Foundations and Trends® in
Information Retrieval
Vol. 4, No. 4 (2010) 247–375

DOI: 10.1561/1500000009

now
the essence of knowledge

Test Collection Based Evaluation of Information Retrieval Systems

By Mark Sanderson

Contents

Foundations and Trends® in
Information Retrieval
Vol. 4, No. 4 (2010) 247–375

DOI: 10.1561/1500000009

now
the essence of knowledge

Test Collection Based Evaluation of Information Retrieval Systems

Mark Sanderson

The Information School, University of Sheffield, Sheffield, UK
m.sanderson@shef.ac.uk

Abstract

Use of test collections and evaluation measures to assess the effectiveness of information retrieval systems has its origins in work dating back to the early 1950s. Across the nearly 60 years since that work started, use of test collections is a de facto standard of evaluation. This monograph surveys the research conducted and explains the methods and measures devised for evaluation of retrieval systems, including a detailed look at the use of statistical significance testing in retrieval experimentation. This monograph reviews more recent examinations of the validity of the test collection approach and evaluation measures as well as outlining trends in current research exploiting query logs and live labs. At its core, the modern-day test collection is little different from the structures that the pioneering researchers in the 1950s and 1960s conceived of. This tutorial and review shows that despite its age, this long-standing evaluation method is still a highly valued tool for retrieval research.

1

Introduction

An examination of the opening pages of a number of Information Retrieval (IR) books reveals that each author defines the topic of IR in different ways. Some say that IR is simply a field concerned with organizing information [210]; and others emphasize the range of different materials that need to be searched [286]. While others stress the contrast between the strong structure and typing of a database (DB) system with the lack of structure in the objects typically searched in IR [262, 244]. Across all of these definitions, there is a constant, IR systems have to deal with incomplete or *underspecified* information in the form of the queries issued by users. The IR systems receiving such queries need to fill in the gaps of the users' underspecified query.

For example, a user typing "nuclear waste dumping" into the search engine of an academic repository is probably looking for multiple documents describing this topic in detail, he/she probably prefers to see documents from reputable sources, but all he/she enters into the search engine are three words. Users querying on a web search engine for "BBC" are probably looking for the official home page of the corporation, yet they fully expect the search engine to infer that specific information request from the three letters entered. The fact that the

content being searched is typically unstructured and its components (i.e., words) can have multiple senses, and different words can be used to express the same concept, merely adds to the challenge of locating relevant items. In contrast to a DB system, whose search outputs are deterministic, the accuracy of an IR system's output cannot be predicted with any confidence prior to a search being conducted; consequently, empirical evaluation has always been a critical component of Information Retrieval.[1]

The typical interaction between a user and an IR system has the user submitting a query to the system, which returns a ranked list of objects that hopefully have some degree of relevance to the user's request with the most relevant at the top of the list. The success of such an interaction is affected by many factors, the range of which has long been considered. For example, Cleverdon and Keen [61, p. 4] described five.

(1) *"The ability of the system to present all relevant documents*
(2) *The ability of the system to withhold non-relevant documents*
(3) *The interval between the demand being made and the answer being given (i.e., time)*
(4) *The physical form of the output (i.e., presentation)*
(5) *The effort, intellectual or physical, demanded of the user (i.e., effort)."*

To this list one could add many others, e.g.:

- the ability of the user at specifying their need;
- the interplay of the components of which the search algorithm is composed;
- the type of user information need;
- the number of relevant documents in the collection being searched;
- the types of documents in the collection;

[1] This is not to say that researchers haven't tried to devise non-empirical approaches, such as building theoretical models of IR systems. However, Robertson [197] points out that a theory of IR that would allow one to predict performance without evaluation remains elusive.

- the context in which the user's query was issued; and
- the eventual use for the information being sought.

Evaluation of IR systems is a broad topic covering many areas including information-seeking behavior usability of the system's interface; its broader contextual use; the compute efficiency, cost, and resource needs of search engines. A strong focus of IR research has been on measuring the *effectiveness* of an IR system: determining the *relevance* of items, retrieved by a search engine, relative to a user's information need.

The vast majority of published IR research assessed effectiveness using a resource known as a *test collection* used in conjunction with *evaluation measures*. Such is the importance of test collections that at the time of writing, there are many conferences and meetings devoted purely to their use: including three international conferences, TREC, CLEF, and NTCIR, which together have run more than 30 times since the early 1990s. This research focus is not just a feature of the past two decades but part of a longer tradition which was motivated by the creation and sharing of testing environments in the previous three decades, which itself was inspired by innovative work conducted in the 1950s. The classic components of a test collection are as follows:

- a collection of documents; each document is given a unique identifier, a *docid*;
- a set of topics (also referred to as queries); each given a query id (*qid*); and
- a set of *relevance judgments* (often referred to as *qrels* — query relevance set) composed of a list of qid/docid pairs, detailing the relevance of documents to topics.

In the possession of an appropriate test collection, an IR developer or researcher simply loads the documents into their system and in a batch process, submits the topics to the system one-by-one. The list of the docids retrieved for each of the topics is concatenated into a set, known as a *run*. Then the content of the run is examined to determine which of the documents retrieved were present in the qrels and

which were not. Finally, an evaluation measure is used to quantify the effectiveness of that run.

Together, the collection and chosen evaluation measure provide a *simulation* of users of a searching system in an operational setting. Using test collections, researchers can assess a retrieval system in isolation helping locate points of failure, but more commonly, collections are used to compare the effectiveness of multiple retrieval systems. Either rival systems are compared with each other, or different configurations of the same system are contrasted. Such determinations, by implication, predict how well the retrieval systems will perform relative to each other if they were deployed in the operational setting simulated by the test collection.

A key innovation in the IR academic community was the early recognition of the importance of building and crucially sharing test collections.[2] Through sharing, others benefited from the initial (substantial) effort put into the creation of a test collection by re-using it in other experiments. Groups evaluating their own IR systems on a shared collection could make meaningful comparisons with published results tested on the same collection. Shared test collections provided a focus for many international collaborative research exercises. Experiments using them constituted the main methodology for validating new retrieval approaches. In short, test collections are a catalyst for research in the IR community.

Although there has been a steady stream of research in evaluation methods, there has been little survey of literature covering test collection based evaluation. Salton's evaluation section [210, Section 5] is one such document; a chapter in Van Rijsbergen's book [262] another; Spärck Jones's edited articles on IR experiments [242] a third. Since those works, no broad surveys of evaluation appear to have been written; though Hearst has recently written about usability evaluation in IR [116, Section 3]. The sections on evaluation in recent IR books provided the essential details on how to conduct evaluation, rather than reviewed

[2] Indeed, it would appear that the academic IR community is one of the first in the Human Language Technologies (HLT) discipline of computer science to create and share common testing environments. Many other areas of HLT, such as summarization, or word sense disambiguation did not start building such shared testing resources until the 1990s.

past work. There are notable publications addressing particular aspects of evaluation: Voorhees and Harman's book detailed the history of the TREC evaluation exercise and outlined evaluation methods used [280]; a special issue of Information Processing and Management reflected the state of IR evaluation in 1992 [98]; another special issue in the *Journal of the American Society for Information Science* provided a later perspective [253]. More recently, Robertson published his personal view on the history of IR evaluation [199]. However, there remains a gap in the literature, which this monograph attempts to fill.

Using test collections to assess the effectiveness of IR systems is itself a broad area covering a wide range of document types and forms of retrieval. IR systems were built to search over text, music, speech, images, video, chemical structures, etc. For this monograph, we focus on evaluation of retrieval from documents that are searched by their text content and similarly queried by text; although, many of the methods described are applicable to other forms of IR.

Since the initial steps of search evaluation in the 1950s, test collections and evaluation measures were developed and adapted to reflect the changing priorities and needs of IR researchers. Often changes in test collection design caused changes in evaluation measures and vice versa. Therefore, the work in these two distinct areas of study are described together and laid out in a chronological order. The research is grouped into three periods, which are defined relative to the highly important evaluation exercise, TREC.

- **Early 1950s–early1990s**, Section 2: the initial development of test collections and measures. In this time, test collection content was mostly composed of catalogue information about academic papers or later the full-text of newspaper articles. The evaluation measures commonly used by researchers were strongly focused on *high recall* search: finding as many relevant items as possible.
- **Early 1990s–early 2000s**, Section 3: the "TREC ad hoc" period. Scale and standardization of evaluation were strong themes of this decade. The IR research community collaborated to build a relatively small number of large test

collections mainly composed of news articles. Evaluation was still focused on high recall search.

- **Early 2000s–present**, Section 4: the post ad hoc period (for want of a better name). Reflecting the growing diversity in application of search technologies and the ever-growing scale of collections being searched, evaluation research in this time showed a diversification of content and search task along with an increasing range of evaluation measures that reflected user's more common preference for finding a small number of relevant items. Run data gathered by TREC and other similar exercises fostered of a new form of evaluation research in this period: studying test collection methodologies. This research is covered in Section 6.

The one exception to the ordering can be found in the section on the use of significance testing. Apart from a recent book [74], little has been written on the use of significance in IR evaluation and relatively little research has been conducted; consequently, I chose to describe research in this area, in Section 5, more as a tutorial than a survey.

Such an ordering means that descriptions of or references to evaluation measures are spread throughout the document. Therefore, we provide an index at the conclusion of this work to aid in their location.

Note, unless explicitly stated otherwise, the original versions of all work cited in this document were obtained and read by the author.

2

The Initial Development of Test Collections

The genesis of IR evaluation is generally seen as starting with the work of Cleverdon and his Cranfield collections, built in the early 1960s. However, he and others were working on retrieval evaluation for most of the 1950s. In his article looking back over his career, Cleverdon [60] stated that along with a collaborator, Thorne, he created a small test collection in 1953. The intention was to test the effectiveness of librarians at locating documents indexed by different library cataloguing systems when faced with information requests from library users. This work was first described by Thorne two years after it was completed [257].

Thorne described the motivation for conducting this testing in terms that have a strong a resonance with the motivations of IR researchers today. "*the author has found the need for a 'yardstick' to assist in assessing a particular system's merits ... the arguments of librarians would be more fertile if there were quantitative assessments of efficiency of various cataloguing systems in various libraries*". In describing their methodology for testing, Thorne stated "*Suppose the questions put to the catalogue* [from users] *are entered in a log, and 100 test questions are prepared which are believed to represent typically such a log. If the*

test questions are based on material known to be included in the collection, they can then be used to assess the catalogue's probability of success".

The paper listed 50 statements of information need that were used to assess a series of library cataloguing systems. Thorne and Cleverdon tests were essentially a form of *known item searching*. To illustrate, the following is an information need taken from Appendix C of the paper: "*The pressure distributions over the nose of a body of revolution of fineness ratio 6 for angles of attack 0° to 8° at high subsonic Mach number* ($RN > 4 \times 10^{\circ}$)." This request was generated by the authors from a document known to be catalogued in a library. Assessments were based not only on success in finding the known item, but also consideration of the costs of implementing the cataloguing system. Note in Salton's writings, this early test collection was often referred to as Cranfield I (though Cleverdon called a different collection by that name).

In the same year of Cleverdon and Thorne's early efforts, Gull (who published the work in 1956, [97]) also reported building a form of test collection. Composed of 15,000 catalogue entries, the collection was built to compare two library cataloguing systems, each built by a separate group. In total, 98 queries (called requests by Gull) were created and searchers from each group worked to locate as many relevant documents for these requests as possible. Each group formed its own relevance judgments independently, which proved to be problematic, as they discovered that their judgments were quite different from each other based on different interpretations of the queries. Gull stated that one group took a more liberal view of relevance than the other. (Cleverdon stated in [60], that after seeing the problems created by independently formed qrels, he decided to centralize relevance judgments for his collections.)

In the 1950s, computers started to be used for searching of library catalogues. An early mention of "*machines*" being involved in IR was by Kent et al. [155], who proposed an evaluation methodology that they called "*a framework of reference for analyzing the performance of an IR system*". The framework described was similar to a modern test collection. Maron et al. [173] as part of their work in experimenting with

probabilistic indexing described a form of evaluation using a collection of 110 documents and 40 queries. Fels [84] detailed a methodology for testing retrieval effectiveness proposed by Mooers [181]. Bryant [35] briefly described the work by Borko [29] who, according to Bryant, constructed a test collection composed of 612 abstracts. In an appendix of his evaluation survey paper Robertson [195] details a number of other early tests; see also the books from Lancaster [159] and from Spärck Jones [242] for more on the early developments in IR evaluation.

Given that the very first uses of computers for searching only date back to the late 1940s,[1] evaluation of searching systems was clearly an early and important priority for IR researchers. These works, however, are little remembered by today's researchers due to the detailed construction of a test collection that Cleverdon started in the late 1950s.

2.1 Cleverdon's Cranfield Collection

In his reflective piece, Cleverdon cited an editorial from American Documentation[2] (now renamed JASIST) stating that "*evaluation of all experimental results is essential in rating the efficiency of* [IR] *systems*". Cleverdon argued that it wasn't good enough for groups to evaluate their own systems, an independently run evaluation was needed. Consequently, he was funded to test four competing indexing approaches on a collection composed of 18,000 papers [57]. The papers were manually indexed using each of the four classification methodologies. Once the indexes were built, the papers were searched with 1,200 "search questions". The questions were designed to retrieve one of the collection

[1] Holmstrom described a "*machine called the Univac*" capable of searching for text references associated with a subject code. The code and text were stored on a magnetic steel tape. Holmstrom stated that the machine could process "*at the rate of 120 words per minute*" [123]. Note, the UNIVAC isn't generally thought to have come into existence until 1951, the date when the first machine was sold, Holmstrom presumably saw or was told about a pre-production version. See also Mooers — creator of the term information retrieval — for further historical references to mechanical searching devices of the early twentieth century [181].

[2] 1955, "The Truth, The Whole Truth..." American Documentation. Vol. 6 p. 58; it was not possible to locate this editorial.

papers; if that paper was retrieved, the search was considered a success. The collection became known as Cranfield I. Cleverdon reported on the results of his comparison of the four methods and from the experience of this collection, decided to develop Cranfield II.

Cleverdon felt that the relatively large size of Cranfield I was not important in ensuring that measurements were reliable. Therefore, the new collection was composed of 1,400 "documents" (titles, author names, and abstracts) derived from the references listed in around 200 recent research papers. The authors of those papers were contacted and asked to write a question that summarized the problem their paper addressed, these became the collection topics. The authors were also asked to rate each reference in their paper on a scale of 1–5 on how relevant the reference was to the stated question and if possible to provide additional references. Cleverdon's students checked all documents against all questions and contacted the authors of each question asking them if they considered any additional documents found to be relevant. All this work resulted in a collection comprising 1,400 documents, 221 topics, and a set of complete variable level relevance judgments.

Cleverdon was not alone in creating test collections, Salton instigated the creation of a series of test collections: collectively known as the SMART collections (named after the experimental retrieval system that Salton and his students built). In 1968, along with Lesk [164], he described research using two collections, the ADI, a collection of short academic papers, and the IRE-3 collection composed of the abstracts of computer science publications. Later, Salton and Yu [215] described two more: Time and MEDLARS, the first is composed of 425 full-text articles from Time magazine; the second composed of 450 abstracts of medical literature. Note, this MEDLARS collection is different from the test collection with the same name built by Lancaster [158] who in an extensive evaluation of the MEDLARS system created a test collection from 410 actual search requests submitted to the system. Another popular test collection was the NPL created by Vaswani and Cameron [263]. To illustrate the scale of these collections, a number of the more commonly used are detailed in the following table. For details on others, see Spärck Jones and Van Rijsbergen's survey [246].

Name	Docs.	Qrys.	Year[3]	Size, Mb	Source document
Cranfield 2	1,400	225	1962	1.6	Title, authors, source, abstract of scientific papers from the aeronautic research field, largely ranging from 1945 to 1962.
ADI	82	35	1968	0.04	A set of short papers from the 1963 Annual Meeting of the American Documentation Institute.
IRE-3	780	34	1968	—	A set of abstracts of computer science documents, published in 1959–1961.
NPL	11,571	93	1970	3.1	Title, abstract of journal papers
MEDLARS	450	29	1973	—	The first page of a set of MEDLARS documents copied at the National Library of Medicine.
Time	425	83	1973	1.5	Full-text articles from the 1963 edition of Time magazine.

2.2 Evaluating Boolean Retrieval Systems on a Test Collection

With the creation of test collections came the need for effectiveness measures. Many early IR systems produced Boolean output: an unordered set of documents matching a user's query; evaluation measures were defined to assess this form of output. Kent et al. [155], listed what they considered to be the important quantities to be calculated in Boolean search:

- n — the number of documents in a collection;
- m — the number of documents retrieved;
- w — the number that are both relevant and retrieved; and
- x — the total number of documents in the collection that are relevant.

[3] Year is either the year when the document describing the collection was published or the year of the first reference to use of the collection.

Inspired by work from Vickery [265, p. 174], Cleverdon and Keen [61, p. 34] produced a *contingency table* of all possible quantities that could be calculated. The table is reproduced below including Kent et al.'s original labels.

	Relevant	Not-relevant	
Retrieved	$a(w)$	b	$a + b(m)$
Not retrieved	c	d	$c + d$
	$a + c(x)$	$b + d$	$a + b + c + d(n)$

Both Kent et al. and Cleverdon and Keen listed measures that could be created out of combinations of the table's cells. The three that are probably the best known are

$$\text{Precision} = \frac{a}{a+b} \qquad \text{Recall} = \frac{a}{a+c} \qquad \text{Fallout} = \frac{b}{b+d}$$

Where *precision* measures the fraction of retrieved documents that are relevant, *recall* measures the fraction of relevant documents retrieved and *fallout* measures the fraction of non-relevant documents retrieved. Although commonly described in IR text books, fallout is by far the least used in published IR research.

Of all the measures that were proposed, two — precision and recall — dominated evaluation from the start. Reporting on his 1953 test collection work, Gull [97] appeared to be the first to describe recall, measuring competing systems by dividing "*actual retrieval*" by "*optimum retrieval*". Precision and recall were first described together by Kent et al. [155]. In their paper, precision was referred to as a "*pertinence factor*"; recall was called "*recall factor*". Kent et al. stated that neither factor could be used on its own; both measures had to be taken into account to determine effectiveness of a retrieval system. Cleverdon, who called the precision and recall measures, respectively, "*relevance ratio*" and "*recall ratio*" described an inverse relationship between the two [58, pp. 71–72], showing that if one issued Boolean searches that precisely targeted relevant documents and avoided the retrieval of non-relevant, precision would likely be high, but recall would be low. If a query could be broadened in some way to improve recall, the almost inevitable consequence was that more non-relevant documents would

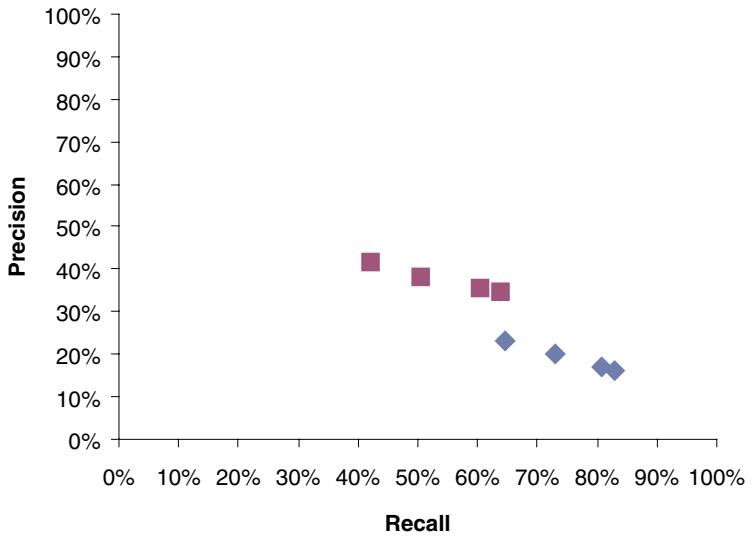

Fig. 2.1 A reproduction of Cleverdon's original recall precision graph, comparing two forms of retrieval.[4]

be returned, causing precision to drop. Using the Cranfield II test collection, he graphed recall/precision data points corresponding to the different Boolean queries; the graph is reproduced in Figure 2.1.

2.2.1 Summarizing Precision and Recall to a Single Value

A great deal of evaluation research addressed the question of how to conflate the two measures into a single value. Van Rijsbergen [261] surveyed a range of such measures. Later in his book, he proposed a measure, which is one minus the weighted harmonic mean of recall and precision, which he called e. Although this measure was sometimes used [73], the weighted harmonic mean was more extensively used in IR literature; it is commonly referred to as f, and is defined as follows.

$$f = \frac{1}{\alpha\left(\frac{1}{P}\right) + (1-\alpha)\left(\frac{1}{R}\right)}.$$

[4] Note, the axes and their labels are changed here from the way that Cleverdon drew the graph, so as to reflect the modern convention in presenting such data.

The tuning constant α is used to indicate a preference for how much influence precision or recall has on the value of f. It is common for α to be set to 0.5, which then allows f to be defined as:

$$f = \frac{1}{\frac{1}{2}\left(\frac{1}{R} + \frac{1}{P}\right)} \quad \text{or the equivalent form} \quad f = \frac{2 \times P \times R}{P + R}.$$

2.3 Evaluating over a Document Ranking

The measures described so far work over an unordered set of retrieved documents as would be returned by a Boolean IR system. The development of ranked retrieval — see for example, Maron et al., [173] — required changes in evaluation as the balance of relevant and non-relevant documents varied over the ranking. For any query that was a reasonable reflection of a user's information need, the retrieved documents that matched the query well tended to be mostly relevant and ranked highly. Relevant documents in the collection that matched the query less well appeared further down the ranking mixed in with progressively greater numbers of non-relevant documents.

