum (CLEF 2002) revised papers. Lecture notes in computer science (LNCS), vol 2785. Springer, Heidelberg, pp 505–511

Moffat A, Zobel J (2008) Rank-biased precision for measurement of retrieval effectiveness. ACM Trans Inf Syst 27(1):Article 2

Robertson S (2008) On the history of evaluation in IR. J Inf Sci 34(4):439–456

Robertson S, Callan J (2005) Routing and filtering. In: Voorhees EM, Harman DK (eds) TREC: experiment and evaluation in information retrieval. MIT Press, Boston, chap 5, pp 99–121

Robertson S, Hancock-Beaulieu M (1992) On the evaluation of IR systems. Inf Process Manag 28(4):457–466

Sakai T (2006) Evaluating evaluation metrics based on the bootstrap. In: Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '06. ACM, New York, pp 525–532

Sakai T (2008a) Comparing metrics across TREC and NTCIR: the robustness to pool depth bias. In: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, pp 691–692

Sakai T (2008b) Comparing metrics across TREC and NTCIR: the robustness to system bias. In: Proceedings of the 17th ACM conference on information and knowledge management, pp 581–590

Sakai T (2014) Metrics, statistics, tests. In: Ferro N (ed) 2013 PROMISE winter school: bridging between information retrieval and databases. Lecture notes in computer science (LNCS), vol 8173 . Springer, Heidelberg, pp 116–163

Sakai T (2016) Statistical significance, power, and sample sizes: a systematic review of SIGIR and TOIS, 2006-2015. In: Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval, SIGIR '16. ACM, New York, pp 5–14

Sakai T, Kando N (2008) On information retrieval metrics designed for evaluation with incomplete relevance assessments. Inf Retr 11:447–470

Sanderson M (2010) Test collection based evaluation of information retrieval systems. Found Trends Inf Retr 4(4):247–375

Sanderson M, Zobel J (2005) Information retrieval system evaluation: effort, sensitivity, and reliability. In: Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '05. ACM, New York, pp 162–169

Savoy J (1997) Statistical inference in retrieval effectiveness evaluation. Inf Process Manag 33(4):495–512

Smucker MD, Allan J, Carterette B (2007) A comparison of statistical significance tests for information retrieval evaluation. In: Proceedings of the sixteenth ACM conference on conference on information and knowledge management, CIKM '07. ACM, New York, pp 623–632

Soboroff I, Robertson S (2003) Building a filtering test collection for trec 2002. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '03. ACM, New York, pp 243–250

Spärck Jones K (1974) Automatic indexing. J Doc 30:393–432

Spärck Jones K (2001) Automatic language and information processing: rethinking evaluation. Nat Lang Eng 7(1):29–46

Spärck Jones K, Bates RG (1977) Report on a design study for the 'IDEAL' information retrieval test collection. Tech. rep., Computer Laboratory, University of Cambridge. Available at http://sigir.org/resources/museum/

Spärck Jones K, Van Rijsbergen C (1975) Report on the need for and provision for and 'IDEAL' information retrieval test collection. Tech. rep., Computer Laboratory, University of Cambridge. Available at http://sigir.org/resources/museum/

Taube M (1965) A note on the pseudomathematics of relevance. Am Doc 16(2):69–72

Tomlinson S, Hedin B (2011) Measuring effectiveness in the TREC legal track. In: Lupu M, Mayer K, Tait J, Trippe A (eds) Current challenges in patent information retrieval. The information retrieval series, vol 29. Springer, Berlin, pp 167–180

Trotman A, Geva S, Kamps J, Lalmas M, Murdock V (2010) Current research in focused retrieval and result aggregation. Inf Retr 13(5):407–411

Turpin AH, Hersh W (2001) Why batch and user evaluations do not give the same results. In: Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '01, pp 225–231

Van Rijsbergen C (1979) Evaluation, 2nd edn. Butterworths, London, chap 7

Voorhees EM (2000) Variations in relevance judgments and the measurement of retrieval effectiveness. Inf Process Process 36:697–716

Voorhees EM (2005) Question answering in TREC. In: Voorhees EM, Harman DK (eds) TREC: experiment and evaluation in information retrieval. MIT Press, Boston, chap 10, pp 233–257

Voorhees EM (2014) The effect of sampling strategy on inferred measures. In: Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval, SIGIR '14. ACM, New York, pp 1119–1122

Voorhees EM, Buckley C (2002) The effect of topic set size on retrieval experiment error. In: Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '02. ACM, New York, pp 316–323

Voorhees EM, Harman DK (2005) The Text REtrieval Conference. In: Voorhees EM, Harman DK (eds) TREC: experiment and evaluation in information retrieval. MIT Press, Boston, chap 1, pp 3–19

Yilmaz E, Aslam JA (2008) Estimating average precision when judgments are incomplete. Knowl Inf Syst 16:173–211

Yilmaz E, Kanoulas E, Aslam JA (2008) A simple and efficient sampling method for estimating AP and NDCG. In: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, pp 603–610

Zobel J (1998) How reliable are the results of large-scale information retrieval experiments? In: Croft WB, Moffat A, van Rijsbergen C, Wilkinson R, Zobel J (eds) Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, Melbourne, Australia. ACM Press, New York, pp 307–314