Swets [249] formally described this situation. He suggested that for any given document ranking, the proportion of relevant to non-relevant documents could be described by two distributions: one for the relevant documents and one for the non-relevant. Swets did not have search output to work with, and so could only speculate on the shape of the distributions: he initially suggested that they would both be normal, though later described other possibilities [250]. Bookstein [28] described potential problems with Swets's model with normal distributions in place. Much later, researchers such as Manmatha, et al. [171] analyzed large sets of ranks and confirmed Swets's formalisms. They found that relevant documents adhered to a normal distribution and the non-relevant followed a negative exponential. Graphs illustrating the distributions of two retrieval systems are shown in Figure 2.2. High scoring documents (on the right of the graphs) are all or nearly all relevant, but for documents that match the query progressively less well (moving to the left), the balance of relevant to non-relevant shifts to a greater proportion of non-relevant being retrieved. The exact nature of the balance

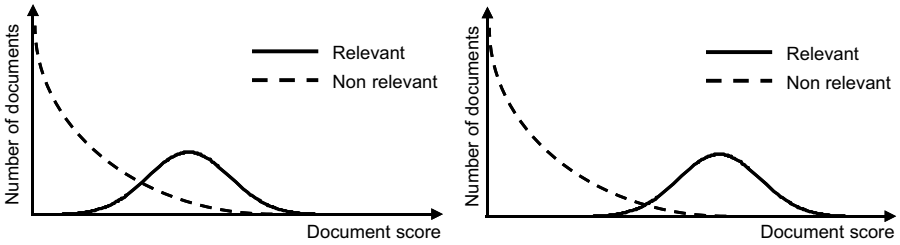

Fig. 2.2 Representations of the balance between relevant and non-relevant documents across a document ranking. The graph on the left represents a less effective retrieval system than the graph on the right.

and the way that it changes across a rank for a particular retrieval system will depend on that system's effectiveness: a good system will produce a ranking that has a strong separation between the distributions (e.g., graph on right of Figure 2.2) and a poor system the opposite (e.g., graph on the left), Swets suggested that this approach could form the basis of an evaluation measure for IR systems, however, the idea was not taken up by the community, who instead choose to focus on adapting precision and recall to ranked retrieval.

2.3.1 Plotting Precision on a Graph

An early popular approach to evaluation of ranked retrieval was to graph precision values, measured over the document ranking averaged across multiple topics. However, as can be seen from the example in Figure 2.3, plotting recall and precision computed at each relevant document for the ranks of two topics results in a scatter plot of discreet points that before being averaged need to be transformed to a pair of continuous functions using interpolation.

Many methods of interpolation were considered by researchers: Cleverdon and Keen [61] defined one; Keen discussed others [151, p. 90]. In the end, one of Keen suggestion's was commonly adopted, which Keen named *semi-Cranfield*, sometimes also called *neo-Cleverdon* Williamson et al. [285]. Here, the interpolated value of precision measured at any recall level (r_i) was defined to be the highest precision measured at any level (r') greater than or equal to r_i. Manning et al. [172]

Rank	Rel	Pr	Rcl
1	1	100%	20%
2	0		
3	1	67%	40%
4	0		
5	1	60%	60%
6	0		
7	0		
8	0		
9	0		
10	0		
∞	10	%	80%
∞	1	0%	100%

Rank	Rel	Pr	Rcl
1	0		
2	0		
3	0		
4	0		
5	0		
6	1	17%	33%
7	0		
8	0		
9	1	22%	67%
10	1	30%	100%

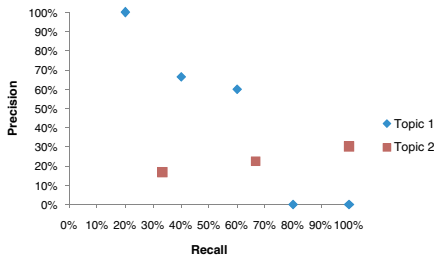

Fig. 2.3 Recall and Precision calculated and plotted for ranks resulting from two topics. In the first topic, there are five relevant documents, two of which were not retrieved; in the second topic, there are three relevant, all of which were retrieved in the top ten of the ranking.

formulated Keen's description thus[5]:

$$P_{interp}(r_i) = \max_{r' \geqslant r_i} P(r').$$

The result of the interpolation method on the points of the two topics in Figure 2.3 can be seen in the graph in Figure 2.4. The values of precision for each function were averaged at a series of pre-chosen recall levels, commonly eleven levels from 0% to 100%; although researchers also used ten levels (dropping the 0%), three levels (25%, 50%, 100%) and twenty-one recall levels (0%, 5%, 10%, 15%, ..., 95%, 100%). The resulting graph of precision averaged across both topics is shown in Figure 2.5. Note, it is common to draw such a graph with a simple interpolated line drawn between the averaged points.

By measuring the precision of every relevant document including those that were not retrieved (implied by measuring precision at recall 100%), there was an assumption in the design of this measure that users were interested in achieving such a high level of recall.

Precision at each of the standard recall levels can itself be averaged and is referred to as *interpolated average precision* or sometimes n-point average precision, where n is the number of recall levels. For

[5] Note, this interpolation measure is sometimes mistakenly thought to be the maximum precision measured between the recall levels r_i and r_{i+1}. See Harmandas et al. [108] and Baeza-Yates and Ribeiro-Neto [18, Section 3] as examples of researchers who made this error.

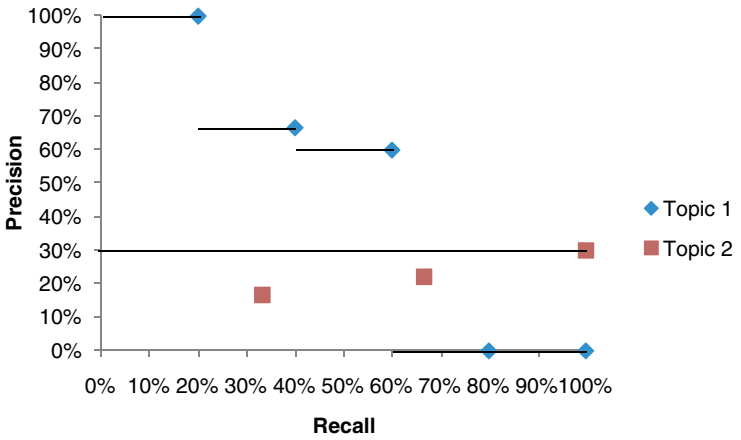

Fig. 2.4 Recall precision graph with interpolation.

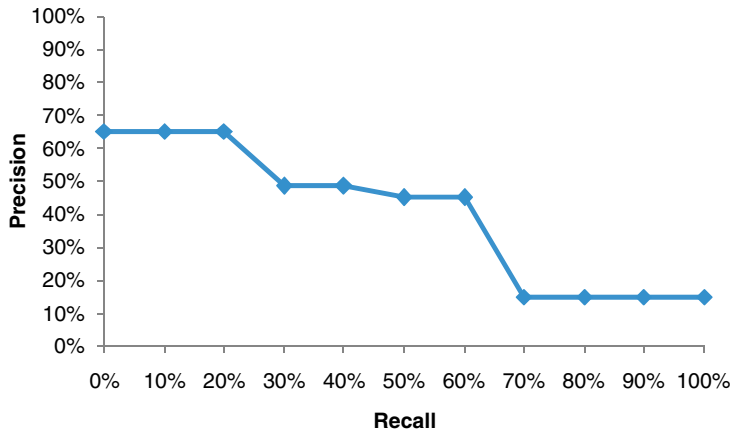

Fig. 2.5 Recall precision graph with precision averaged over two topics.

example, from the graph in Figure 2.5 one could compute 11-point average precision.

Recall precision graphs were a common form of reporting effectiveness: in his book, Salton mentioned little else; in Van Rijsbergen's evaluation section, [262], much space was devoted to the ways of computing such graphs.

At first glance, it might appear that the interpolation was an unusual choice as it ensured that the continuous function could only be

a flat or a monotonically decreasing line. Williamson et al. [285] stated that it was chosen as the standard used by the SMART retrieval system [211]. Ten years later, and it would appear quite independently, Van Rijsbergen also declared Keen's interpolation as the most appropriate to use [262, Section 7]. Both Van Rijsbergen and Williamson stated that they preferred this method over others as it was a conservative interpolation that did not inflate the values of precision for a topic. For topic 1 in Figure 2.4, this would appear to be the case; however, for topic 2, the interpolation would appear to be anything but conservative.

Keen appeared to explain this feature of the interpolation by stating that it computed "*the theoretical maximum performance that a user could achieve*". Van Rijsbergen, in a later personal communication stated that his reasoning for choosing the interpolation was one of normalization against potential errors. The general trend of precision recall graphs was that of a monotonically decreasing line, the increasing precision of topic 2 went against that trend and so should be viewed as an error to be normalized. Van Rijsbergen also stated that at the time, retrieval systems ranked documents based on similarity scores with a coarse granularity. Often, sets of documents were assigned exactly the same score. The order that such documents were ranked by was commonly the order the documents were stored in the IR system; Cooper [65] described this form ranking as a *weak ordering*. Van Rijsbergen's concern was that these blocks of weakly ordered documents caused many of the increases in precision seen for topics. Consequently, he chose the interpolation function as it would normalize such increases.

2.3.2 Another Early Evaluation Measure

Cooper was interested in evaluating ranked retrieval using a single evaluation measure, but was not convinced that finding every relevant document was always the priority of searchers. In his 1968 publication, he stated "*most measures do not take into account a crucial variable: the amount of material relevant to [the user's] query which the user actually needs*"; he went onto say "*the importance of including user needs as a variable in a performance measure seems to have been largely*

overlooked". He proposed a measure called Expected Search Length (ESL) [65], which determined the amount of a ranking that had to be observed by a searcher in order to locate a pre-specified quantity of relevant documents. Cooper was aware that the ranked retrieval algorithms of the time commonly sorted retrieved documents into weak orderings with large numbers of documents being given the same score. Consequently he ensured the measure provided an effectiveness score that accounted for these blocks of equally retrieved documents. Although relatively un-used by researchers at the time, ESL was later influential, most notably in the cumulative gain family of measures from Järvelin et al. described in Section 4.2.1.

2.4 Challenging the Assumptions in the Early Collections

There were a number of common themes to the test collections created in the first few decades of IR research, particularly regarding topics and the definition of relevance. Topics tended to be sentence length statements that mimicked the types of information request issued to librarians. Although not explicitly stated in the literature at the time, there was an assumption that users would query a computer-based IR system in the same way they used the service of a librarian: with a written detailed natural language information request. Early on some researchers pointed out that the assumption was often wrong: Fairthorne stated that users could issue a query that was "*an exceedingly ambiguous phrase*" [83]. However, this disparity between test collection queries and actual user queries wasn't addressed for some considerable time (see Section 4.1 for more detail).

From the initial testing work of Cleverdon and Thorne and of Gull, relevance was assumed to be a form of *topical relevance*, where a user would judge a document as relevant to an information need if its content either partially or wholly addressed the need. Further, search engine users were assumed to be people who would want to find as much relevant information on a particular topic as possible.

Although this straightforward notion of relevance persisted in the test collections that were built, many researchers were aware early on of the potential limitations of these assumptions. Verhoeff et al. [264]

stated that it was highly likely that different users could issue the same query but consider different documents as relevant. They proposed that retrieved documents should be ranked based on the probability of a population of users judging those documents as relevant: the more users who considered a particular document relevant, the higher its rank. Goffman, then considered the interdependence of relevance, pointing out that a document may not be viewed as relevant if retrieved documents containing the same information were previously seen by the user [95]. Such pioneering views on the importance of considering diversity in relevance and redundancy of information was only taken up in earnest much later; see Section 4.1.2.

Cooper [66] proposed that there should be a distinction between topical relevance (in his paper he referred to this as *logical relevance*) and what he called *utility*. He stated that ultimately an IR system needs to be useful and while it is possible to conceive of systems that retrieve a wide range of documents that have some level of topical relevance to an information need, the more important question to ask was which of those documents were actually useful to the user? Cooper pointed out that the credibility of a source or the recency of a document might be important factors in determining the utility of a relevant document to a user. Later [67] he suggested that writing style or even a human assessment of document quality could be a factor in utility. See Saracevic for a broader survey [224] of relevance research.

With hindsight, it can be seen that such suggestions were important features to consider when designing both test collections and search engines. However, it was many decades before such ideas were put into practice and in the intervening period the challenges made to the core assumptions of test collections were largely ignored.

2.5 Assessor Consistency

One of the primary concerns about test collection formation was that the relevance assessments made to form qrels were subjective judgments made by individuals rather than objectively true decisions. An early critic of test collections, Faithorne [83] argued for relevance judgments made by groups rather than individuals. Katter stated "*A recurring*

finding from studies involving relevance judgments is that the inter- and intra-judge reliability of relevance judgments is not very high" [145]. With such low levels of agreement, the concern was that effectiveness scores calculated from test collections using a single set of judgments were not accurate or representative. See also Burgin [41] who detailed past studies on the wide range of influences shown to affect assessment, including the order and way in which documents are presented, the definitions of relevance used, instructions given to assessors, the experience of assessors, etc.

Lesk and Salton [214] studied the question of assessor consistency by gathering pairs of relevance judgments for a test collection composed of 1,268 abstracts and 48 queries: judgments from the creator of the query were paired with those of a "subject expert" who assessed documents independently. The researchers evaluated three configurations of a search engine using different combinations of the paired judgments, determining which configuration was the best. The conclusion of their work was that regardless of the judgments used, the ranking of the different versions of the engine always came out the same. Lesk and Salton analyzed the reasons for this consistency and found that although on average, assessor consistency was low; the disparity between assessors was largely to be found for lower ranked documents. They stated that the reason for this result was that top-ranked documents tended to be most similar to the query, therefore judgments about such documents were easier to make. Most of the effectiveness measures used to assess search engines were more influenced by the rank position of top-ranked documents, therefore the ranking of the three configurations tested was consistent across the different qrel sets.

A similar experiment was conducted by Cleverdon [59] who worked with the Cranfield II collection and its multiple relevance judgments. Like Lesk and Salton, Cleverdon found variations in assessments but also found that they did not impact on the ranking of different configurations of a retrieval system. Burgin [41] also looked at a collection with multiple relevance judgments and again confirmed the Lesk and Salton result, though he cited one work (that we have been unable to locate) that was said to show variations across assessors could impact on ranking of systems. Harman [100] mentioned briefly

an assessor consistency experiment that she reported showed an 80% overlap in relevance judgments.

As a final footnote to this section, it is worth noting that the work here focused on *absolute* judgments of relevance. Rees and Schultz [193], examined the consistency of users at making judgments of documents *relative* to each other. In this test, they reported "*It is evident that the Judgmental Groups agree almost perfectly with respect to the relative ordering of the Documents.*" (p. 185). This early important observation was noted by a number of other researchers, but little work on capturing or exploiting relative judgments was reported until recently, see Section 6.3.4.

2.6 The Practical Challenges of Creating and Using Test Collections

In the early years of IR research, there were a series of practical challenges that faced test collection builders.

2.6.1 The Challenge of Obtaining Documents

The only digitized materials widely available for collection construction were catalogue information about document collections. It would appear that obtaining large quantities of full-text was virtually impossible. The only early test collection with complete documents was the Time collection built by Salton's group. It would appear that Salton got students at his institution to transcribe news articles from copies of the actual magazine. The issues used were preserved by researchers at NIST in the United States and are pictured in Figure 2.6. Later, Salton and Yu implied another collection, (MEDLARS) was also created through manual transcription [215].

2.6.2 Distribution of Collections

Although IR researchers were keen to share test collections, the practicalities of sharing could be challenging. Although in the 1970s and 1980s many institutions were connected to each other via some form of network, across the world, a range of different protocols were used to

Fig. 2.6 The copies of TIME magazine, the articles of which were manually transcribed to form the TIME test collection.

transfer data. Broad adoption of the Internet's TCP/IP beyond North America did not occur until the late 1980s. Removable storage devices such as magnetic tapes were a data transfer option, but a range of formats existed and often only a few devices to read each format were found in an organization such as a University. Because of these obstacles, sharing of collections between research groups was patchy and ad hoc.

No broadly applicable solution to distribution was found until the early 1990s, when the first large-scale distribution of test collections was achieved with the creation of the Virginia Disc One: a CDROM containing many of the commonly used test collections [86]. Several hundred copies of the discs were distributed world-wide.[6] With the increased ubiquity of networks and data transfer protocols, by the early 1990s, networked-based distribution of collections started; an early example of which is the University of Glasgow IR group's test collections web page, created in 1994 by Sanderson.[7]

2.6.3 The Challenge of Scale — Limited by qrels

Commercial IR systems were by the early 1960s searching the subject keywords of several tens of thousands of documents [78]. By the mid-1970s, it is recounted that searching hundreds of thousands of documents was routine [23], yet the test collections of the time were orders of magnitude smaller. A key reason for this appears to be researchers'

[6] A personal communication to the author.
[7] http://ir.dcs.gla.ac.uk/resources/test_collections/ (accessed April 2010).

insistence on knowing the total number of relevant documents for a query.

Cleverdon, when building Cranfield II, employed people to manually scan the full collection for all relevant documents for all topics. Others investigated less resource intensive approaches. Kent et al. [155] and later Bornstein [32] proposed collection sampling to locate some of the missing relevant and estimate the numbers remaining unfound. Maron et al. [173, p. 79] described a method of using multiple queries generated by searchers to create a set of retrieved documents that were then assessed for relevance, this approach was also suggested by Salton [210, p. 294]. Lancaster [158] used subject experts to both search and draw on their knowledge of existing documents to build up what he called the "recall base". Fels [84] stated that Mooers [180] proposed creating a form of known item search test collection. The methodology, which Fels tested, involved randomly sampling documents from a collection and creating topics that would be highly relevant, relevant or irrelevant to the sampled document. These topics would then be issued to a searching system and tests of success or failure would be determined by the presence or absence of the known items.

Despite a plethora of suggestions listed above, however, there appears to be little evidence of researchers actually trying these suggestions to build bigger test collections. Neither does there appear to be a willingness to give up on the notion of finding all relevant documents as advised by Cooper [67]. The conclusion amongst IR researchers at the time was that a way had to be found to produce larger test collections while at the same time locate as many relevant documents as possible.

Spärck-Jones and Van Rijsbergen proposed [245] a methodology for building larger test collections (what they referred to as an *ideal test collection*). Their proposal was motivated by concern that existing test collections were not only small but often carelessly built and/or inappropriately used by researchers (p. 3 of their report). They proposed the creation of one or more large test collections built using well-founded principles and distributed to the community by a common organization. Addressing the problem of assessors not being able to judge every document in large collections they proposed a solution using a technique they referred to as *pooling*.

Spärck-Jones and Van Rijsbergen suggested that for a particular topic, assessors judge the documents retrieved by "*independent searches using any available information and device*" [245, p. 13]. Pooling would create a small subset of documents containing a sufficiently representative sample of relevant documents. The relationship of pool size and its impact on the accuracy of comparisons between retrieval systems was analyzed later in some detail by Spärck-Jones and Bates [244, p. A31]. In order to manage the number of relevance judgments needing to be made, the later report also described random sampling from a pool (see pp. 20–21).

The impact of the ideal test collection report was initially limited. Some further small collections were built using exhaustive relevance judgments, such as the CISI [85] and the LISA collections [76]. Some collections were built using pooling but did not appear to be aware of Spärck-Jones and Van Rijsbergen's work. Katzer built the INSPEC test collection, composed of 12,684 abstracts [146, 147]. Katzer stated that 84 topics were created for the collection and relevance judgments made on a pool of documents composed of the union of seven searches conducted by search intermediaries. Later Salton augmented the pool with the ranked document lists of two retrieval systems configured to use different ranking algorithms [213, p. 1030]. Although earlier evaluation was conducted using approaches similar to pooling, such as Lancaster's MEDLAR tests, INSPEC appears to be the first shared test collection built using pooling. Fox described another collection, the CACM, composed of 3,204 documents where Katzer et al.'s seven search variations were used to build a pool for the collection [85].[8] Fuhr and Knorzs built a 300 query, 15,000 document collection with pooling [91]. Blair and Maron [25] later Blair [24], constructed a test collection for estimating the true recall of a Boolean search engine, using a series of broad searches to locate as many relevant documents as possible. None of this work cited Spärck-Jones and Van Rijsbergen.

Details of these somewhat larger collections are provided in the following table. By contrast, commercial search engines were by that

[8] Note, the literature is a little confused on how the CACM collection's relevance judgments were formed. A later article briefly mentioned the CACM stating that all documents were examined for relevance [213, p. 1030].

time routinely searching multi-million document collections and calls were made by the industry for the research community to start testing on larger data sets [160]. A few research groups obtained such collections: researchers in Glasgow used 130,000 newspaper articles for testing a new interface to a ranked retrieval system [218]; IR researchers at NIST conducted experiments on a gigabyte collection composed of 40,000 large documents [107]; and Hersh et al. [118] released a test collection composed of around 350,000 catalogue entries for scholarly articles.

Name	Docs.	Qrys.	Year[9]	Size, Mb	Source document
INSPEC	12,684	77	1981	—	Title, authors, source, abstract, and indexing information from Sep to Dec 1979 issues of Computer and Control Abstracts.
CACM	3,204	64	1983	2.2	Title, abstract, author, keywords, and bibliographic information from articles of Communications of the ACM, 1958–1979.
CISI	1,460	112	1983	2.2	Author, title/abstract, and co-citation data for the 1,460 most highly cited articles and manuscripts in information science, 1969–1977.
LISA	6,004	35	1983	3.4	Taken from the Library and Information Science Abstracts database.

Spärck-Jones and Van Rijsbergen's *ideal test collection report* is often cited for its introduction of the idea of pooling, however, the researchers had more ambitious goals. On page 2 of the report can be found a series of recommendations for the IR research community:

(1) "*that an ideal test collection be set up to facilitate and promote research;*

[9] Year is either the year when the document describing the collection was published or the year of the first reference to use of the collection.

(2) *that the collection be of sufficient size to constitute an adequate test bed for experiments relevant to modern IR systems...*
(3) *that the collection(s) be set up by a special purpose project carried out by an experienced worker, called the Builder;*
(4) *that the collection(s) be maintained in a well-designed and documented machine form and distributed to users, by a Curator;*
(5) *that the curating (sic) project be encouraged to, promote research via the ideal collection(s), and also via the common use of other collection(s) acquired from independent projects.*"

This vision of larger test collections built by a curating project that fostered their use in research was finally realized in the formation of TREC.

3

TREC and Its Ad Hoc Track

In 1990, the US government agency DARPA funded the National Institute of Standards and Technology (NIST) to build a large test collection to be used in the evaluation of a text research project, TIPSTER. In 1991, NIST proposed that this collection be made available to the wider research community through a program called TREC — the Text REtrieval Conference. The annual evaluation event started in November, 1992. Operating on an annual cycle, the multiple goals of TREC were to:

- create test collections for a set of retrieval tasks;
- promote as widely as possible research in those tasks; and
- organize a conference for participating researchers to meet and disseminate their research work using TREC collections.

In the early years, TREC organizers annually created gigabyte-sized test collections, each with 50 topics and a set of qrels built using pooling, see Voorhees and Harman [280] for a detailed history of the exercise. As can be seen in the overlap between the TREC goals and the Spärck-Jones and Van Rijsbergen recommendations, the ideal test collection document appeared to have influenced the construction

of TREC, however, its initiators, headed up by Harman, still had to instantiate them. Key to making TREC a success was their solution to gathering the independent searches that Spärck-Jones and Van Rijsbergen described.

Harman and her colleagues appear to be the first to realize that if the documents and topics of a collection were distributed for little or no cost, a large number of groups would be willing to load that data into their search systems and submit runs back to TREC to form a pool, all for no cost to TREC. TREC would use assessors to judge the pool. The effectiveness of each run would then be measured and reported back to the groups. Finally, TREC would hold a conference where an overall ranking of runs would be published and participating groups would meet to present work and interact. It was hoped that a slight competitive element would emerge between groups to produce the best possible runs for the pool.

The benefits of the "TREC approach" were that research groups got access to new test collections; and at the conference, compared their methods against others. The benefits of the approach to TREC were that their chosen area of IR research became a focus of interest among the research community. The benefit of the approach to all was that new test collections were formed annually for all of the IR community to use. Other fields of Human Language Technology used such collaborative/competitive approaches before TREC: e.g., the Message Understanding Conference [96]. However, the continued running of TREC, now approaching its third decade, is a testament to the particular success of the approach Harman and her colleagues applied to IR.

TRECs had a profound influence on all aspects of evaluation, from the formatting of test collection documents, topics, and qrels, through the types of information needs and relevance judgments made, to the precise definition of evaluation measures used. In particular, the first eight years of TREC, when the *ad hoc track* was run, established the norms on which a great deal of other TREC and broader IR evaluation work was based. Consequently that period in TREC is described here in some detail. The section starts with an explanation of how each of the three components of a test collection was created followed by a detailing of some of the tasks that TREC chose

to focus on in its early years. Finally, the evaluation measures used are explained.

3.1 Building an Ad Hoc Test Collection

The TREC ad hoc collections were built with the searching task of an information analyst in mind: a person who was given topics to search for on behalf of someone else Harman [105]. The topic given to them was well described and the analyst was expected to locate as much relevant material as possible. The topics for the ad hoc track were created by members of the TREC document assessment team at a rate of 50 per year. The numbers and exact procedures for forming the topics varied over the eight years of TREC ad hoc, Voorhees and Harman [280, p. 28]. However, certain aspects of the method remained constant. The creators of the topics would create a set of candidate topics, these were then trialed by searching on the ad hoc collections to estimate how many relevant documents each topic would return. Topics with too many or too few relevant documents were rejected [99, 278].

TREC topics were structured to provide a detailed statement of the information need that lay behind the query, which was intended to ensure that the topic was fully understood. The topics were formatted into an XML-like scheme (Figure 3.1), the structure of which varied over the years, but its main components were:

- a topic *id* or number;
- a short *title*, which could be viewed as the type of query that might be submitted to a search engine;
- a *description* of the information need written in no more than one sentence; and
- a *narrative* that provided a more complete description of what documents the searcher would consider as relevant.

Obtaining large quantities of text to build a collection involved persuading the copyright owners of a large corpus of material to allow their content to be used. Through connections with news publishers, TREC organizers obtained US and UK newspaper and magazine articles, as

```
<top>

<num> Number:  200

<title> Topic: Impact of foreign textile imports on U.S. textile industry

<desc> Description:  Document must report on how the importation of foreign
textiles or textile products has influenced or impacted on the U.S. textile
industry.

<narr>  Narrative:  The impact can be positive or negative or qualitative.
It may include the expansion or shrinkage of markets or manufacturing volume
or an influence on the methods or strategies of the U.S. textile industry.
"Textile industry" includes the production or purchase of raw materials;
basic processing techniques such as dyeing, spinning, knitting, or weaving;
the manufacture and marketing of finished goods; and also research in the
textile field.

</top>
```

Fig. 3.1 Example TREC ad hoc topic.

well as US government documents. TREC standardized the gathered documents in a similar XML scheme as used in the topics.

The documents and topics were sent out to participating groups who were given a limited time to generate and return a series of runs. Each run contained a maximum of 1,000 top-ranked documents retrieved for each of the TREC topics. The top n documents (most often $n = 100$, or more recently 50) from each run were merged into the pool to be judged. In order to make pool judgment tractable, TREC organizers sometimes had to limit the number of runs that contributed to the pool. In such situations, participating groups nominated a subset of submitted runs to be assessed.

TREC defined two types of run:

- *automatic* runs, defined as runs where no manual intervention took place between the submission of topics to a group's retrieval system and the outputting of the run.
- *manual* runs, where any amount of human intervention in the generation of search output was allowed. For some manual runs, the list of documents submitted was a concatenation of the best results from multiple queries. For details of how individual manual runs were created, see Voorhees and Harman's overview of one of the years of TREC (e.g., [99,

100, 101, 277, 278]). Although such runs appeared to have limited scientific value, TREC organizers encouraged their submission as they were found to be rich sources of relevant documents for the pool. Kuriyama et al. [156], showed the importance of manual searching in effective pool formation.

In order to be seen to be fair to all participants, TREC assessors viewed all top n documents in the pool; documents were sorted by docid so that the rank ordering of documents did not impact on the assessment. TREC organizers tried to ensure that the creator of the topic was also the assessor of its qrels. Unlike a number of earlier test collections, which had degrees of relevance, in TREC, assessors made a binary relevance judgment. They were instructed that "*a document is judged relevant if any piece of it is relevant (regardless of how small the piece is in relation to the rest of the document)*",[1] which resulted in a liberal view of what documents were viewed as relevant.

TREC, particularly, its ad hoc collections continue to have a profound impact on the academic IR community. The collections are used extensively: a search on a well-known scholarly search engine (conducted in May 2010) revealed that the phrase "TREC collection" occurred in nearly 1,210 papers; in a small survey conducted for this paper, examining 40 of the 60 papers in ACM SIGIR 2004, 28 of the papers used TREC collections (70%), 17 of which used at least one of the ad hoc collections (43%). These multi-gigabyte data sets became the de facto standard on which many new ideas were tested.

3.2 Classic TREC Ad hoc Measures

TREC was not only influential on test collections used by researchers but also on the effectiveness measures used. Although many measures were calculated by TREC organizers, three, MAP, R-precision, and MRR are commonly used in the broader community. Precision measured at a fixed rank, $P(n)$, although used before TREC was another measure adopted by the evaluation exercise and an important result comparing the properties of MAP and $P(n)$ was described for the first

[1] http://trec.nist.gov/data/reljudge_eng.html (accessed April 2010).

time at TREC meetings. The measures are detailed here. Following the chronological ordering of this review, more recent ad hoc measures are described later in Section 4.2.

3.2.1 Average Precision

In the first year of the exercise, TREC organizers calculated 11-point average precision using Keen's interpolation function. However, perhaps because weak orderings of rankings were less common by the early 1990s, this was soon replaced by a non-interpolated version. The first reference to this measure was in Harman [99], where the measure was called *non-interpolated average precision* (AP). It is defined as follows:

$$AP = \frac{\sum_{rn=1}^{N} (P(rn) \times rel(rn))}{R}.$$

Here, N is the number of documents retrieved, rn is the rank number; *rel(rn)* returns either 1 or 0 depending on the relevance of the document at rn; *P(rn)* is the precision measured at rank rn; and R is the total number of relevant documents for this particular topic. Simply, the measure calculates precision at the rank position of each relevant document and takes the average. Note, by summing over the N and dividing by R, in effect, precision is measured as being zero for any unretrieved relevant documents. This measure is similar to normalized precision [210, p. 290].

If one calculates AP for each of a set of topics and takes the mean of those average precision values, the resulting calculation is known as *Mean Average Precision* (MAP). Harman's original definition was published with a mistake, replacing the denominator R, with the number of relevant documents retrieved; a journal version of the paper contained the same error [102]. Voorhees appeared to be the first to describe the measure as mean average precision [267], though it took several years for MAP to become its universally accepted name. MAP became one of the primary measures used in many evaluation exercises as well as a large quantity of published IR research.

Note that, interpolated average precision (described in Section 2.3.1.) was often in older literature referred to as average precision or AP, which can cause confusion for the modern reader. Occasionally,

more recent papers and books appear to use interpolated AP where it would appear that the authors were unaware of the existence of the more established non-interpolated version.

3.2.2 Measuring Precision at a Fixed Ranking

A common option for measuring precision is to decide that a user will choose only to examine a fixed number of retrieved results and calculate precision at that rank position.

$$P(n) = \frac{r(n)}{n},$$

where $r(n)$ is the number of relevant items retrieved in the top n. The choice of n is often influenced by the manner of their display, P(10) being the commonest version. Sometimes the measure P(1) is used and referred to as the *Winner Takes All* (WTA) measure. Precision measured at a fixed rank has long been described in the literature. Both Salton [210], and Van Rijsbergen [262] mentioned calculating precision at a fixed rank. However, both described the measure in the context of producing graphs of precision over a range of ranks. Neither described the measure in the way it was used later on: a single value measured at one point in the ranking.

Note P(n) ignores the rank position of relevant documents retrieved above the cutoff and ignores all relevant below. Also if a topic has fewer than n relevant documents in the collection being searched, P(n) for that topic will always be < 1. However, there is little evidence these features of the measures are problematic.

However, there is one feature that is worth noting, the importance of which was described in one of the later years of TREC. Computing precision at a fixed rank ignores the total number of relevant documents there are for a topic. This number can affect the expected value of precision calculated at a fixed rank n. To illustrate, if we calculate P(10) on the ranks resulting from retrieval on two topics, one with 1,000 relevant documents, the other with 10. For both topics, the balance of relevant and non-relevant documents will change across the resulting ranks. However, in the first topic, it should be relatively straightforward for a retrieval system to place a great many relevant documents in the

top 10. For the second topic, the chances are that there will be fewer easy to retrieve relevant documents; consequently, it will be harder for a retrieval system to fill the top 10 with just relevant. In other words, IR systems are likely to get a high P(10) on the first topic and a low P(10) on the second. An improved system finding one more relevant documents for the first topic will score the same increase in P(10) as another system that finds one more relevant for the second topic, even though locating the extra relevant for the second topic was most likely algorithmically harder to achieve.

When evaluating within a particular test collection there does not appear to be any published evidence that this feature of the measure causes problems. However, there have been evaluations across two test collections, where differences in measurement arose. The effect was highlighted during the running of the *Very Large Collection* (VLC) track of TREC-6 [115]. Participating groups applied their retrieval systems to a 20 GB ad hoc collection and a 2 GB subset. Effectiveness was measured using P(20). Across all seven participating groups, P(20) was higher for searches on the 20 GB collection than on the subset; on average 39% higher. Hawking and Thistlewaite noted the increase, but did not study it. The following year the track was re-run, using a larger 100 GB collection along with 1% and 10% subsets, a similar increase in P(20) was noted [112].

Reasons for the increase were discussed at the TREC meeting. Consequently, Hawking and Robertson [114] studied the results in detail, postulating a number of hypotheses. They concluded that the core reason was that searching on a larger collection resulted in there being more relevant documents per topic and a consequent increase in the number of such documents that could be highly ranked. As a means of final confirmation, Hawking and Robertson examined the effectiveness of the systems participating in the VLC track using MAP, which measures precision across all retrieved and un-retrieved relevant documents. With this measure, no increase in effectiveness was observed.

Given that P(20) behaved completely differently from MAP, one might ask, is one measure better than the other? Hull discussed the qualities of the two approaches [125]. He pointed out that the answer to which is the best, depends on how users are going to use a search

engine. If the user is (like most web searchers) focused on obtaining a few relevant documents and examined only the first page of results, then a fixed rank version of precision seems more appropriate. If a search engine was able to increase the size of the collection it retrieved over, such users would view the resulting increase in relevant documents in the first page of a search engine as a clear improvement.

The situation would be different if the users of the system were, for example, patent searchers whose goal was to locate every relevant document in the collection. When the collection being searched was increased in size, such searchers would probably value the growth in the total number of relevant documents, but they might not view the engine as having improved because across all the relevant documents viewed by such a thorough searcher, the proportion of non-relevant to relevant would be unchanged.

Here, the rank cutoff version of precision appears to be the better choice in most situations. As will be seen later in Section 6.4, it would be rash to assume one version is always better than another: when the measures are compared in other contexts or used for alternate purposes, different conclusions are often drawn.

3.2.3 R-precision

Instead of calculating precision over the same fixed rank for a set of topics, one could use a different cut off for each topic; R-precision uses this approach. It is calculated as P(R), where R is the total number of relevant documents in the collection for a particular topic. Precision is calculated at the rank position where a perfect retrieval system would have retrieved all known relevant documents, a more consistent recall level than a fixed rank. The measure was first used in TREC-2 ([99], Appendix A).

Note that at the point R, the number of relevant documents ranked below R will always equal the number of non-relevant documents ranked above R, which has led others to refer to R as the *equivalence number* and called R-precision *missed@equivalent* [189]. Note also that all forms of AP and, as pointed out by Aslam et al. [14], R-precision approximates the area under a recall precision graph.

3.2.4 Searching for a Single Relevant Document

Known item search describes retrieval based on topics that have one relevant document in the collection being searched. It was first described in Thorne's original test collection paper [257]. Mean Reciprocal Rank (MRR) was created by Kantor and Voorhees[2] [143] to assess such retrieval. The measure calculates the reciprocal of the rank of the first relevant document in a ranking.

The Reciprocal Rank (RR) calculated over the four example rankings shown in Figure 3.2, is respectively, 1, 0.5, 0.5, and 0. Note how the measure is particularly sensitive to small changes in the location of the relevant document at top ranks: the RR for the second example is half of that of the first. Conversely, the measure is insensitive to large difference in low rank. Because any other relevant documents in a ranking are ignored, the RR is the same for the second and third examples. The MRR has been used in some evaluations, for example, it was used in a known-item search task in TREC [279].

3.2.5 Standardizing Measure Calculation

The organizers of TREC recognized that another important role it could play was to act as a supplier of a standard tool to calculate the effectiveness of a retrieval run. Thus it did with the public release of *trec_eval*: an application that, given a run and a set of qrels, calculates an extensive range of effectiveness measures over the run. The tool is used by many research groups and is generally viewed as holding

Rank	**Rel**	**Rank**	**Rel**	**Rank**	**Rel**	**Rank**	**Rel**
1	1	1	0	1	0	1	0
2	0	2	1	2	1	2	0
3	0	3	1	3	0	3	0
4	0	4	0	4	0	4	0
5	0	5	1	5	0	5	0

Fig. 3.2 Four example documents ranks.

[2] See also Kantor and Voorhees [144] for a more complete description of the work.

the definitive definition of many of the measures used by the IR community.[3]

3.3 The Other TREC Tracks and Uses of TREC Collections

Although the test collections and other associated data resulting from TREC ad hoc is the strongest legacy of TREC in the 1990s, test collections for many other documents types were developed at the same time. Voorhees and Harman [280, pp. 8–9] detailed both the names and nature of them in their book, some of the more significant tracks focused on:

- categorizing and/or retrieving streamed text, as addressed in the routing and filtering tracks [200];
- medical scholarly articles in the TREC-based genomics collection where matching to variants of gene names became a part of the search task [119];
- search across languages with English queries retrieving Spanish and Chinese documents, as covered in the cross language search tracks [230]; and
- retrieval of noisy channel data, output by OCR and speech recognizer systems, addressed in the Confusion [144] and Spoken Document Retrieval tracks [92].
- Later, within the field of distributed IR, groupings of TREC ad hoc collections were established into a set of commonly used collections by that research community. Shokouhi and Zobel [229] detail six such collated collections and the researchers who initially built them.

Some tracks established their own evaluation measures or methods — see the measures in the filtering track overview [200] or the differing methods of the interactive track [186]. However, across the majority of the TREC tracks in this period, the type of search task established in the ad hoc track was highly influential on the

[3] There does not appear to be any paper or technical report describing trec_eval. It can be downloaded from the following URL: http://trec.nist.gov/trec_eval/ (accessed April 2010).

topic design, definition of relevance, evaluation measures, and pooling methodologies used.

3.4 Other Evaluation Exercises

The success of TREC inspired many others to start similar evaluation exercises:

- CLEF[4] — The annual Cross Language Evaluation Forum focuses on search across European languages; though in recent years it diversified into other languages including Persian and some of the languages of the Indian sub-continent [33]. Search of other objects such as images has also been addressed: imageCLEF [64].
- NTCIR — The NII Test Collection for IR Systems is an evaluation exercise held every 18 months in Japan. NTCIR has focused on cross-language search for Asian languages such as Japanese, Chinese, and Korean. A particular focus was on patent search [141]. The first NTCIR evaluation exercise used a collection of the title and abstracts of several hundred thousand scholarly articles [142].
- TDT — Topic Detection and Tracking was an exercise examining the automatic detection and tracking of important emerging stories in streaming text [5].
- INEX — The INitiative for the Evaluation of XML Retrieval examines the retrieval of semi-structured documents, in particular focusing on retrieval of document fragments [157].
- TRECVID — an evaluation exercise focused on video retrieval [232].
- A number of other smaller and/or newer evaluation exercises were created, a number of which presented their work at the First International Workshop on Evaluating Information Access [219].

[4] Pronounced "clay", from the French word for key.

3.5 TREC's Run Collection

In addition to the test collections, topics and qrel sets generated each year by TREC, the runs (ranked lists of the documents retrieved for each topic) submitted by participating groups for each track were also archived. In the ad hoc track for TRECs 2–8, nearly 500 runs were archived. As will be seen in Section 6, this archive opened up new opportunities for research examining the impact of different evaluation measures and for exploring the effectiveness of test collection formation methodologies. Other evaluation exercises also archived their run data, see Sakai [206] for use of NTCIR runs and Sanderson et al. [221] for an example of use of runs from CLEF.

3.6 TREC Ad Hoc: A Great Success with Some Qualifications

TREC and its spin-off evaluations had a profoundly positive impact: providing large-scale test collections, a pooling method, evaluation measures, and other data sets to a research community that up to the formation of TREC did not appear to have the appetite or resources to build its own. The collections, particularly those from the ad hoc track, are extensively used. Ten years after the track stopped, it is still common to see the collections exploited in high impact research. The inaugural running in 2008 of the Indian languages evaluation exercise, FIRE (Forum for IR Evaluation) used collections and a topic design strongly influenced by the TREC ad hoc paradigm [170]. Because the ad hoc collections and methodologies continue to be widely used, it is worth reviewing some of the methods employed in those early years, focusing on collections, topics, and relevance.

3.6.1 Collections

The documents of test collections in this period were commonly newspaper articles. This tradition started with Salton's TIME collection, but was carried on by TREC and later other evaluation campaigns. TREC started in 1992. In late 1993, public web search engines were being created [154, p. 152]; by the summer of 1994, a large number were

in existence including the relatively well-known Lycos and Excite. However, TREC and the broader academic IR research community maintained their focus on search from almost nothing but newspaper and government documents for much of the 1990s. Using this form of content had a strong influence on the type of topics that were created for the test collections, which in itself limited the range of search tasks that were addressed by IR researchers in the 1990s. Some have criticized TREC organizers for not moving more quickly to the study of web search.

However, it is worth considering at least one of the contributing factors to this situation. The building of web and particularly enterprise collections for use in a large-scale evaluation is a difficult legal and privacy challenge. A web crawler has no automatic way of knowing the copyright status of the pages it is downloading. Just because an item is placed on a freely accessible web page does not mean the creator of that page is giving permission to others to copy and redistribute it. Although now, precedent has established that crawling and storing most web content is unlikely to be a copyright violation, in the 1990s this was not as clear. TREC took a cautious legal approach to such matters, which was undoubtedly a factor in the delay in adapting to Web tasks.[5]

3.6.2 Topic Selection

As a general rule one would expect the components of a test collection to be a sample of the documents and queries typically submitted to an IR system. When creating their original testing environment, Thorne [257] suggested sampling from a log of queries for example. The processes used for topic selection in the ad hoc track of TREC were

[5] One might view such caution as an inhibition to research. However, ignoring such concerns is not without risk: in a personal communication with the author, the head of a research group who were regularly crawling blogs (respecting conventions and protocols) hit problems when a blog owner went to the press claiming the researchers were spying, causing the head significant work in placating his University and the blog owner. A better-known example was the release of query logs from AOL in 2006. Although the logs were anonymized, sufficient session information was preserved to allow some people to be identified [21]. The consequence of this privacy failure was the sacking of two employees and the resignation of a company executive [292].

intended to obtain a representative sample. However, there was no log of existing searches on the document collections being built, therefore, topics had to be created. As described in Section 3.1, certain topics with too few or too many relevant documents were avoided; the later because of concerns that relevance assessors would be overwhelmed with documents to assess. However, by focusing only on topics that had a middling number of relevant documents associated with them, there was a danger of introducing bias into the topic sets.

Although TREC topics had a title, which mimicked a short user query, groups commonly submitted runs that were based on a combination of the topic's title and description fields. However, Rose and Stevens [204] described research based on query logs of web search engines showing that 53% of queries consisted of just one word; far shorter than the TREC topic titles of the time. As a reaction to this, the length of topic titles became shorter in subsequent years of TREC and participating groups were encouraged to submit runs based only on titles; as detailed by Voorhees and Harman [280, p. 39]. Nevertheless many still used the full-text of the topics in experiments despite their apparent lack of realism.

It is notable that while the length of topics was addressed, ambiguity was consciously avoided: in the TREC-5 overview, it was stated that "*Candidate topics were ... rejected if they seemed ambiguous*" and this approach persisted for many years, not just by TREC, but by most other evaluation campaigns. Attempts at building such a collection in the interactive track of TREC were made in the 1990s [187], however, the methodologies used there were not picked up by others in TREC or the wider search evaluation community until much later.

3.6.3 Binary Relevance

For many of TRECs early years, the focus was on topics locating as much information as possible even to the point of seeking documents with only a single relevant sentence. Sormunen [240] re-examined TREC relevance judgments for 38 topics from TREC-7 and 8 using three relevance grades: not relevant, marginally relevant and highly relevant. He reported that about 50% of the relevant documents were

marginally relevant and questioned if it was right for these commonly used test collections to be so strongly composed of this form of relevant document. While giving a talk about test collections and the TREC definition of relevance in 2000, a member of the audience who, at the time, was working for part of the UK intelligence community claimed (to me) that TREC's definition was the same as the one used by intelligence analysts in his organization. It would appear that this liberal definition of relevance was an appropriate definition for the information analyst task that TREC ad hoc was originally created for. Whether it was an appropriate definition for many of the information-seeking tasks carried out by other users is perhaps open to question.

3.6.4 Summing Up

It is worth reiterating that these qualifications are in the context of a highly successful on-going evaluation exercise. It is in many ways because of the success of TREC that the issues are highlighted here. TREC test collections particularly ad hoc were not only used but also imitated. Although as will be seen in the next section, TREC organizers and many others moved to address the problems listed here, there is a danger that other test collection creators sometimes too faithfully reproduced the early TREC approach. The user of any test collection would do well to examine overview documentation to understand fully the way the collection they intend to experiment on was built.

3.7 Conclusion

In this section, the initial development of large-scale test collections using a pooling approach for building qrels was described and the measures used to assess effectiveness were described. The innovation primarily from TREC of keeping run data and encouraging its use for a new form of experimentation was also described, before finally detailing some of the concerns over ad hoc test collection design.

4

Post Ad Hoc Collections and Measures

Although ad hoc-style test collections continue to be used and created, from the early 2000s, development of new test collections started with different document content, addressing new tasks, and beginning to employ novel evaluation measures. One of the motivations for this was a realization that TREC's ad hoc design had limitations.

An example of this appeared when some unusual results emerged from an early TREC web test collection: [110, 111]. Although the collection content was different from classic ad hoc collections, the topics were similar to those used in the past. One striking result from these collections was that link-based methods, such as PageRank, appeared to provide little or no value. Broder pointed out that many web queries were different from the classic information seeking view of search [34]. So-called navigational queries where users seek a home page did benefit from link-based methods, but they weren't present in the existing TREC web collections. Consequently, the organizers introduced different types of topics into subsequent collections, focusing particularly on locating named home pages. This was later generalized into the so-called *topic distillation* task: finding a series of home pages relevant to a particular topic.

It was now clear that testing different searching applications required different document collections and different types of searching task. In this section, the tasks and collections that were introduced in this post ad hoc period are described, followed by details of the new measures.

4.1 New Tasks, New Collections

At the same time that the TREC web track was being developed, the *Question Answering* (QA) track was created. Here the search task was for a QA system to interpret an inputted question and locate passages within documents that answered the question [269]. Searching for passages within documents was also explored in other QA tracks run in other evaluation exercises [169] as well as the novelty track of TREC [103]. Another evaluation exercise to explore search of document parts was INEX (INitiative for the Evaluation of XML Retrieval), where search of different collections of XML data was examined [90]. Here the focus was on evaluating searching systems for not only their ability to retrieve relevant structured documents, but also to locate the best point in a document's structure where a user could start reading the relevant part of the document.

Search of new web-based content was examined in the blog track of TREC, where several hundred thousand blog feeds were crawled to form collections composed of millions of postings. Here the search tasks examined a form of topic distillation to locate relevant feeds addressing a particular topic. Blog topics were sampled from a blog search engine log. In addition, detection of the opinions expressed in blogs became part of the search task [185].

Search tasks associated with organizations were explored in the Enterprise track, which examined search of email discussion threads as well as using a multi-faceted collection of documents from an organization to create a search task for location of experts in particular topics [70]. Multimodal search tasks started to be explored first off in the video track of TREC — TRECVid [233, 231] — and then in the image searching track of CLEF — ImageCLEF [64]. In both cases, topics were specified as a combination of text and examples of the media sought.

New types of high recall search were also examined. In 2002, NTCIR started a long-term examination of patent retrieval [129], considering a range of different search tasks. In this domain, location of all relevant past material was important. Aspects of patent search were more recently taken up by CLEF evaluation exercise [202]. Search supporting e-discovery was examined in the TREC legal track [22], another area of IR where the users (e.g., lawyers) wish to find all relevant items.

4.1.1 Moving Beyond Binary Relevance

From the start of test collection development, there were collections with multiple levels of relevance, Cleverdon's Cranfield I collection [58] had ternary judgments: relevant, partially relevant or not relevant. However, the qrels of most ad hoc style test collections used binary relevance judgments; see Kando et al. [142] and Oard et al. [184], as notable exceptions. This situation changed as researchers started to report that retrieval techniques that worked well for retrieving highly relevant documents were different from the methods that worked well at retrieving all known relevant documents [271].

There was also a realization that degrees of relevance were commonly being used in the internal test collections of web search companies. To illustrate, White and Morris [284, p. 256] mentioned a form of test collection within Microsoft with relevance judgments "*assigned on a six-point scale by trained human judges*". Carterette and Jones [50] described a collection within Yahoo! used for advertising retrieval with five levels of relevance ("*Perfect, Excellent, Good, Fair, and Bad*"); and Huffman and Hochster [124] described work in Google where assessors judged relevance of retrieved documents on a "*graduated scale*". Note, although the dates of these publications are more recent, they are the best examples found of the companies revealing some aspects of their testing work; it is thought very likely that this approach to relevance has been used for sometime in search companies.

Consequently, degrees of relevance become more common in test collections. For example, a ternary scheme was used in the web track of TREC [71]; degrees of relevance were used in TREC's Enterprise track as well as the Blog track.

4.1.2 Diversity

As described at the start of Section 2, the origins of IR test collections lay in methods developed for measuring the effectiveness of library cataloguing systems, where users wrote detailed information requests for librarians who would do the actual searching. However, Verhoeff et al. [264], Fairthorne [83] and Goffman [95], respectively, pointed out that users' definitions of relevance were diverse, queries were commonly ambiguous, and that the relevance of a document to a query could be strongly influenced by the documents already retrieved. However, test collections topics continued to follow the tradition of being detailed unambiguous statements of an information need for which one view of relevance was defined and the relevance of a document was assessed independent of other documents.

An early attempt to create topics with multiple distinct notions of relevance was the interactive track of TREC, which over several years, built a collection composed of 20 topics, each with relevance judgments that addressed multiple aspects [187]. This collection was widely used in diversity research. Following on from this, the novelty track of TREC [103] and the QA track [272] both encouraged the building of systems that retrieved fragments of documents (sentences for the novelty track, so-called *nuggets* for QA) that were both relevant and had not previously been seen. More recently, a collection addressing search of ambiguous person names was created [11]. Both Clarke et al. and Sanderson et al. described re-using existing test collections for diversity research: Clarke re-using a TREC QA collection [55]; Sanderson, an image search collection [221]. Liu et al. [167] described building a web test collection of ambiguous queries, where they examined search engine effectiveness against different levels of ambiguity in the queries.

This relatively small number of test collections addressing diversity is likely to change in the coming years, as web, blog, and image collections were used in the 2009 runs of TREC and CLEF.

4.2 Post Ad hoc Measures

With the new collections, tasks and relevance judgments, came a need for new effectiveness measures. Note, only the prominent measures are

described here, others exist, see Demartini and Mizzaro for a tabulation of a great many [77].

4.2.1 Grades of Relevance

Measures such as Precision can only be used with binary judgments, though using a threshold, one can map n-ary judgments to a binary scheme. Cleverdon used such a mapping in his early test collection work (the two sets of points in the graph in Figure 2.1 reflect results calculated with different thresholds). Researchers were aware that commonly used evaluation measures failed to consider the degrees of relevance and suggested alternatives: Pollock [191] conducted notable early work in this area, aspects of which are described later in this sub-section.

Assuming that one can transform relevance judgments on documents to numerical values, Järvelin and Kekäläinen [132] proposed a suite of measures that evaluated the effectiveness of a retrieval system regardless of the number of levels of relevance. Their simplest measure, *Cumulative Gain* (CG), is the sum of relevance values (*rel*) measured in the top n retrieved documents. Note, Cooper's ESL measure [65] operated in a similar way, though the number of non-relevant documents, instead of relevant, was counted.

$$CG(n) = \sum_{i=1}^{n} rel(i).$$

Examining some example rankings: in the left hand rank in Figure 4.1, CG(5) = 6. Because CG ignores the rank of documents we see that this is also the value of CG in the poorer rank on the right in Figure 4.1. However, *Discounted Cumulative Gain(DCG)*, where the relevance values are discounted progressively as one moves down the document ranking, used a log-based discount function to simulate users valuing highly ranked relevant documents over the lower ranked.

$$DCG(n) = rel(1) + \sum_{i=2}^{n} \frac{rel(i)}{\log_b(i)}$$

Järvelin and Kekäläinen suggested setting b to 2. The ranks in Figure 4.2 show the values of the discount function (in the denominator)

Rank	Rel	Rank	Rel
1	2	1	1
2	1	2	0
3	2	3	2
4	0	4	1
5	1	5	2

Fig. 4.1 Two document rankings.

Rank	Rel	Disc	Rel/Disc	DCG	Rank	Rel	Disc	Rel/Disc	DCG
1	2	1.00	2.0	2.0	1	1	1.00	1.0	1.0
2	1	1.00	1.0	3.0	2	0	1.00	0.0	1.0
3	2	1.58	1.3	4.3	3	2	1.58	1.3	2.3
4	0	2.00	0.0	4.3	4	1	2.00	0.5	2.8
5	1	2.32	0.4	4.7	5	2	2.32	0.9	3.6

Fig. 4.2 Document rankings with discount values.

Rank	Rel	Disc	Rel/Disc	IDCG
1	2	1.00	2.0	2.0
2	2	1.00	2.0	4.0
3	1	1.58	0.6	4.6
4	1	2.00	0.5	5.1
5	0	2.32	0.0	5.1

Fig. 4.3 Perfect ordering of relevant documents.

and the DCG scores for each rank position. We see that DCG(5) of the left hand rank in Figure 4.2 is 4.7 and 3.6 in the right hand poorer rank. In a follow-up paper Järvelin and Kekäläinen [133] added a third measure, *normalized* DCG (nDCG). Here DCG was normalized against an ideal ordering of the relevant documents, IDCG, see Figure 4.3.

$$nDCG(n) = \frac{DCG(n)}{IDCG(n)}.$$

The value of nDCG ranges between 0 and 1. The nDCG(5) of the left and right rankings in Figure 4.2 are $4.7/5.1 = 0.92$ and $3.6/5.1 = 0.71$. Note, Pollock worked on graded relevance, he proposed a measure that computed normalized cumulative gain [191].

Al-Maskari et al. [3] pointed out that in certain circumstances nDCG can produce unexpected results. To illustrate, for both rankings in Figure 4.4, there are only three known relevant documents, though the topic on the left has three highly relevant documents, the other has three partially relevant documents. For both topics the rankings

Rank	Rel	Disc	Rel/Disc	DCG	IDCG	Rank	Rel	Disc	Rel/Disc	DCG	IDCG
1	2	1.00	2.0	6.7	6.7	1	1	1.00	1.0	4.6	4.6
2	2	1.00	2.0	8.7	8.7	2	1	1.00	1.0	5.6	5.6
3	2	1.58	1.3	10.0	10.0	3	1	1.58	0.6	6.3	6.3
4	0	2.00	0.0	10.0	10.0	4	0	2.00	0.0	6.3	6.3
5	0	2.32	0.0	10.0	10.0	5	0	2.32	0.0	6.3	6.3

Fig. 4.4 Rankings from two different topics that result in the same *nDCG*.

are ideal, so the nDCG (DCG ÷ IDCG) in both cases is 1, which is perhaps a counterintuitive result.

Burges et al. [40] described a version of nDCG, for which the DCG component more strongly emphasized the high ranking of the most relevant documents:

$$DCG(n) = \sum_{i=1}^{n} \frac{2^{rel(i)} - 1}{\log(1+i)}.$$

A series of other graded relevance measures were proposed in recent times: Sakai [205] detailed and compared the properties of a number of them including Average Weighted Precision (AWP) and Q-measure. The measures were used by NTCIR organizers as graded relevance was a common feature of the test collections produced by that evaluation exercise. Järvelin and Kekäläinen reviewed many other proposed measures [133].

Moffat and Zobel [179] argued that the log-based discount function in DCG was not the best model of users' behavior when browsing a ranked list of documents. They constructed a model based on the probability p that a user progresses from one document in the ranking to the next. A high p models a persistent searcher; a low p models a fleeting one. The probability was incorporated into a geometric discount function forming the Rank-Biased Precision (RBP) measure

$$RBP = (1-p) \cdot \sum_{i=1}^{d} rel(i) \cdot p^{i-1},$$

where d was the depth of rank one wished to compute the measure over.

Although it is rarely discussed in the literature, when one has grades of relevance in a set of qrels, one could view measuring the effectiveness

of a retrieval system as a comparison of ranked lists: the retrieved ranking compared with the qrels ranked by their relevance. This idea was suggested by Pollack [191]. Joachims implemented the idea using Kendall's τ [153] to measure the correlation between the two ranks so as to obtain a measure of effectiveness [137].

4.2.2 Managing Unjudged Documents

When pooling approaches were introduced into test collection formation, the relevance judgments that accompanied the collections were composed of lists of documents that were judged relevant or not relevant. There was, however a third class of document: the unjudged, documents not in the pool. An early mention of unjudged documents can be found in Hersh et al. [118], where the researchers stated that they chose to ignore such documents when calculating effectiveness measures. A more common approach was to assume that unjudged documents were not relevant. Büttcher et al. [43] pointed out that in many situations this is a sensible approach.

However, Buckley and Voorhees [38] stated that there are other situations where unjudged documents were a potential problem. They were concerned that the size of pools relative to the size of collections was reducing as test collections grew. A related concern was that some test collections were created from sets of documents that were subsequently updated. If no new judging was done after an update, the pool would effectively reduce in size relative to the collection. Therefore, Buckley and Voorhees considered if a new evaluation measure could be devised that better estimated the effectiveness of a system when there were a large number of unjudged documents. They devised BPref, so-called as it "*uses binary relevance judgments to define the preference relation*". It is defined as follows:

$$BPref = \frac{1}{R}\sum_{r}\left(1 - \frac{|n \text{ ranked higher than } r|}{\min(R, N)}\right),$$

where R is the number of documents judged relevant for a particular topic; N is the number of documents judged not relevant; r is a relevant retrieved document; and n is a member of the first R irrelevant retrieved

documents. The measure considers a bounded number of judged non-relevant documents, determining the fraction of these documents that are ranked higher than r. The numerator captures relevance in terms of preference (n ranked higher than r). Although not the first measure to consider preference — see Frei and Schäuble's usefulness measure [88] — BPref is the first commonly used measure to which preference judgments could be easily applied.

In their 2004 paper, Buckley and Voorhees stated that the formulation of BPref was arrived at empirically after a number of experiments. Note the definition shown is from Soboroff's paper [237]. It supersedes the original definition for BPref and its variation BPref-10 and is the default definition of BPref used in recent years by TREC and its trec_eval system (from version 8 onwards). However, such is the popularity of the earlier paper the previous definitions persist in the research community.

Buckley and Voorhees devised a series of simulated experiments comparing the *stability* of BPref with P(10), R-precision, or MAP. They stated that BPref's greater stability was due to its ignoring the increasing numbers of unjudged documents, when the other measures treated these documents as not relevant.

Soon after, Yilmaz and Aslam [289] described a number of alternative effectiveness measures also built to handle un-judged documents including *induced AP* (indAP) and *inferred AP* (infAP). In Bpref, Buckley and Voorhees's aim was to create a measure that mimicked MAP as closely as possible, which was also Yilmaz and Aslam's aim. Unlike BPref, however, their two AP measures were more directly related to the formulation of MAP and in the presence of complete relevance information, resulted in the same score as MAP. Their second measure, infAP is the more widely used of the two and is described here in more detail.

Yilmaz and Aslam split the unjudged documents of a run into two sets: based on whether the documents would or would not have contributed to the test collection's pool had the run been used to build the collection. For the later set, the unjudged documents were assumed to be non-relevant; for the former set, infAP calculated the proportion of judged relevant and non-relevant documents in the document ranking

for that topic and assumed that this proportion was the probability that unjudged documents were relevant. For example, in most of the TREC test collections, pools were formed from the top 100 documents of each submitted run. For such a collection, infAP measured at rank position k would be formulated as follows:

$$infAP(k) = \frac{1}{\mathrm{R}} \sum_{r} \left[\frac{1}{k} + \frac{(k-1)}{k} \left(\frac{|d100|}{k-1} \bullet \frac{|rel| + \varepsilon}{(|rel| + |nonrel| + 2\varepsilon)} \right) \right].$$

Here the definitions of R and r are the same as BPref; $|d100|$ is the number of judged documents found above rank k plus the number of unjudged documents above rank k that would have contributed to the document pool; $|rel|$ is the number of documents above rank k that are judged relevant; and $|nonrel|$ is the number above k that were judged not relevant and ε is a smoothing constant.

In similar stability experiments to those conducted by Buckley and Voorhees, Yilmaz and Aslam showed that all of their new measures, in particular infAP, were substantially more stable than BPref. A number of evaluation campaigns adopted infAP using it in conjunction with pool sampling to streamline their relevance assessment process: TRECVID started in 2006 [188] as did the TREC Terabyte track [42]. For more on pool sampling see Section 6.3.1.

These were not the only example of such effectiveness measures, a tranche of similar measures were proposed and further analyses conducted. Sakai [207] tested a number of alternatives including one that considered graded judgments (RPref). Büttcher et al. [43] described their measure, RankEff, which inferred the relevance of an unjudged document based on its textual similarity to judged documents. A number of variants directly inspired by infAP were described in the literature, statAP, which is used in the Million Query track of TREC is probably the best known [51]. Bompada et al. [27] compared BPref, infAP, and nDCG under a wide range of situations where qrels were incomplete. They found nDCG (which simply ignores unjudged documents) to be the most stable measure. See also Sakai and Kando [209] for another detailed comparison of these types of measures.

Probably because it was the first such measure, BPref started to be used quite widely in the IR research community, however, given more

recent research questioning its stability compared to alternatives, this popularity may be brief. There is not yet a sufficiency of definitive work on which alternative is best.

A different approach to dealing with unjudged documents was suggested by a number of researchers: assessing potential error in a measure. Baillie et al. [20] proposed that the number of unjudged documents retrieved should be considered when making comparisons between runs. They found that if the number of unjudged between such runs was different, there was a danger that comparisons were unreliable. Moffat and Zobel similarly examined error rates when comparing systems on collections with unjudged documents [179].

4.2.3 Relevance Applied to Parts of Documents

Most IR evaluation focuses on retrieval of whole documents. It is to those whole units that judgments of relevance or non-relevance are applied. In early IR research, the "documents" being retrieved tended to be at most abstract-sized texts. However, as full-sized documents started to be retrieved, passage-based retrieval was increasingly studied and effective means of its evaluation was considered. Initial work [117, 212] focused on using passage retrieval to improve document retrieval, which meant that existing document test collections could be used unaltered. In passage retrieval, the aim was to identify accurately the passages of a document that were relevant to a user's request. Consequently, an adapted form of test collection was built with more detailed information in the qrels on the location of relevant passages.

Passage retrieval was part of the tasks in the TREC HARD track and was an integral part of the INEX evaluations of XML retrieval. A broad range of evaluation measures were developed within HARD [281] and INEX [149, 148, 140]. Many of the measures were extensions of existing approaches to evaluation of document retrieval, such as precision, recall, MAP, and DCG. Many of the measures were task specific and no single measure emerged that is used more than others or used beyond the evaluation exercises that created it. Here we illustrate the working of one of these measures taken from Kamps et al. [140]. Here,

document passages (called parts in INEX) p_r were ranked at positions r to collectively form a ranking retrieved in response to a query q. Precision at rank r was defined as:

$$P(r) = \frac{\sum_{i=1}^{r} rsize(p_i)}{\sum_{i=1}^{r} size(p_i)},$$

where $rsize(p_i)$ was the length of any segment of the document part that was judged relevant and $size(p_i)$ was the total number of characters in p_i. Recall at rank r was:

$$R(r) = \frac{\sum_{i=1}^{r} rsize(p_i)}{Trel(q)},$$

where $Trel(q)$ was the total quantity of relevant text in the collection for the query q. Kamps et al. [140] went on to describe an interpolated version of P to allow precision recall graphs to be plotted and an MAP-like measure MAiP to be calculated.

4.2.4 Dependence and Diversity in Rankings

All the evaluation measures described so far assumed that users judged the relevance of a document independent of all other documents. Many measures discounted the importance of documents retrieved lower down the ranking, but the level of discount was always determined by rank and not the quantity of relevant documents already retrieved. Measures that took a dependent view of relevance were developed in the context of diversity and novelty. For diversity, coverage of different aspects of relevance in rankings and individual documents was a priority. For novelty, the goal was to score higher IR systems that prevented repetition of the same relevant content in a ranking.

The initial work in measuring the effectiveness of diverse retrieval systems appears to be from the TREC-6 interactive track [186]. Documents relevant to the collection topics were expected to be relevant to one or more distinct *aspects*,[1] which Over stated were "*roughly one of*

[1] In later TRECs aspects were called *instances* [121].

many possible answers to a question which the topic in effect poses". When assessing runs submitted to the track, *aspectual precision* and *aspectual recall* were calculated. Respectively, Over defined them as "*the fraction of the submitted documents which contain one or more aspects*" and "*the fraction of total aspects ... for the topic that are covered by the submitted documents*". As defined, aspectual precision appears to be the same as precision.

Zhai et al. [293] defined a diversity-specific version of precision and formalized Over's definition of aspectual recall, choosing to instead to call it sub-topic recall (S-recall). Considering a topic with n_A subtopics and a ranking of documents, $d_1, \ldots, d_m$, S-recall calculated the percentage of subtopics retrieved by the documents up to rank position K:

$$S\text{-}recall(K) = \frac{\left|\bigcup_{i=1}^{K} s(d_i)\right|}{n_A}.$$

Here, $s(d_i)$ was the number of subtopics covered in d_i. The measure gave a higher score to runs that covered the largest number of subtopics. Several years later, Chen and Karger [53] proposed the measure k-call(n), which counted if at least one relevant document was retrieved by rank n. By measuring effectiveness in this way, IR system designers looking to optimize for this measure would be motivated to produce systems with diverse rankings.

Both measures ignored the rank of retrieved documents. Clarke et al. [55] proposed an adaptation of nDCG (Section 4.2.1) called α-nDCG, which included this aspect in a diversity measure. The researchers re-defined the function *rel(i)* from nDCG as:

$$rel(i) = \sum_{k=1}^{m} J(d_i, k)(1-\alpha)^{r_{k,i-1}},$$

where m is the number of distinct *nuggets* (the researchers' term for aspects or subtopics), $n_1, \ldots, n_m$, relevant to a particular topic; $J(d_i, k) = 1$ if an assessor judged that document d_i contained nugget n_k;

$$r_{k,i-1} = \sum_{j=1}^{i-1} J(d_j, k)$$

was the number of documents ranked before document d_i that were judged to contain nugget n_k; and the constant α represented the probability that the user of the retrieval system observed prior relevant documents. Note, if α was set to zero and the number of distinct nuggets $m = 1$, the measure reverted to standard DCG. Clarke et al. [56] also created the NRBP diversity metric based on Rank-Biased Precision (RBP). Agrawal et al. [2] pointed out that some of the sub-topics of a query could be more popular or important than others. They found sources of information to estimate a user's probable intended meaning when entering ambiguous topics. To assess their system, they adapted a number of conventional measures (nDCG, MAP, MRR) to handle diversity and to be *intent aware*.

Chapelle et al. [52] described Expected Reciprocal Rank (ERR). While a version of the measure that deals with diversity was described, ERR could also be used simply to promote novelty in search. The measure was defined as follows:

$$ERR = \sum_{r=1}^{n} \frac{1}{r} \prod_{i=1}^{r-1} (1 - R_i) R_r,$$

where n was the number of documents in the ranking and R_i was the probability that the document at rank i was relevant. At each rank position, r, the probability that a user missed each of the relevant documents retrieved higher up the ranking $(1 - R_i)$ was used as a discount on the impact that R_r had on the final score.

The α-nDCG and intent aware precision measures were used in a recent TREC diversity evaluation track [54]; and cluster recall was used in ImageCLEF evaluations [10]. A consensus on a common diversity effectiveness measure is yet to emerge.

4.3 Are All Topics Equal?

Commonly, when an evaluation measure is defined in the literature, its formula is presented as a calculation over a single topic. It is assumed that when summarizing the values computed across a set Q of test collection topics, the arithmetic mean of the values is taken.

Alternatives have been discussed and occasionally tried. Cooper [65] suggested the use of the geometric mean and a weighted average when aggregating scores across queries for his ESL measure, but settled on the arithmetic mean. Later, Voorhees described GMAP, which uses the geometric mean of AP scores [273]. Robertson described the formulation of GMAP [198] and suggested a more computationally tractable version of the formula, which took the arithmetic mean of the log values of AP, Robertson referred to this as AL, this formula produced the same ranking of runs as GMAP though with different values. The two approaches to calculating geometric mean (GMAP, AL) of average precision (AP) values computed over a set of topics Q are as follows:

$$GMAP(Q) = \sqrt[|Q|]{\prod_{k=1}^{|Q|}[AP(Q_k) + \varepsilon]} \quad AL(Q) = \frac{1}{|Q|}\sum_{k=1}^{|Q|}\ln(AP(Q_k) + \varepsilon).$$

Note, if $AP = 0$ for any topic, GMAP becomes zero and AL undefined, therefore, Robertson added a small value ε to avoid such problems. Robertson discussed this measure in some detail pointing out that using geometric mean emphasized improvements in topics that had a low AP score. As Robertson stated...

> "*GMAP treats a change in AP from 0.05 to 0.1 as having the same value as a change from 0.25 to 0.5. MAP would equate the former with a change from 0.25 to 0.3, regarding a change from 0.25 to 0.5 as five times larger.*"

This property of GMAP caused it to be created for use in the Robust Track of TREC [273], where there was a particular focus on so-called poorly performing topics. Beyond the robust track, it is little used. Whether the method is a more effective averaging approach than the arithmetic mean is yet to be determined.

The quality of GMAP to emphasize some changes in topics over others was explored using alternate approaches. Given a large historical set of run scores for the topics of a given test collection, one can compute the score of a new run in relation to the historical scores, determining

on a per topic basis if the new run is better or worse than the past runs. Webber et al. [282] proposed such an approach, employing a methodology used in human testing called *score standardization* also known as *z-score standardization*. For each topic (t) in a collection, the score of a new run s is computed as m_{st}. The mean (μ_t) and standard deviation (σ_t) of the scores for the topic is computed from the historical data and a standardized score (m'_{st}) for s is computed as follows:

$$m'_{st} = \frac{m_{st} - \mu_t}{\sigma_t}.$$

It was long assumed that the topics of a test collection contributed equally to the measuring of the effectiveness of a search engine. Bodoff and Li [26] used Classical Test Theory (CTT) to examine how TREC ad hoc topics ranked the runs submitted to particular years of TREC. They showed how CTT identified potential outlier topics that ranked runs differently from the majority in a collection. The implication from Bodoff and Li's work was that these outliers were potentially, introducing noise into the test collection. Whether such topics were noise or important outliers was not examined by them. The same year, Mizzaro and Robertson [177] reported a study on TREC ad hoc run data looking to find redundant topics in test collections: topics that ranked runs similarly. Mizzaro and Robertson stated that such topics could be eliminated from a test collection and that one "*could do reliable system evaluation on a much smaller set of topics*". However, they only found this small set of topics through an exhaustive search of all possible combination of topics using the run data from TREC.

4.4 Summing Up

In the decade following the development of ad hoc test collections, IR evaluation research explored an increasingly wide range of collection types and search tasks. There was a re-discovery of evaluation ideas and practices described in the past, including use of logs to source topics, gathering, and measurement of graded relevance judgments and increasing consideration of diversity in search results. There was also an exploration of relatively new topics such as management of unjudged documents. As shown in this section, these topics produced an extensive

body of work. However, this was only one strand of evaluation research; work in studying statistical significance was addressed; as well as a more introspective examination of the methods that underpin the creation and use of test collections also became a major part of evaluation research. These two topics are described next.

5

Beyond the Mean: Comparison and Significance

Whichever evaluation measure one uses, the effectiveness of one run will almost always be compared to another, often considering the absolute or relative difference between the runs. However, simply considering the Δ between two values can hide important detail, which is illustrated with an examination of three runs: a, b, c.[1] The MAP for the three runs is 0.281, 0.324, and 0.373, respectively. With similar sized gaps between the runs, one might view the comparisons to be revealing similar differences. However, if one graphs a & b and b & c — plotting the AP scores across each of the 50 topics used in the collection — a more complex picture is revealed; see Figures 5.1 and 5.2. The order of topics in both graphs is sorted by the topic effectiveness scores of the b run.

It can be seen that there is great variation in AP ranging from 0.01 to 0.87. In Figure 5.1, the effectiveness of a follows a similar pattern and has a similar range of scores. Harman and Buckley [106] reported on a detailed study of run comparisons and stated that for most runs, the relatively similar performance on topics shown here is typical. Having

[1] The runs are from TREC-8, INQ604, ok8alx, and CL99XT. The later is a manual run, which probably explains the more erratic difference between it and the others in Figure 5.2.

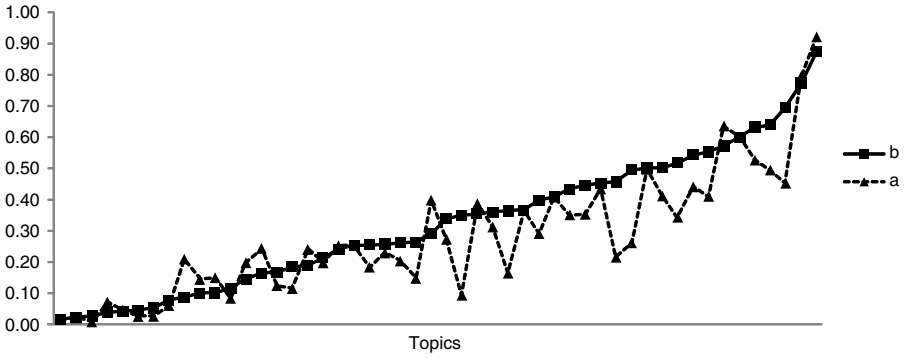

Fig. 5.1 Topic-by-topic comparison of two TREC-8 runs based on average precision scores.

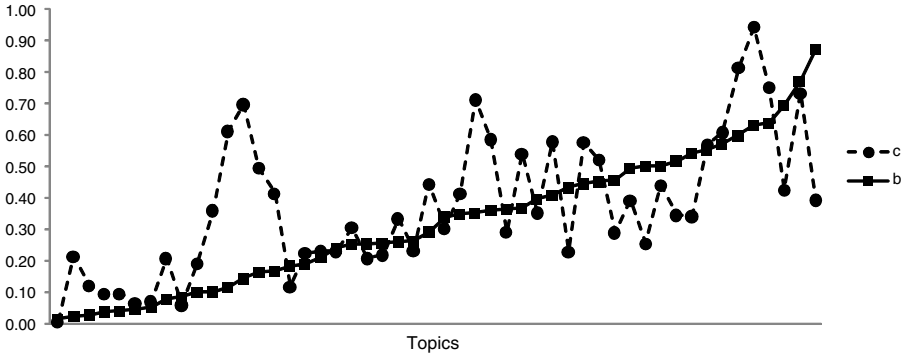

Fig. 5.2 Topic-by-topic comparison of two TREC-8 runs based on average precision scores – taken from Harman's work [104].

the worse run, in this case a, being the same or a little better for some topics (16 of the 50 topics here) is also common. The absolute difference between a and b is 4.3% and relative difference is 15.4%. Both the difference in scores and an examination of the graph in Figure 5.1, would lead most to agree that in this case system b is better.

Harman [104] highlighted the comparison shown in Figure 5.2 where it is arguably harder to determine the better run, yet the absolute and relative differences between b and c are 4.9% and 15.1% respectively, similar to the two runs above. While Harman and Buckley's work [106] showed that most run comparisons were like the case in Figure 5.1, situations such as those found in Figure 5.2 are not exceptional. Therefore,

it is necessary to examine more than the single value effectiveness measure calculated for a particular run.

5.1 Significance Tests

A common approach to more fully understanding a difference Δ measured between two runs is to use one or more significance tests. The tests estimate the probability p of observing a difference at least as large as Δ given that a so-called *null hypothesis* (H_0) is true. In the context of IR experiments, H_0 states that the systems producing the two runs under examination have effectively the same retrieval characteristics and that any difference between the runs occurred by random chance. The convention when using such tests is that if it is found that p is below a certain threshold — typically either 0.05 or 0.01 — it is concluded that H_0 is unlikely and consequently should be rejected. Although it is not universally agreed upon, the common interpretation of rejecting H_0 is to conclude that an alternate hypothesis, H_1 is true. This hypothesis states that the two IR systems have different retrieval characteristics, leading the experimenter to conclude that a significant difference was observed. The exact nature of H_1 depends on whether a *one-* or a *two-tailed* test is chosen. This topic is discussed below in Section 5.1.2.

The tests are not infallible and can make errors, which have been classified into Type I and Type II errors. Type I errors are false positives: leading the experimenter to incorrectly reject H_0 and conclude that H_1 is true; Type II errors are false negatives: leading the experimenter to incorrectly conclude that they cannot reject the null hypothesis. In IR parlance, Type I measures the precision of the test, Type II measures its recall. Different significance tests tend to produce a different balance between these two errors. For example, the sign test is known for its high number of Type II errors whereas the t-test is known for producing Type I.

The creators of the tests make assumptions on the underlying data being examined and it is important for experimenters to be aware of a test's assumptions before applying it. Tests that have fewer assumptions tend to generate more Type II errors and are said to have less *power*. So-called *non-parametric tests*, fit this profile, the best known in

IR research are the *Wilcoxon signed-ranks test* and the *Sign test*. More powerful tests generate fewer Type II errors but make more assumptions about the data being tested. If such tests, known as *parametric tests*, are applied to data in violation of such assumptions Type I errors can result. The best-known parametric test used in IR is the *t-test*.

Assuming each run is conducted on the same topics in a test collection, the significance test is usually operated as a paired test. If, less commonly, the runs being compared use different sets of topics an independent or two-sample version of the tests can be used. Although there is a wide range of tests available to the IR researcher, the three mentioned so far are the most often used. Lesk [162] discussed use of the sign and the t-test in IR experiments and Ide used the Wilcoxon and t-test to examine the significance of retrieval results [127].

At the same time, Saracevic urged caution on the use of a wide range of statistical tests. He found that no parametric statistical test could be used with confidence on data emanating from effectiveness evaluations because such "*data does not satisfy the rigid assumptions under which such tests are run*" [223, p. 13]. In addition, he pointed out that conditions set for use with non-parametric tests were also likely violated by the data output from test collection evaluations. Saracevic did not suggest that these problems should be viewed as being the end of the matter, instead he called for studies on the applicability of statistical tests in IR evaluation. Later Van Rijsbergen detailed the problems with using these test [262, Section 7]. For the t-test, results needed to be drawn from a "*normally distributed population*", which Van Rijsbergen stated did not occur with the output of test collection based retrieval experiments. For Wilcoxon and Sign, he pointed out these tests could only be used if data was drawn from continuous distributions; yet retrieval experiments produce discrete distributions. Despite the violations, Van Rijsbergen suggested "*conservative*" use of the Sign test.

In later work, Hull [125] suggested that such prudence was most likely excessive. Hull argued that with "*sufficient data, discrete measures are often well-approximated by continuous distributions*". Further, he stated that "*the t-test assumes that the error follows the normal distribution, but it often performs well even when this assumption is violated*". Hull went on to discuss the properties of a range of

significance tests, including the three examined by Van Rijsbergen as well as variations of the ANOVA and Friedman tests. Robertson had earlier conducted a theoretical analysis of a set of tests [196], estimating the topic set sizes needed in order to obtain statistical significance in retrieval experiments. He examined the Mann–Whitney U-test, the Chi-squared and the t-test. However, neither Hull nor Robertson empirically tested their comparisons.

Empirical examinations of the tests appeared to have been first conducted by Keen [152] who described a small-scale comparison of the Sign and Wilcoxon tests, the results of which showed the tests gave "*a similar picture*". Keen stated that Wilcoxon tended to indicate significance more readily. Zobel described a study of which of three significance tests was the best suited for IR experiments [295]. Splitting the topics of a test collection in half: one-half used as a mini test collection; the other as a simulation of an operational setting, he tested the paired t-test, the paired Wilcoxon's signed rank test and ANOVA (his method is described in detail in Section 6.4). Zobel concluded that use of all three tests resulted in accurate prediction of which was the best run, though he expressed a preference for the Wilcoxon test "*given its reliability and greater power*".

Later, Sanderson and Zobel [222] using Zobel's [295] methodology examined the t-test, Wilcoxon and Sign tests. Their conclusions were that use of the t-test and the Wilcoxon test allowed for more accurate prediction over the sign test of which run was better. However, the differences between the t-test and Wilcoxon test were small. They also pointed out that even if one observed a significant difference between two runs based on a small number of topics (≤ 25), one should not be confident that the same observed ordering of runs will be seen in the operational setting. Using a variant methodology, Cormack and Lynam [68] reported significant differences resulting from the t-test failing to be observed in an operational setting particularly for topics with a small number of relevant documents (≤ 5).

Savoy [226] proposed use of the *bootstrap* test,[2] as assumptions of normality, or continuous distributions are not required with this test. Despite Savoy's promotion, the test was little used by the IR

[2] Savoy cites Efron and Tibshirani [81, 82]; Léger et al. [161] as the originators of the test.

community until more recent times, when Cormack and Lynam [68], Sakai [206] and Jensen [134] applied the test. Sakai [208] provided a detailed explanation of how two forms of the test can be used.

A comparison of a number of such newer significance tests was conducted by Smucker et al. [235] who compared the Wilcoxon, sign, and t-tests with the bootstrap and the *randomization* or *permutation* test. Like the bootstrap, almost no properties of a data set must hold before the randomization test can be applied. Smucker et al. compared the values of p obtained across every possible pairing of runs submitted to 6 years of TREC's ad hoc track. They found the t-test, bootstrap and randomization produced similar p values across the pairs, with the Wilcoxon and sign tests producing quite different p values. Smucker et al. stated that the Wilcoxon and sign tests were simpler versions of the randomization test and so argued that in comparison with the randomization test, the different sets of p-values were indicative of errors in the two non-parametric tests, leading the researchers to argue that use of the Wilcoxon and sign should cease. Although their experiments did not show any difference between the remaining three tests, they argued from references to past work that the randomization test was likely to be the best to use in test collection experiments.

5.1.1 Not Using Significance

Spärck Jones [241] suggested a simple approach to determining the significance of comparisons stating that "*in the absence of significance tests, performance differences of less than 5% must be disregarded*". Voorhees and Buckley [276] later clarified that Spärck Jones was referring to was an absolute percentage difference. Given the example of the two TREC runs graphed above, with an absolute difference of 4.9%, one might be tempted to agree with this view. She went on to state that she would "*broadly characterize performance differences, assumed significant, as noticeable if the difference is of the order of 5–10%, and as material if it is more than 10%*". Use of simple tests like this are rarely reported in the research literature.

Voorhees and Buckley [276] when comparing effectiveness measures (see Section 6.4) chose to treat Spärck Jones's threshold as a form of simple significance test: requiring a 5% difference between runs. The

question that Voorhees and Buckley addressed was how many topics were needed in a test collection in order for a 5% difference (measured in the collection) to accurately predict which of a pair of runs was the better in the operational setting. They found that around 50 topics were needed before one could be confident that a 5% absolute difference measured between two runs on a test collection would predict which was the better run in an operational setting. Their result was in stark contrast to Zobel's 1998 work who found that using 25 topics in conjunction with the common significance tests was more than sufficient [295]. Given that the methodologies and data sets used between the two works were similar, it would suggest that, just measuring the magnitude of difference in effectiveness scores has limited utility.

Another work of note is that of Frei and Schäuble [88], who proposed a new evaluation measure that as part of its calculation, computed an error probability to indicate the stability of the value measured. The measure, usefulness, was never widely adopted, although, another feature of the work, that it used relative relevance judgments, proved to be influential on others later.

5.1.2 One or Two-Tail Tests?

So far H_0 has been described, but H_1 has not. There are two types of hypothesis that can be chosen for H_1, which correspond to different types of test: a one- and a two-*tailed* test (also known as a one- or two-*sided* test). In a two-tailed test, H_1 states that the two systems under examination are not equal, e.g., from the runs in Figure 5.1, H_1 would state that system a does not have the same retrieval characteristics as system b. Comparing a and b, a two-tailed t-test of H_0 returns $p = 0.002$; the Wilcoxon signed-rank test returns $p = 0.004$; and the sign test $p = 0.015$. Assuming a 5% threshold, regardless of which test was used, the experimenter would reject the null hypothesis and consider the difference between a and b to be significant.

Since IR experiments are often concerned with determining if a new type of IR system is better than an existing baseline, experimenters sometimes use a form of significance test that focuses only on the question of difference in one direction between two runs: this

is the one-tailed test. Here the experimenter predicts before conducting the experiment that one of the systems will be better than the other and sets H_1 to reflect that prediction. Taking this time the comparison from Figure 5.2, if system b is a baseline and system c is a new system under test, the experimenter sets H_1 to predict that $c > b$. A one-tailed t-test returns $p = 0.036$; the Wilcoxon $p = 0.026$; and the sign test, $p = 0.102$.[3] Despite the lack of significance in the sign test, most experimenters would consider the improvement of c over b as significant. The one-tailed test is recommended for use in IR experiments by Van Rijsbergen [262, Section 7] and more recently by Croft et al. [74], however, it is worth noting that in some areas of experimental science the one-tailed test is viewed as almost always inappropriate [8, p. 171].

The one-tailed version of a significance test has a p value that is half that of the two-tailed version, which makes it a tempting choice for experimenters as its use doubles the chance of finding significance. Note for example that all of the two-tailed tests comparing b and c would have failed to reject H_0. However, if using the one-tailed test, it is important to understand what its use entails. If from Figure 5.1 an experimenter had incorrectly predicted that system $a >$ baseline b and chose to use a one-tailed test; and upon discovering that a $< b$, the experimenter would *have* to conclude that they had failed to reject H_0. In other words, the experimenter would be obliged to report that a and b had the same retrieval characteristics, no matter how much worse a was compared to b; to many a strange conclusion to draw. The experimenter could of course conduct the one-tailed test in the opposite direction, but this second test could *only* be conducted on a new data set.

Recalling the example in Figure 5.2, an abuse of significance tests would arise if an experimenter decided to use a two-tailed test, when comparing c and b, found no significance and so switched to a one-tailed test in the favorable direction in order to search out significance.

[3] It is also worth noting in the c and b comparison how the Wilcoxon and t-tests produced p values below the 0.05 threshold but the sign test did not. The former tests were more influenced by the substantial improvements of c over b in some topics. The sign test ignored the size of a difference; considering only the sign of the difference.

The choice of a one- or two-tailed test needs to be made before analyzing the data of an experiment and not after. If you are not sure of the direction of difference you wish to test for when comparing two systems, a two-tailed test is the appropriate choice. If you are certain that you only wish to test for a difference in one pre-selected direction, the one-tailed test can be used. It is important that the experimenter always states which "tailed version" they used when describing their work.

5.1.3 Consider the Data in More Detail

It is also always worth remembering that although the use of significance tests can help the IR researcher better understand the difference Δ between two runs, the tests are not oracles, they are merely a generic statistical tool constructed for the purpose of estimating the probability p of observing at least Δ if the null hypothesis, H_0, is true. The value of p is calculated on the sample of topics, documents, and relevance judgments in the test collection. If that test set is a representative sample of the broader population of queries, documents and judgments made by users in an operational setting, then the conclusions on whether H_0 can be rejected or not should apply to the population. If, however, as is often the case, the sample is not representative, then conclusions drawn may be unreliable. Sanderson and Zobel [222] showed a number of examples where a range of significance tests produced p values ≤ 0.05 for a pair of runs on one sample test collection, but for the pair using a different sample collection produced p values > 0.05. Voorhees, using larger topic sets [275] went further, occasionally finding examples where one system was better than another, but on a different topic set, the system ordering was swapped and in both cases the differences were significant.

Even if the experimenter compareds two runs using the collection and found $p \leq 0.05$, there remains the question, is the result *practically significant*? Comparing the runs in Figure 5.2, if b was a baseline system already installed at an organization, even though c is significantly better (according to a one-tailed test), the manager of the existing baseline system might argue it is questionable if users would welcome the new system c given that it is notably worse than b on 10 of the 50 topics

(20%). In a different setting, a manager might conclude that c is worth installing because it appears to improve substantially on topics that b performed very poorly on, but only reduces somewhat b's top performing topics and those reductions would be acceptable to his/her users. Such issues, which could be critical in deciding the value of one system over another, are not addressed by significance tests and can only be answered by a more detailed understanding of the uses and users of an IR system.

It is also important to consider the magnitude of difference between two systems: i.e., if a significant difference is substantial enough. In an operational setting, as Cooper pointed out [67], a better system might require more compute resource than the baseline and the benefits of the new might not outweigh the disadvantages of the additional resource needed. Alternatively, experimentation comparing a new system with a baseline might fail to reject H_0. However, if the new system uses fewer resources than the baseline, the new system could still be the better choice. The question of what constitutes a sufficiently large improvement in retrieval effectiveness continues to be examined in some detail by the IR community, details of which can be found in Section 6.5.

A significance test provides a binary decision on whether there is something of note in data or not. On finding significance, researchers may not feel it is necessary to examine their data further. Perhaps of more concern is researchers who fail to find significance, may not look further at their data, which may prevent them from learning what had gone wrong and/or how to fix any problem. Webber et al. explored this situation, pointing out that one possible explanation for failing to reject H_0 is that the experimental setup did not have sufficient *statistical power* for such a difference to be reliably observed: i.e., a Type II error occurred [283]. The researchers described the means of measuring such power in test collections and detailed a method for incrementally adding topics to a test collection until the required power was achieved in order to avoid such errors.

While significance tests are without doubt a popular statistical data analysis method, it is worth remembering that many statisticians feel the tests are overused and that their use discourages researchers from examining their data in more detail. As pointed out by Gigerenzer,

a wealth of other statistical analysis methods exists to allow different forms of analysis to be conducted [94]. One popular alternative is the *confidence interval* (CI) which can be used to compute an interval around a value, commonly displayed in graphs using an error bar. If when comparing two values, the error bars don't overlap, a researcher can state that the difference between the values is of note. In some scientific fields, confidence intervals have replaced significance tests to become the default method for analyzing experimental data. It would appear they were chosen because their use encouraged more analysis of the properties of the data, than significance testing does. Confidence intervals are sometimes used in IR literature; in describing statMAP, Carterette et al. defined how to compute CIs over that measure [48]. Cormack and Lynam [68] described how to calculate an interval on MAP.

6

Examining the Test Collection Methodologies and Measures

Cleverdon's report of his design and use of the Cranfield test collection [58] — with its set of documents, topics, and qrels — concluded a decade of preliminary work in testing IR systems and established a methodology for evaluation that, now approaching its 6^{th} decade, continues to be widely used. The changes in computer technology in that time have been profound, causing IR systems to transform from slow searchers of limited collections to engines capable of searching billions of documents across different media, genres, and languages. With such enormous change, it is striking that the test collection methodology has altered little over that time.

In the decade of research conducted after the development of ad hoc test collections, there was a wide ranging examination of all aspects of test collection methodology, helped greatly by the run data sets produced by TREC and NTCIR. Exploitation of these sets prompted a re-examination of the impact of assessor consistency on measurements made on test collections; an exploration of pooling including consideration of effective use of relevance assessors' time; determining which were the best topics to use in a test collection; establishing which is the best evaluation measure; and perhaps most importantly, determining

if test collections actually predict how users will use an IR system. The work is now described.

6.1 Re-checking Assessor Consistency

As highlighted in Section 2.5, an early concern of some researchers was the potential of inconsistent relevance assessments resulting in poor quality qrel sets, which could mean that measurements made on test collections could be inaccurate. Early on, both Lesk and Salton as well as Cleverdon tested the potential for such error and found that variations in assessments although high did not affect the relative ranking of runs. Voorhees conducted a larger scale study using TRECs 4 and 6 run data [268], for which multiple relevance assessments, qrels, were available.

Voorhees established a method of correlating ranks of runs that became widely used by many other IR researchers. The method involved ranking the runs submitted to the two TRECs using different qrel sets. Voorhees measured the correlation between a rank of runs using one qrel set and the same runs ranked using a different qrel set. A high degree of correlation meant the qrel sets ranked IR systems similarly; and low correlation implied that each qrel set was ranking runs differently.

In order to determine the similarity between ranks, Voorhees used Kendall's Tau (τ) [153]. For any two rankings of the same n set of items, τ is a linear function of the number of pairs of items which are in different orders in the two rankings. This function is constrained such that $\tau = 1$ if the two rankings are in identical order, and $\tau = -1$ if the order is reversed. There are a number of τ variations, we illustrate the one originally defined by Kendall. It is as follows:

$$\tau = \frac{2(n_C - n_D)}{n(n-1)}.$$

Here n_C is the number of pairs in the two rankings that are in the same order (i.e., concordant) and n_D is the number of pairs that are in different order (i.e., discordant). Note that every possible pair of items in the rankings (i.e., there are $n(n-1)$ such pairings) is compared when calculating n_C and n_D. If one was using τ to measure the correlation

between two rankings of ten runs from TREC and the two rankings were identical, then $n_C = 45$, $n_D = 0$, and $\tau = 1.0$. If there was a swap anywhere in one ranking of two adjacent items, then $n_C = 44$, $n_D = 1$, and $\tau = 0.96$. If the swap was between the first item and the last item of one ranking, then $n_C = 28$, $n_D = 17$, and $\tau = 0.24$.

When considering τ calculated over a set of rankings, a common question that is asked is at what level of correlation can one view two rankings as effectively equivalent? Voorhees suggested $\tau \geq 0.9$ as such a threshold [270], which was adopted by a number of subsequent researchers.[1] Using Kendall's τ Voorhees found the rankings of runs, though not identical, were very similar, leading Voorhees to conclude that variations in assessment did not impact noticeably on retrieval effectiveness. The second part of Lesk and Salton's work on examining the consistency of judgments against the rank of the documents being judged was also repeated in later years: both Sanderson [216] and later Voorhees [270] using different TREC data sets, showed that inter-assessor agreement was higher for top-ranked documents.

In the experiments conducted up to this point, the assessors used to generate different qrels were all assumed to be capable of judging the relevance of the documents. In later work based on TREC Enterprise track data, Bailey et al. [19] drew from different sets of assessors based on their knowledge about the test collection topics. The assessors were classed as gold, silver, and bronze judges. Gold and silver were subject experts with the gold judges having a more intimate knowledge of the data set being searched. The bronze judgments were made by the participants in the TREC track: presumably motivated, but non-subject experts. Like the previous results, Bailey et al. showed that the

[1] Sanderson and Soboroff [220] pointed out that the items in a ranking are sorted by a score and the range of the scores of the items in a list, impacts on the value of τ. They showed that if the range is large, there is a greater likelihood of finding high τ correlation scores. This quality is common to all rank correlation measures and makes use of absolute thresholds difficult. Another criticism of τ is that it measures correlation equally across a ranking and for many IR tasks, correlation in the top part of a ranking (i.e., the runs of the best performing systems) is generally more important than the bottom. Yilmaz et al. [290] produced a new correlation coefficient τ_{ap} that addresses this failing. They also cite a number of other proposed τ variants. Carterette has also described an alternative to τ [46]. See also Melucci [175] on concerns about use of τ in IR experiments.

qrels from the gold and silver judges produced similar rankings of runs. However, the rankings from gold and bronze judges were different.

These works show that while test collections are more resilient to assessor variation than was originally feared, there are limits to this resilience and the appropriateness of the assessors used to make judgments needs to be carefully considered when forming qrels.

It is often thought that differences in assessment are an indication of some sort of human error. However, Chen and Karger [53] showed that one can view the differences as simply two distinct, but valid, interpretations of what constitutes a relevant document. They used the TREC-4 and 6 multiple assessments to test a retrieval system that returned diverse search results. Viewing the two sets of assessments as representing two legitimate interpretations of relevance for the topics in TRECs 4 and 6, Chen and Karger showed that supporting diversity ensured that more documents were retrieved that satisfied at least one of the assessors. For more on diversity evaluation, see Section 4.1.2.

6.2 Does Pooling Build an Unbiased Sample?

The aim of pooling is to locate an unbiased sample of the relevant documents in a large test collection, as made clear by Spärck Jones [243]. She was confident in the validity of pooling in part due to earlier work by Harman [102] and later Zobel [295] who tried to estimate the quantity and impact of relevant documents missing from TREC pools. In the early years of the TREC ad hoc collections, pools were formed from the union of the top 100 documents retrieved by each submitted run for each topic. Harman examined a pool formed by the documents in ranks 101–200 for a sample of the runs and topics in TREC-2 and all the runs and topics in TREC-3. She reported that a further 11% of relevant documents were discovered in the TREC-2 pool and a further 21% in TREC-3. Harman stated that "*These levels of completeness* [in the pools] *are quite acceptable for this type of evaluation*". Zobel examined the relationship of the number of relevant documents found (n) to the depth of rank used to form a pool (p) and found the relationship to match well to the following power law distribution:

$$n = Cp^s - 1,$$

where C and s were constants. So strong was the fit that Zobel felt confident to extrapolate the curve beyond the depth 100 pools of TREC. Extrapolating what n would be when $p = 500$, he stated that the number of extra relevant documents would be double that found when $p = 100$. The prospect of so many unfound relevant documents caused Zobel to consider if it was possible for runs to retrieve the majority of the unknown relevant and consequently receive an unfairly low effectiveness score. He, therefore, explored how the contributing runs to TREC would have been ranked if they had not contributed to the formation of the pool. This was achieved by, in-turn, removing from the pool the relevant documents unique to a particular run, re-forming the reduced qrels and then comparing a ranking of all runs with the original run rank. Zobel found relatively small changes in the way that left out runs were ranked, the pools appeared to be an unbiased sample. Zobel stated the result boded well for the reusability of test collections.

More recently, Büttcher et al. [43] explored two adaptations of Zobel's "*leave one run out*" experiment. They pointed out that it was common for a research group to submit multiple runs to TREC, leaving a single run out, as Zobel had done, was perhaps not the best of simulations as other runs from the same group would have been left in. Therefore, borrowing a technique described by Voorhees and Harman [277] the researchers used a "*leave one group out*" approach. Testing their work on the TREC 2006 Terabyte track runs and qrels, the researchers found that removing a whole group made a relatively small difference to the way runs from the held out group were ranked. In a second test, the researchers held out relevant documents uniquely found in manual runs from the pools. The manual runs are generally the richest source of relevant documents. By leaving them out Büttcher et al. were attempting to simulate a situation where after a test collection was formed, a new substantially better run was tested on the collection. Examining how well these runs were ranked by the reduced pools, Büttcher et al. reported that the runs were ranked somewhat differently compared to when the full TREC pools were used.

Such a result was undoubtedly concerning. However, the extent that it represented a problem for experiments is yet to be fully determined. There would not appear to be much in the way of evidence that users of

test collections have found new retrieval systems poorly scored, though admittedly few researchers actually analyze for such potential problems. There appears in the literature to be just one example of a run that if it had not contributed to the pool of a test collection, it would have been poorly scored on that collection. So unusual was this particular run that it was studied in some depth by Buckley et al. [36]. They found that within the pool of runs used to form the test collection was a particular bias: most relevant documents contained at least one word from the title of the test collection topics. However, the errant run contained many documents that were both relevant and did not contain words in the topic title.

Buckley et al. studied this unusual run and the properties of the collection to try to better understand the causes of this anomaly. It would appear that the size of the collection was an important factor. It is normal for many of the relevant documents in a test collection to contain words from the title of a topic and it is also to be expected that such documents would be highly ranked. The number of such documents is finite. In the smaller test collections it would appear that the size of the pools assessed was larger than the number of relevant documents containing terms from the topic title. However, in the collection under study, the pool size was not large enough to encompass all the relevant documents containing a topic title word and to also find enough of the relevant documents without.

Whether this one example was an outlier or an indicator of a broader problem with current approaches to pooling in large test collections is yet to be determined.

Examining a different aspect of potential bias in pooling, Azzopardi and Vinay [17] studied if within large collections there are documents that are almost never retrieved by any search engine. Loading large collections into a conventional ranked retrieval system, they ran hundreds of thousands of queries on the collection. The queries were either single word terms that occurred >5 times in the collection, and bigrams that occurred >20 times. The researchers' aim was to understand if through all those queries there were documents in the collections that persistently failed to be highly ranked. Their conclusions were that a notable number of such documents existed in these collections. Such a

conclusion could be of concern since pools are built through querying. However, Azzopardi et al. did not examine the relevance of the poorly retrieved documents. It is unclear at the moment if this probable bias is important with respect to locating a representative sample of relevant documents.

It would appear that despite concerns of some in the IR community that pooling risks the creation of test collections a biased sample of qrels, studies have largely shown such concerns are unfounded. However, the effort required to build pools is substantial and as described in Section 6.3, attempts are being made to produce smaller pools, which might introduce new forms of bias. This is a topic, therefore, that is likely to be returned to in the future.

6.3 Building Pools Efficiently

Reflecting on the first eight years of TREC, Voorhees and Harman [278] detailed that the average number of documents assessed per topic in the ad hoc tracks of TREC was 1,464 (averaged from data in Table 4 of that paper). With 50 topics per year, approximately 73,000 judgments were made each year. At a rate of two judgments per minute — Voorhees and Harman [279, Section 2.1.3] estimate — 8 hours of work per day, judging the TREC ad hoc pool each year took just over 75 person days. This was the limit of human resource TREC organizers were able to supply; similar limits applied to other large evaluation exercises such as CLEF or NTCIR.

Assessors in evaluation exercises tend to be used in a relatively straightforward and similar manner. Here we highlight the way that assessors were used in TREC ad hoc. When groups submitted a run to TREC, the top 100 documents for each topic in the run were extracted and merged into a pool for each topic and sorted by document id. The assessor for a topic was generally the person who created that topic. They examined every document from every system in the pool for that topic. This straightforward approach to examining a pool was justified by Voorhees and Harman [278] and later by Soboroff and Robertson [239] by stating that it was important for the pools to contain a set of documents that were not biased in any way toward one particular

retrieval strategy or a particular type of document. This was judged vital so that the qrels could be used not only to fairly determine the relative effectiveness of runs submitted to TREC, but also that later users of the test collections could be confident that the effectiveness of a new retrieval strategy will be accurately and fairly measured.

A number of researchers examined other ways of selecting or sampling parts of a pool to judge so as to use fewer human assessments. The approaches are grouped here into examinations of the way that pools are scanned; research assessing if pool depth or topic breadth is more important; approaches to solve the assessment resource problem by distributing assessment; an examination of an approach that opened the possibility of avoiding use of assessors at all; and finally exploitation of existing data to simulate assessments.

6.3.1 Scanning the Pools in Different Ways

Zobel [295] pointed out that some topics in a test collection will have more relevant documents in them than others and suggested that as topics were being assessed for pools, the number of relevant found so far could be noted and for those topics richer in relevant documents, more assessor effort could be focused on examining a larger pool for them.

In the same year that Zobel's made his suggestion, Cormack et al. [69] proposed a number of alternate strategies. In a similar vein to Zobel's ideas of focusing assessor effort on the richest sources of relevant documents, they pointed out that the runs from certain retrieval systems contributing to the pool contained many more relevant documents than others. They proposed focusing assessor effort onto those runs richer in relevant documents. This approach they called local move to front (MTF) pooling. The researchers also tested an approach that included Zobel's ideas: prioritizing assessment of both the most fruitful runs and the most fruitful topics, which they called global MTF pooling. The authors tested the approaches and showed that they could assess 10% of the full TREC pool and produce a qrel set that ranked runs in an almost identical manner to the ranking achieved using the TREC baseline pool. Global MTF appeared to be more effective than local MTF. See also Moffat et al. [178] for related work.

Cormack et al.'s paper contained one other approach to building qrels, which they called *Interactive Search and Judge* (ISJ). Here they proposed an alternative role for the relevance assessor, instead of judging a long list of documents, the assessor would search the test collection, issuing multiple queries, noting down relevant documents found and searching until they could find no more relevant for a particular topic. Cormack et al. reported that a group of ISJ assessors was given the task of locating relevant documents for one of the years of TREC. In just over 13 person days (compared to TREC's 75 person days) a qrel set was formed that was shown (through experimentation) to be of comparable quality to that produced by TREC.

The work was a re-examination of the approach to pooling used by Katzer et al. [146] and earlier still by Lancaster [158]. TREC implicitly uses ISJ through its encouragement for manual runs to be submitted to its tracks [217]. It has also been used explicitly by a number of evaluation campaigns such as CLEF [63, 93] and NTCIR [156]. Recognizing that TREC assessors preferred to assess rather than search, Soboroff and Robertson [239] described an alternative approach to ISJ where relevant documents located by assessors were used as positive examples for a relevance feedback system to locate more items for assessment. Soboroff reported that the approach had worked well in reducing assessor time and effort. Oard et al. [184], detailed using this technique, which called they *search-guided assessment*, to build a test collection.

Carterette et al. [47] pointed out that certain documents in the pool were better at distinguishing runs from each other and these should be targeted for assessment. For example, if one compared two runs using P(10) and all that one wished to determine was which run was better, only the top ten documents needed to be examined and any documents in common could be left unjudged. Through such targeting, the researchers took eight runs from six different searching systems retrieving across 60 topics and in 6 hours of assessor time were able to produce a ranking of those runs that with 90% confidence was the same as the ranking produced by a baseline TREC top 100 pooling approach.

Another approach to reducing assessment is sampling of pools, an analysis of which was first described by Spärck Jones and Bates [244,

pp. 20–21]. Aslam et al. [13] described experiments with pool sampling showing that one could sample as little as 4% of a TREC ad hoc pool and still produce accurate results. Lewis [165] used stratified sampling to the pools of the TREC filtering track. More recently this approach was exploited in the TREC million query track [51] and the legal track [22]. When sampling pools, the number of unjudged documents is likely to increase, consequently, the measures described in Section 4.2.2, tend to be used, particularly, *infAP* and *statAP*.

When compared to the original approach used to form qrels for the TREC ad hoc collections, it would appear that at least some of the methods described here could be used with confidence. For some of the newer approaches involving more extreme forms of sampling, although experimental results have shown the methods to be reliable when tested on historical data, there is still the question of how re-useable such collections will be in experiments run in the future.

6.3.2 Narrow and Deep, or Wide and Shallow?

One question considered by test collection creators is should assessors be focused on judging deeply the pools of a small number of topics, or the shallow pools of a larger number of topics? The convention was to limit the number of topics and examine their pools in some detail. For many years in TREC, the top 100 retrieved documents of submitted runs were assessed, though in recent years this was reduced to the top 50. Sanderson and Zobel [222] speculated on the number of topics that could be assessed if a smaller part of a run was drawn into a pool. They estimated that if only the top 10 of each run was assessed, a test collection with between 357 and 454 topics could be created using the same amount of assessor effort as with 50 topics examined to depth 100. They also pointed out that the top part of runs generally have a greater density of relevant documents, consequently such a strategy would in all likelihood find between 1.7 and 3.6 times more relevant documents than a conventional pooling approach.

Bodoff and Li [26] pointed out that sources of score variation in test collection based evaluation can be attributed to different IR systems, different topics, and to interactions between systems and topics. The

researchers analyzed TREC run data using Generalizability Theory to identify where the main source of variation was, concluding that topics were the highest source. This led the researchers to conclude that building test collections with more topics was a priority. Webber et al. [283] applied their analysis of statistical power in test collection based experiments to also study this question. They found that greater statistical power would result from a wide and shallow approach to pooling.

At around this time, publications from the commercial search engine community were produced showing that internal test collections had substantially larger numbers of topics than existed in publically available ones: White and Morris [284] mentioned a collection at Microsoft with 10,680 "*query statements*"; Carterette and Jones [49] described a collection in Yahoo! with 2,021 queries; at the same company Chapelle et al. [52] mentioned several internal collections, one with over 16,000 queries, judged to depth five.

In reaction to this, the academic community looked to build its own collections with many more topics. Carterette et al. [51] described work on the so-called Million Query Track, which, using both Carterette et al.'s just-in-time approach and the pool sampling methods associated with *statAP*, created a test collection with 1,755 topics. With their larger data set Carterette et al. were able to empirically test the proposal by Sanderson and Zobel that "wide and shallow" was better that "deep and narrow". In their data set Carterette et al. found that 250 topics with 20 judgments per topic were the most cost-effective in terms of minimizing assessor effort and maximizing accuracy in ranking runs. Voorhees expressed concern that wide and shallow test collections might not be as reusable as ones built using the deep and narrow approach [274]. However, Carterette [45] provided evidence that such collections were more reusable than was perhaps previously thought.

6.3.3 Distributing Assessment

The relevance assessors are generally paid by the organizers of evaluation campaigns. There have been attempts to reduce or remove such

costs. Both INEX and the enterprise track of TREC explored making relevance assessment a necessary part of groups being able to participate in the campaign. There is relatively little research on the accuracy of such "coerced" judgments. However, Bailey et al.'s work [19] on gold, silver, and bronze judgments suggested that such an approach to gathering judgments was not without its risks.

Another potentially large source of human assessors can be found through crowd sourcing. Alonso et al. [7] described their use of the Amazon Mechanical Turk system to obtain relevance judgments. Mechanical Turk is a market place where workers are paid small amounts of money (typically ranging from 1 cent to $2) to conduct short run tasks, called *HITS*. The tasks on offer in the market place include writing short reviews, adding metadata to images, and judging documents for relevance. Because the workers on systems like Mechanical Turk are anonymous, it is hard to know the motivation of those conducting the task. It is reasonable to assume that some workers will attempt to earn money for little or no work. Alonso et al. described their attempts to ensure that the anonymous workers chosen were motivated and appropriate for the task.

Work in this area is still relatively novel and the success of such approaches requires more study.

6.3.4 Absolute vs. Relative/Preference Judgments

Virtually every test collection built has gathered its qrels using absolute judgments: asking an assessor to determine a document's relevance relative to a topic independent of other documents seen. Researchers have sometimes asked if these so-called *absolute* judgments are the most reliable approach to gathering qrels, suggesting instead that *relative* or *preference* judgments made between pairs of documents are sought instead. Some research addressed this question. In the context of concern about consistency of assessor judgments, Rees and Schultz [193] stated "*It is evident that the Judgmental Groups agree almost perfectly with respect to the relative ordering of the Documents.*" (p. 185). In contrast, Katter reported on an experiment the results of which showed that absolute judgments were more reliable [145].

A concern with relative judgments was that gathering a complete set required showing assessors every possible pairing of documents under consideration, an $O(n^2)$ problem. Rorvig conducted experiments examining the tractability of building test collections with relative judgments [203]. His results indicated that collections could be built from such judgments as preference judgments appeared to be transitive, which meant that some preferences could be reliably inferred, substantially cutting the number of judgments needing to be assessed. He proposed a methodology for building a test collection. More recently, Carterette et al. [48] showed that relative judgments drawn from users produced more reliable results than absolute. They also found that 99% of judgments gathered were transitive and went on to build on Rorvig's methods for reducing the number of preference judgments that needed to be made.

There does not as yet appear to have been a public test collection built from relative judgments. As will be seen in Section 7.2, however, deriving relevance judgments using preference from query logs is increasingly common.

6.3.5 Don't Use Assessors at All?

Soboroff et al. [238] examined the possibility of not using human input of any kind when creating relevance assessments. They hypothesized that judgments could be generated simply by randomly selecting commonly retrieved documents from the pool of runs used to form a test collection. The researchers examined the way in which runs submitted to several years of TREC ad hoc were ranked using such automatically generated qrels and compared the ranking using the standard qrels of TREC. They found that the two rankings were correlated relatively well. However, the most effective runs were ranked as poor or middle performing runs by the automated qrels. Consequently, the approach was judged to not work.

Aslam et al. [12] later pointed out that the method was ranking runs by their similarity to each other; those runs retrieving the most popular documents amongst the set of runs were ranked highest. The best runs retrieved relevant documents that few other runs found, such

documents were not in the automatic qrels, consequently the best runs were scored poorly. Soboroff et al.'s method was explored further, Wu and Crestani [287]; Shang and Li [227]; Can et al. [44] and later Nuray and Can [183]. As yet, no success has been found in fixing the important failing in Soboroff et al.'s approach.

Others suggested automatically creating both qrels and topics. A recent exploration of this area was described by Azzopardi et al. [16] who examined means of creating known item topics for a range of collections and languages; they were inspired by earlier work from Tague and Nelson [252]. The work demonstrated the potential for this approach, but the authors acknowledged that more investigation was needed.

6.3.6 Exploiting Structure or Other Data Sets

Although pooling was the predominant means of forming qrels, alternative approaches to seeking relevant documents were tried. What follows is a list of some proposals.

- Sheridan et al. [228] described building a small spoken document test collection of broadcast news items. To make relevance assessments easier, queries are referred to events that occurred on a specific date. This allowed the researchers to concentrate assessment to items broadcast on or after the date of the event.
- Using pre-existing manual organization of documents has been used on a number of occasions. Harmandas et al. [108] described the building of a web image test collection where assessors were encouraged to use the topical classifications present on many websites to locate relevant items. Haveliwala et al. [109] used a similar approach using the topical grouping of the Open Directory website to locate relevant related documents.
- Cleverdon building his Cranfield II collection composed of scientific papers used references in papers to identify potentially relevant material. This approach was further developed by Ritchie et al. [194].

- When working in the enterprise web search domain, Hawking et al. [113] described using the sitemaps of a website (a page that maps out for users the location of important pages on a website) as a source of relevance judgments for known item searching. The known item being the identified page, the query for that item being a title extracted from the sitemap page.
- Amitay et al. [9] proposed *trels*. For each topic in a test collection it was proposed to manually form a set of words that defined what was and was not a relevant document. Once a stable set of trels was formed, unjudged documents were assessed against the trels to determine relevance. Amitay and her collaborators showed trels to be successful in a simulation on TREC data. This approach was also used to build reusable question answering test collections, see Lin and Katz [166].
- Jensen et al. [135], in the context of web search, tested the combining of manual relevance judgments with judgments mined from a website taxonomy, such as DMOZ. They were able to show that the additional judgments improved evaluation accuracy.
- In the field of personalized search, use of bookmark or URL tagging data has been used as an approximation to relevance judgments in a personalized searching system. See for example, Xu et al. [288].
- An ever increasing body of work has examined the use of search engine logs to help determine relevance of items. Morita and Shinoda [182] explored using the time that a retrieved item was viewed as a way to infer the relevance of the item. More common was the use of click data in logs to determine relevance, e.g., Fox et al. [87]. So much log data is being generated particularly within large web search engines, that there is extensive research in analyzing log data and exploiting it. A description of the work in this area is described in Section 7.

With the exception of using query logs, none of the methods described has been as thoroughly tested as pooling.

6.4 Which is the Best Effectiveness Measure?

Perhaps surprisingly, for a research field that so values evaluation, it would appear that for many decades there was no quantitative research into the relative merits of different effectiveness measures. This was rectified in recent years through two forms of study: calculating the correlation between evaluation measures and assessing the stability of measures.

6.4.1 Correlating Measures

Tague-Sutcliffe and Blustein [254] were the first to quantitatively compare evaluation measures, establishing a methodology that became the standard for most subsequent research. Taking archived runs TREC, Tague-Sutcliffe and Blustein used different precision-based evaluation measures to each rank the runs. Correlations measured between the ranks showed strong similarities across the measures. The researchers concluded that there was little value in calculating different precision-based measures. However, more recent investigations, e.g., Buckley and Voorhees [39] and Thom and Scholer [255], showed that high precision measures, such as P(10) and P(1), correlated less well with measures such as MAP or R-precision.

6.4.2 Measuring Measure Stability

Zobel [295] devised a method to test the predictive power of evaluation measures. The core role of a test collection is to determine which retrieved method will produce the best effectiveness when used in an operational setting. Zobel simulated this testing and operational setup by splitting the topics of a test collection in half: one-half was treated as a mini test collection, the other half was a simulation of the operational setting. Using TREC-5 run data, Zobel took pairs of runs and determined which was the best on the mini collection and then measured

if the winning run was still the best in the operational setting. If it was, then a correct prediction was made using the reduced collection, if the pairs had swapped order, the result measured on the collection was a mistake. Using this *swap method*, Zobel determined which of four precision based the measures produced better predictions. He reported that although P(10) and $P(100)$ were worse at predicting than 11 point interpolated AP, in his judgment, the difference between the measures was too small to be of concern.

Using an alternative method, Buckley and Voorhees exploited the TREC-8 query track test collection [37], which had 21 so-called *query sets*: manually generated variations of each of the 50 topics in the collection. Each of the sets was run against a range of different retrieval systems resulting in 9 runs for each of the 21 sets. Buckley and Voorhees sought an evaluation measure that ranked the runs consistently over the query sets that also produced the smallest number of ties. They reported that measures such as MAP, R-precision, and P(1000) were the most stable; P(10) and P(1) the least. Buckley and Voorhees judged MAP to have the best balance between high stability and few ties.

In a separate study Voorhees and Buckley [276] applied Zobel's swap method to a wide range of TREC test collections and again confirmed that rank cutoff measures like P(10) were less accurate at predicting the effectiveness of runs than measures like MAP. One possible reason for the difference between Zobel's ambivalent and Voorhees and Buckley's more emphatic conclusions about a measure like P(10) was that Zobel used his measures in conjunction with a significance test, Voorhees and Buckley did not.

Sanderson and Zobel [222] pointed out that when comparing two measures, such as, MAP and P(10), the effort required to judge the relevant documents for MAP was substantially higher than that required to assess P(10); where only the top 10 documents from each run need be examined. Analyzing nine years of TREC data, Sanderson and Zobel showed that P(10) required between 11% and 14% of the assessor effort required to calculate MAP. The researchers concluded that P(10) was far more stable than MAP per equal quantity of assessor effort. If the qrels of a test collection already exist, then Sanderson and Zobel's point on the value of P(10) over MAP was not important. If one was

evaluating retrieval systems without a test collection, where assessors still had to judge the relevance of documents, then consideration of assessor effort was critical.

In contrast to most papers suggesting that MAP produces stable ranks of runs, Soboroff [236] used the swap method on a test collection with a small number of relevant documents per topic: the TREC 2003 topic distillation collection. He found that P(10) was noticeably more stable than R-precision and MAP. Soboroff also showed that MRR can be stable when used in a collection with a large number of topics (≥ 100). Further means of testing stability were described by Bodoff and Li [26] using Cronbach's alpha (a statistic that measures co-variance); and Sakai [208] who used the bootstrap test to count statistical significance between pairs of runs when measured with a particular evaluation measure.

It is worth remembering that the work on measure stability, while valuable, has its limitations. An "evaluation measure" could be created that ranks the runs of different retrieval systems by an alphabetical sorting of the run's name: e.g., a run labeled "Inquery" would be ranked higher than a run labeled "Okapi", which would be ranked higher than "Terrier". Under every stability test described here, this useless measure is perfectly stable; Sakai's significance count methodology would result in the maximum number of observable significant differences, and the Cronbach's alpha approach would show perfect co-variance.

Ignoring questions of stability, Aslam et al. [15] used a maximum entropy-based method to explore the degree to which an evaluation measure predicted the distribution of relevant and non-relevant documents across a retrieved list. Essentially, in this work the aim was to understand how well the single value from an evaluation measure summarized the distribution of relevant and non-relevant documents. Aslam et al. found that average precision was the better measure compared to R-precision and precision measured at fixed rank.

In this section, the assessment of measures was achieved by comparing a relatively simple property of each measure against some ideal. In the following section, the outputs of test collections and evaluation measures were compared with models of user searching behavior.

6.5 Do Test Collections or Measures Predict User Behavior?

A series of experiments were conducted to measure how well predictions made using test collections or evaluation measures correlated with a range of user behaviors when searching on systems under test. Results from this work are contradictory; the research described here is broken into those that concluded that little or no correlation existed, those that showed some link and those that showed a stronger link. Finally the apparent contradictions between these sets of work are discussed.

6.5.1 Little Correlation Found

In testing the impact of a searching system on user behavior, one can choose to measure effectiveness scores of users searching on an operational system and look for correlations between the scores and some aspect of user behavior or an outcome from the search. A number of studies took this approach. Tagliacozzo [251] showed that 18% of ∼900 surveyed MEDLINE (a medical literature search engine) users did not appear to be satisfied with search results despite them containing a large number of relevant retrieved documents. Su [247] attempted to correlate many measures of IR performance with user satisfaction. She found that precision did not correlate significantly with satisfaction and examined this issue in more detail later [248]. Hersh et al. [120] examined medical students' ability to answer clinical questions after searching on MEDLINE. Expert assessors were used to calculate recall and precision of the students' search outputs looking for correlations between these measures and the scores students attained for the questions. The researchers reported no correlation. Similar work was conducted by Huuskonen and Vakkari [126] producing similar negative results.

Hersh et al. [122] were the first to try to correlate test collection-based results with user behavior. They examined a pair of IR systems measured as significantly different on a small test collection; when subjects used one of the pair of systems, no significant difference in user behavior was observed. This experiment was repeated on another small collection with the same perhaps surprising conclusion [258]. See also

a more detailed examination of the experiments [259]. Using a method of artificially creating ranked document lists each with a different level of MAP, Turpin and Scholer [260] described a larger experiment that showed some small significant differences in user behavior when there were large differences in MAP between the artificial ranks.

Smith and Kautor [234] engaged 36 users to each search 12 information gathering topics on two versions of a web search engine: one the normal searching system, the other a version of the engine which displayed results starting from rank 300, presumably much worse. No significant difference in user success in finding relevant items was observed. Smith et al. reported that users adapted to the poorer system by issuing more queries; this change appeared to mitigate the smaller number of relevant documents retrieved in each search.

To many researchers, the totality of this work highlighted the artificiality of test collections. Ingwersen and Järvelin [128, p. 234] provided a detailed survey of past work that outlined the limitations of what an experimental result on a test collection can tell the researcher. The collective results from these works were viewed by some as strong evidence that there was a problem with the test collection methodology.

6.5.2 Some Correlation Found

Allan et al. [6] studied the problem of locating relevant text fragments, called *facets*. The researchers created artificial document rankings displaying fragments and links to full document texts. The rankings were formed starting by randomly degrading a perfect ranking. Users were asked to identify within the rankings, sections of documents that were relevant to a topic. Subjects were given many hours to complete the task. Allan et al. measured the time users took to complete their task, their error rate, and their facet recall. Unlike previous work, the researchers found a correlation between user behavior and test collection-based evaluation measures.

Huffman and Hochster [124] addressed the question of how effectively a test collection can be used to predict user satisfaction. They described getting two sets of assessors to judge the search results of 200 queries: the first assessors judged the relevance of the top three results;

and the second set of assessors judged user satisfaction with the overall results. The researchers reported finding a correlation between DCG measured on the relevance judgments and user satisfaction.

Al-Maskari et al. [3] conducted a small study measuring correlations between user satisfaction measures and different evaluation measures based on examinations of Google searches. She showed that there was a strong correlation between user rankings of results and the ranking produced by the evaluation measures she tested. She found that Cumulative Gain (CG) correlated better with user measures than P(10), DCG, and nDCG. Later, she and others used a test collection to select a pair of retrieval systems that had noticeably different effectiveness scores on a particular topic [4]. They then measured how well groups of users performed on those two systems for that topic. Fifty-six users searched from a selection of 56 TREC topics. The researchers showed a correlation between test collection experiments and user behavior, though they noted that user satisfaction was harder to predict than more objective measures such as the number of relevant documents saved.

6.5.3 Strong Correlation Found

When conducting an analysis of click log data, Joachims claimed that "*It appears that users click on the (relatively) most promising links ... independent of their absolute relevance*" [136]. He described experimental results showing that users, given different versions of an IR system, clicked at almost the same average rank position, despite there being differences in the effectiveness of the three versions. Joachims highlighted Rees and Schultz's [193] past work on relative relevance judgments and proposed an alternative approach for measuring user interaction with different systems. His suggestion was to interleave the outputs of the different systems into a single ranking and observe if users tended to click more on results from one ranking over another. The results of this *preference*-based experiment showed that users chose the results from the better ranking in a statistically significantly measurable way. This work was repeated later by Radlinski et al. [192], showing the same results.

Inspired by Joachims, Thomas and Hawking [256] presented a different preference methodology that allowed users to express a preference for not only the ranking of a retrieval system but potentially its interface as well. In their methodology, two versions of a search engine result were presented side-by-side to users. Users could query the two engines and interact with them as normal. Thomas et al. presented in the two panels, the top 10 results of Google and the presumably worse Google results in ranks 21–30. The researchers observed a clear statistically significant preference for the results from the top ranks over the lower-ranked results.

6.5.4 Discussion

The work showing little correlation might lead some to question the value of test collections; however, it is notable that many of the studies in the opening sub-section failed to show statistical significance in the user-based tests. A lack of significance can mean that there is no measurable difference or it can mean that the experiment was not powerful enough to allow such differences to be measured (see Section 5). The challenge of accurately measuring users was pointed out by Voorhees [274] who suggested that the experiments, such as those from Hersh and Turpin et al., concluding failure in test collections may in fact have failed to measure their users' behavior accurately enough. Perhaps the strongest conclusions to draw from these collective works is that faced with a poor search, or worse a poor IR system, users either make do with the documents they are shown or they work around the system to manage to achieve their search task successfully. The last tranche of studies contrasts with the former as user's performance was assessed in a relative instead of an absolute way. From that work, it would appear that given a choice between two systems, users prefer to use the better system as a source of retrieval results.

6.6 Conclusions

This section examined in some detail the range of research that tested many aspects of the test collection method. Assessor consistency was

re-examined and was generally found to be un-problematic. Pooling was found to produce a sample of relevant documents to effectively rank runs. Means of building collections more efficiently were proposed and a number of those methods adopted. Evaluation measures were examined in detail and the importance of selecting the right measure for the right task was highlighted. Finally, the consistency with which test collection results predicted user behavior on operational system was examined. Perhaps the simplest conclusion to draw here is that measuring users accurately requires care.

7

Alternate Needs and Data Sources for Evaluation

As shown in Section 6, over the past decade, a detailed examination of the construction and use of test collections was conducted that by and large found the long-standing evaluation methodology to be a valid approach to measuring the effectiveness of IR systems. However, during that period, the needs of at least part of the IR research community changed and at the same time, new potential sources of information about the relevance of documents became more accessible to researchers. In this section, the new need is described and the data sets created for it are outlined. Also two new evaluation data sources are introduced. As much of this work is beyond the scope of a test collection review article, it is described here briefly.

7.1 Learning to Rank

Test collections and evaluation measures are commonly used for the purposes of comparison: deciding if one approach or one retrieval system is better than another. However, there is a related use, retrieval function optimization. The ranking functions of IR systems are increasingly complex, containing a wide range of parameters for

which optimal settings need to be found. A common approach to finding such values is to use a machine learning approach known as *Learning To Rank* (LTR). The study of LTR has its origins in the late 1980s [89]; see [168] for other early LTR papers. Although a resurgence of interest started around 2005, from the point of view of evaluation, work is still in its infancy. There are two key evaluation areas to consider: data sets and evaluation measures.

7.1.1 Data Sets

As with any machine learning approach, data is needed to train an LTR retrieval function, which is then tested on a separate data set. In LTR, the data are generally composed of the classic components of a test collection: documents, topics, and qrels. It is notable that in his pioneering work, Fuhr stated that a key concern was the approach "*needs large samples of relevance feedback data for its application*", by which he meant training and testing data. Fuhr used a test collection with > 240 usable topics [91]. The first shared LTR data set was the LETOR benchmark [168]. It was composed of two existing IR test collections: OHSUMED and TREC web, which together had a similar number of topics to Fuhr's earlier collection. A series of features were extracted from all relevant and top-ranked documents in relation to each topic in the data set. Machine learning groups who were not interested in extracting such features from the documents could simply apply the features to their learning algorithms.

The collection quickly became a standard for use in LTR experiments. However, it is relatively new and recent publications have suggested that certain biases exist within it [176]. It is likely that adjustments to LETOR to correct these biases will arise as will the creation of new LTR collections. However, it is not clear if adapting existing IR test collections will produce large enough data sets for the LTR community. Web search companies, such as Yahoo!, have released custom built LTR data sets[1]; exploiting sources such as query logs to build data sets are an active area of research, which are described in Section 7.2.

[1] http://learningtorankchallenge.yahoo.com/ (accessed April 26, 2010).

7.1.2 Evaluation Measures

An LTR function is trained with respect to a particular evaluation measure. Liu et al. [168] described training using a series of common measures: P(n), nDCG, and MAP. It was assumed that the measure to use when optimizing was the one that reflected, most accurately, a model of the user in the operational setting one is optimizing for. Recent research, however, from Yilmaz and Robertson [291] showed that measures that make the greatest use of available training data can in fact be the better measure to employ. For example, although one might argue that P(10) is a more accurate model of a typical casual search engine user. If one optimizes on that measure, relevance information from only the top 10 documents will be used. If instead, one optimizes on MAP, relevance information from across the document ranking will be used. Yilmaz and Robertson showed that LTR systems trained on MAP and tested on P(10) produced better rank optimization than systems trained and tested on P(10).

Because the range of parameters in a retrieval functions can be very large, it is impossible to exhaustively explore every possible combination. In order to optimize an LTR system effectively, techniques drawn from the machine learning community, such as gradient ascent, are used. However, Robertson and Zaragoza [201] showed that the current suite of existing evaluation measures are not ideal for use with gradient ascent and related learning techniques. In their paper they argued that new measures need to be built to ensure that optimization can be achieved more successfully. This is likely to be an area that will come to factor more significantly in future evaluation surveys.

7.2 Query Logs — Modeling Users

For as long as automated searching systems existed, logs of activities on those systems were gathered and studied. An early example is Meister and Sullivan [174], who, studying the NASA/RECON citation search engine, examined both the volume of searches and the number of retrieved items that were viewed. Inductive studies of user behavior as recorded in *query logs* continued from that time, growing considerably

with the introduction of web search engines and the selective release of large public data sets from them; see [130] for an overview of that research. Such use of logs in this way was influential in IR researchers' understanding of user behavior, from the shortness of query length to the prevalence of spelling mistakes.

Section 6.3.6 briefly mentioned research exploiting data in logs to help generate conventional test collections: Fox et al. [87] showed that it was possible to use clicks as indicators of relevance. However, more recent research showed that in order to use such data, noise and bias needed to be removed. Noise was introduced to logs by automated programs repeatedly querying a search engine either to gather information or to try to deliberately spam the search engine in some way. Simple methods for identifying the activities of information gathering systems were found to be relatively straightforward: Jansen et al., for example, removed search sessions that had > 100 queries [131]. Detecting spam data, which will be engineered to be as similar to user interactions as possible, is harder to spot. Description of that work is beyond the scope of this monograph.

Bias in the query logs arises from the way that users interact with search engines. Joachims et al. [138] identified two forms of user bias, what they called *trust bias* (in other publications, this was called *presentation bias*) and *quality bias*. Trust bias was given its name due to users' willingness to trust the search engine to find the most relevant item in the top ranks. Joachims demonstrated the strength of this bias by manipulating search results, deliberately placing non-relevant documents in top ranked positions and showing that users still commonly clicked on the top position. With the second form of bias, Joachims showed that when the overall quality of search results was poor, users appeared willing to click on less relevant documents.

Joachims et al.'s conclusions were that extracting absolute relevance judgments from a query log was hard and as an alternative, proposed that relative or *preference judgments* should be extracted. For example, if a user clicked on the item in the 2^{nd} rank position but not the 1^{st}, one would infer that the item at rank 2 was more relevant than the item at rank 1. Joachims later showed that how such preference judgments were used in LTR [139], as did Zheng et al. [294].

Agichtein et al. [1] used query logs to learn how to customize search results for individual users. They removed trust bias from query logs by building a model of the typical bias toward certain rank positions and then subtracted that bias from the query and click log data of the user under study. This work was notable as it was one of the first to use the technique of *click prediction*. Here the researchers split the query log into two parts. They trained their system to a particular user (in this case using the first 75% of the log) and then used the system to predict which result that user would click on for the queries they submitted in the remaining part of the log, thus determining if the user model was accurate See also Piwowarski et al. [190] for further work in this area.

Joachim's observations of bias in user clicks were an initial attempt to model user behavior when examining search results. A series of models were subsequently proposed and tested on extensive collections of search log data often using click prediction. Craswell et al. [72] showed how modeling behavior simply based on document rank was not ideal. They introduced what they referred to as a *cascade model* where the probability of a click on a search result was dependent on the probability of the current result being relevant and of the higher ranked results not being relevant. See Dupret and Piwowaeski [80] and Chapelle et al. [52] for further extension to and testing of the cascade model.

Query logs were also used to validate evaluation measures. Chapelle et al. [52] compared a range of evaluation measures, using a combination of assessor-based relevance judgments and click data from a large query log.

7.3 Live Labs

The involvement of users in evaluation of IR systems has long been advocated and conducted as is recorded and promoted in the works of Ingwersen and Järvelin [128], Saracevic [225], and Borlund [30, 31]. A key limiting factor in the experimental methods promoted by such researchers is the challenge of finding a sufficiency of users. Given there are now very large numbers of people who have high speed access to the Internet, new forms of search evaluation are possible, using what

has sometimes been called *live labs*. This rather broad term covers a range of experimental methodologies, which we outline here.

An early example was the work of Dumais et al. [79] who as part of the testing of their desktop search engine, Stuff I've Seen, deployed a working version of the system, which was installed by 234 users (employees of a large organization) who used the system as their desktop search tool. The search engine was instrumented to log certain information about user interaction, which enabled the researchers to understand how the system was used and how often. Unbeknownst to the employees, the researchers randomly deployed different versions of the search interface and using the logs were able to determine how the versions affected searcher behavior. This approach of deploying software to willing volunteers/users was used by others, e.g., [75].

When working with services accessed over a network, such as a search engine, it is possible to make changes to the searching system at regular intervals without the users of the engine to have to install any updates. Such activity was described at a conference panel by Cutting [163] where he stated that several updates to commercial search engines in a single day was not an unusual occurrence. After each change, search logs could be examined to observe any change in user behavior. Joachims appeared to be the first to publish on this topic (see Section 6.5.3), describing a methodology for measuring user preferences for two different versions of a search engine. A key part of Joachims' approach was that users were unaware they were being given a choice between two different searching systems.[2] In a later paper working with Radlinski et al. [192], Joachims, deployed this methodology to a popular academic paper searching system for a month and was able to observe user behavior for over 20,000 queries.

A number of IR researchers were inspired by von Ahn's ESP game [266], where users label images as part of their activities while playing a multi-user game. Clough et al., [62] keen to study cross language searching created an image finding system built on top of Flickr and

[2] Cooper [67] proposed an evaluation methodology where experimenters would go to the site where an IR system was being used and observe a random sample of users conducting their search tasks on either an existing system or a new trial system. The users would not know which system they were being shown.

through user interactions with the game were able to study interaction. Kazai et al. [150] created a game that involved players making relevance judgments on documents.

While enticing as approaches to creating large-scale evaluations or data sets for evaluation, all three methods are challenging to implement. The first requires the software being deployed to be of a high standard before users will willingly engage with it for a long period of time. The second method requires the experimenter to have access to a popular search engine so as to manipulate its results. The third requires high-quality software to be developed where the game play is enticing enough for a sufficient number of people to participate.

There is, however, another approach, as mentioned in Section 6.3.3, it is possible, using services like Mechanical Turk, to pay people to conduct short-run tasks. The small amount of money they are willing to work for means many people can be employed. The example task described in the earlier section was that of judging the relevance of a document for the purposes of building a test collection, however, the potential range of tasks is broader than this: annotating corpora, seeking user opinion of search interfaces, and comparing result rankings are just some of the possibilities. Exploiting systems like Mechanical Turk for IR research is relatively new with little research to review as yet. There are challenges to using such services, but nevertheless, the service is likely to be increasingly used.

8

Conclusions

This monograph presented a brief history of the development of test collection-based evaluation from the earliest works through to the highly influential TREC exercises. Next a series of prominent evaluation measures were described and research testing the properties of those measures was detailed. The need for and use of significance tests in IR experiments was outlined next. One can see that the IR community still uses the model for evaluation initiated by the pioneering work of Thorne, Cleverdon, and Gull in the early 1950s and consolidated by Cleverdon's Cranfield collections of the early 1960s. Most of the evaluation measures used by the community are closely related to the measures created by Gull and Kent et al. in the 1950s. The commonest significance tests used in research papers today are the same as those used by IR researchers in the late 1960s.

One might expect for the research community to discover flaws in such a long-standing methodology. A great deal of research conducted in the past decade has tried specifically to determine if such flaws exist. However, the results of the research are some new evaluation measures; some useful alternatives to the means by which test collections are built; but ultimately the research has validated the test collection

approach. The components of a test collection — a set of documents, a set of topics, and a list of qrels — while a somewhat artificial construct remains at the core of experimental validation of new methodologies in IR. It is clear that query logs offer a means of constructing noisy though vast testing sets that are particularly helpful in new lines of IR research such as LTR. However, it is likely that this approach will not be a replacement, instead offering a complementary methodology to the long standing and proven approach of measuring the effectiveness of an IR system on a test collection.

Acknowledgments

I am most grateful to Paul Clough and Peter Willett for their helpful comments while preparing this monograph and two of my students Azzah Al-Maskari and Shahram Sedghi, both of whom were working on evaluation topics and whose work and conversations were particularly stimulating while writing. In addition, Evangelos Kanoulas, Ian Soboroff, Chris Buckley, and Ellen Voorhees at TREC were continually helpful in charting out more recent developments in evaluation and being willing to discuss some of the finer (and admittedly quite nerdy) points of IR evaluation. Stefan Rüger, Peter Bath, and Andrew Holmes were a tremendous help in guiding my understanding of significance tests. Finally, I wish to thank the reviewers for their detailed and invaluable comments after examining the earlier versions of this monograph.

Obtaining primary sources in the early history of IR evaluation research would have been much harder had it not been for the invaluable and timely help of Donna Harman, whose formation of a digital library of the key IR texts made access to the old Cornell and Cranfield reports trivially easy. In addition Bill Maron and Keith van Rijsbergen helped me obtain copies of the early reports from Maron, Kuhns,

and Ray. Tefko Saracevic kindly scanned in some pages from one of his early IR reports. Peter Willett and Micheline Beaulieu's personal collection of books, preprints, and technical reports was an invaluable resource also. Finally, my efforts to locate old articles were helped by the long tradition of IR research conducted in my department in Sheffield, which meant that in a dark, lower basement corner of the University of Sheffield Western Bank Library, a wealth of 1940s, 1950s, and 1960s journals, books, proceedings and reports lay in wait for me and my photocopy card.

References

[1] E. Agichtein, E. Brill, S. Dumais, and R. Ragno, "Learning user interaction models for predicting web search result preferences," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3–10, New York, NY, USA: ACM, 2006.

[2] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, "Diversifying search results," in *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pp. 5–14, ACM, 2009.

[3] A. Al-Maskari, M. Sanderson, and P. Clough, "The relationship between IR effectiveness measures and user satisfaction," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 773–774, New York, NY, USA: ACM Press, 2007.

[4] A. Al-Maskari, M. Sanderson, P. Clough, and E. Airio, "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*, pp. 59–66, New York, NY, USA: ACM, 2008.

[5] J. Allan, *Topic Detection and Tracking: Event-based Information Organization,* (The Kluwer International Series on Information Retrieval, vol. 12). Springer, 1st ed., 2002.

[6] J. Allan, B. Carterette, and J. Lewis, "When will information retrieval be "good enough"?," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 433–440, New York, NY, USA: ACM, 2005.

[7] O. Alonso, D. E. Rose, and B. Stewart, "Crowdsourcing for relevance evaluation," *ACM SIGIR Forum*, vol. 42, no. 2, pp. 9–15, 2008.

[8] D. G. Altman, *Practical Statistics for Medical Research*. Chapman & Hall/ CRC, 1st ed., 1990.

[9] E. Amitay, D. Carmel, R. Lempel, and A. Soffer, "Scaling IR-system Evaluation using Term Relevance Sets," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information retrieval*, pp. 10–17, New York, NY, USA: ACM, 2004.

[10] T. Arni, P. Clough, M. Sanderson, and M. Grubinger, "Overview of the ImageCLEFphoto 2008 photographic retrieval task," *Evaluating Systems for Multilingual and Multimodal Information Access*, Lecture Notes in Computer Science, *5706/2009*, 500-511. doi:10.1007/978-3-642-04447-2_62, 2009.

[11] J. Artiles, S. Sekine, and J. Gonzalo, "Web people search: Results of the first evaluation and the plan for the second," in *Proceedings of the 17th International Conference on World Wide Web*, pp. 1071–1072, New York, NY, USA: ACM Press, 2008.

[12] J. A. Aslam, V. Pavlu, and R. Savell, "A unified model for metasearch, pooling, and system evaluation," in *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, pp. 484–491, New York, NY, USA: ACM, 2003.

[13] J. A. Aslam, V. Pavlu, and E. Yilmaz, "A statistical method for system evaluation using incomplete judgments," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 541–548, ACM, 2006.

[14] J. A. Aslam, E. Yilmaz, and V. Pavlu, "A geometric interpretation of r-precision and its correlation with average precision," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 573–574, New York, NY, USA: ACM, 2005.

[15] J. A. Aslam, E. Yilmaz, and V. Pavlu, "The maximum entropy method for analyzing retrieval measures," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 27–34, New York, NY, USA: ACM, 2005.

[16] L. Azzopardi, M. de Rijke, and K. Balog, "Building simulated queries for known-item topics: An analysis using six European languages," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 455–462, New York, NY, USA: ACM, 2007.

[17] L. Azzopardi and V. Vinay, "Retrievability: An evaluation measure for higher order information access tasks," in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pp. 561–570, ACM Press: New York, NY, USA, 2008.

[18] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison Wesley, 1999.

[19] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz, "Relevance assessment: Are judges exchangeable and does it matter," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 667–674, New York, NY, USA: ACM, 2008.

[20] M. Baillie, L. Azzopardi, and I. Ruthven, "A retrieval evaluation methodology for incomplete relevance assessments," in *Advances in Information Retrieval*, Lecture Notes in Computer Science, vol. 4425, pp. 271–282, 2007.

[21] M. Barbaro and T. Zeller Jr, "A face is exposed for AOL Searcher No. 4417749," *The New York Times*. Retrieved from http://www.nytimes.com/2006/08/09/technology/09aol.html, August 9 2006.

[22] J. R. Baron, D. D. Lewis, and D. W. Oard, "TREC-2006 legal track overview," in *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, NIST Special Publication, vol. 500, pp. 79–98, Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, 2006.

[23] S. Bjørner and S. C. Ardito, "Online before the internet, Part 1: Early pioneers tell their stories," *Searcher: The Magazine for Database Professionals*, vol. 11, no. 6, 2003.

[24] D. C. Blair, "STAIRS redux: Thoughts on the STAIRS evaluation, ten years after," *Journal of the American Society for Information Science*, vol. 47, no. 1, pp. 4–22, 1996.

[25] D. C. Blair and M. E. Maron, "An evaluation of retrieval effectiveness for a full-text document-retrieval system," *Communications of the ACM*, vol. 28, no. 3, pp. 289–299, doi:10.1145/3166.3197, 1985.

[26] D. Bodoff and P. Li, "Test theory for assessing IR test collections," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 367–374, New York, NY, USA: ACM, 2007.

[27] T. Bompada, C. C. Chang, J. Chen, R. Kumar, and R. Shenoy, "On the robustness of relevance measures with incomplete judgments," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 359–366, New York, NY, USA: ACM Press, 2007.

[28] A. Bookstein, "When the most "pertinent" document should not be retrieved — An analysis of the Swets model," *Information Processing & Management*, vol. 13, no. 6, pp. 377–383, 1977.

[29] H. Borko, *Evaluating The: Effectiveness of Information Retrieval Systems* (No. Sp-909/000/00). Santa Monica, California: Systems Development Corporation, 1962.

[30] P. Borlund, "The concept of relevance in IR," *Journal of the American Society for Information Science and Technology*, vol. 54, no. 10, pp. 913–925, 2003.

[31] P. Borlund and P. Ingwersen, "The development of a method for the evaluation of interactive information retrieval systems," *Journal of Documentation*, vol. 53, pp. 225–250, 1997.

[32] H. Bornstein, "A paradigm for a retrieval effectiveness experiment," *American Documentation*, vol. 12, no. 4, pp. 254–259, doi:10.1002/asi.5090120403, 1961.

[33] M. Braschler and C. Peters, "Cross-language evaluation forum: Objectives, results, achievements," *Information Retrieval*, vol. 7, no. 1–2, pp. 7–31, 2004.

[34] A. Broder, "A taxonomy of web search," *SIGIR Forum*, vol. 36, no. 2, pp. 3–10, doi:10.1145/792550.792552, 2002.

[35] E. C. Bryant, "Progress towards evaluation of information retrieval systems," in *Information Retrieval Among Examining Patent Offices: 4th Annual*

Meeting of the Committee for International Cooperation in Information Retrieval among Examining Patent Offices (ICIREPAT), pp. 362–377, Spartan Books, Macmillan, 1966.

[36] C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees, "Bias and the limits of pooling for large collections," *Information Retrieval*, vol. 10, no. 6, pp. 491–508, 2007.

[37] C. Buckley and E. M. Voorhees, "Evaluating evaluation measure stability," in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 33–40, New York, NY, USA: ACM, 2000.

[38] C. Buckley and E. M. Voorhees, "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, New York, NY, USA: ACM, 2004.

[39] C. Buckley and E. M. Voorhees, "Retrieval system evaluation," in *TREC: Experiment and Evaluation in Information Retrieval*, pp. 53–75, MIT Press, 2005.

[40] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proceedings of the 22nd International Conference on Machine Learning*, pp. 89–96, Bonn, Germany, 2005.

[41] R. Burgin, "Variations in relevance judgments and the evaluation of retrieval performance," *Information Processing & Management*, vol. 28, no. 5, pp. 619–627, doi:10.1016/0306-4573(92)90031-T, 1992.

[42] S. Büttcher, C. L. A. Clarke, and I. Soboroff, "The TREC 2006 terabyte track," in *Proceedings of the Fifteenth Text Retrieval Conference (TREC 2006)*, vol. 500, pp. 128–141, Maryland, USA: Gaithersburg, 2006.

[43] S. Büttcher, C. L. A. Clarke, P. C. K. Yeung, and I. Soboroff, "Reliable information retrieval evaluation with incomplete and biased judgements," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 63–70, New York, NY, USA: ACM Press, 2007.

[44] F. Can, R. Nuray, and A. B. Sevdik, "Automatic performance evaluation of Web search engines," *Information Processing and Management*, vol. 40, no. 3, pp. 495–514, 2004.

[45] B. Carterette, "Robust test collections for retrieval evaluation," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 55–62, New York, NY, USA: ACM Press, 2007.

[46] B. Carterette, "On rank correlation and the distance between rankings," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 436–443, ACM, 2009.

[47] B. Carterette, J. Allan, and R. Sitaraman, "Minimal test collections for retrieval evaluation," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 268–275, New York, NY, USA: ACM, 2006.

[48] B. Carterette, P. Bennett, D. Chickering, and S. Dumais, "Here or There," in *Advances in Information Retrieval*, pp. 16–27, 2008. Retrieved from http://dx.doi.org/10.1007/978-3-540-78646-7_5.

[49] B. Carterette and R. Jones, "Evaluating search engines by modeling the relationship between relevance and clicks," in *Proceedings of the 21st Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 217–224, 2007.

[50] B. Carterette and R. Jones, "Evaluating search engines by modeling the relationship between relevance and clicks," *Advances in Neural Information Processing Systems*, vol. 20, pp. 217–224, 2008.

[51] B. Carterette, V. Pavlu, E. Kanoulas, J. A. Aslam, and J. Allan, "Evaluation over thousands of queries," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 651–658, New York, NY, USA: ACM, 2008.

[52] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan, "Expected reciprocal rank for graded relevance," in *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 621–630, New York, NY, USA: ACM Press, 2009.

[53] H. Chen and D. R. Karger, "Less is more: Probabilistic models for retrieving fewer relevant documents," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 429–436, New York, NY, USA: ACM, 2006.

[54] C. L. A. Clarke, N. Craswell, and I. Soboroff, "Preliminary report on the TREC 2009 Web track," *Working notes of the proceedings of TREC 2009*, 2009.

[55] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon, "Novelty and diversity in information retrieval evaluation," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 659–666, New York, NY, USA: ACM, 2008.

[56] C. L. A. Clarke, M. Kolla, and O. Vechtomova, "An effectiveness measure for ambiguous and underspecified queries," in *Advances in Information Retrieval Theory: Second International Conference on the Theory of Information Retrieval, ICTIR 2009 Cambridge, UK, September 10–12, 2009 Proceedings*, pp. 188–199, New York Inc: Springer-Verlag, 2009.

[57] C. W. Cleverdon, "The evaluation of systems used in information retrieval (1958: Washington)," in *Proceedings of the International Conference on Scientific Information — Two Volumes*, pp. 687–698, Washington: National Academy of Sciences, National Research Council, 1959.

[58] C. W. Cleverdon, *Report on the Testing and Analysis of an Investigation Into the Comparative Efficiency of Indexing Systems*. ASLIB Cranfield Research Project. Cranfield, UK, 1962.

[59] C. W. Cleverdon, *The Effect of Variations in Relevance Assessments in Comparative Experimental Tests of Index Languages*, (Cranfield Library Report No. 3). Cranfield Institute of Technology, 1970.

[60] C. W. Cleverdon, "The significance of the cranfield tests on index languages," in *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3–12,

Chicago, Illinois, United States: ACM Press New York, NY, USA, 1991. doi:10.1145/122860.122861.

[61] C. W. Cleverdon and M. Keen, "Factors Affecting the Performance of Indexing Systems," Vol 2. *ASLIB, Cranfield Research Project. Bedford, UK: C. Cleverdon*, pp. 37–59, 1966.

[62] P. Clough, J. Gonzalo, J. Karlgren, E. Barker, J. Artiles, and V. Peinado, "Large-scale interactive evaluation of multilingual information access systems — The iCLEF flickr challenge," in *Proceedings of Workshop on Novel Methodologies for Evaluation in Information Retrieval*, pp. 33–38, Glasgow, UK, 2008.

[63] P. Clough, H. Muller, T. Deselaers, M. Grubinger, T. M. Lehmann, J. Jensen, and W. Hersh, "The CLEF 2005 cross-language image retrieval track," in *Accessing Multilingual Information Repositories*, Lecture Notes in Computer Science, vol. 4022, pp. 535–557, 2006.

[64] P. Clough, M. Sanderson, and H. Muller, "The CLEF cross language image retrieval track (ImageCLEF) 2004," *Lecture notes in Computer Science*, pp. 243–251, 2004.

[65] W. S. Cooper, "Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems," *American Documentation*, vol. 19, no. 1, pp. 30–41, doi:10.1002/asi.5090190108, 1968.

[66] W. S. Cooper, "A definition of relevance for information retrieval," *Information storage and retrieval*, vol. 7, no. 1, pp. 19–37, 1971.

[67] W. S. Cooper, "On selecting a measure of retrieval effectiveness," *Journal of the American Society for Information Science*, vol. 24, no. 2, 1973.

[68] G. V. Cormack and T. R. Lynam, "Statistical precision of information retrieval evaluation," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 533–540, New York, NY, USA: ACM, 2006.

[69] G. V. Cormack, C. R. Palmer, and C. L. A. Clarke, "Efficient construction of large test collections," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 282–289, New York, NY, USA: ACM, 1998.

[70] N. Craswell, A. de Vries, and I. Soboroff, "Overview of the trec-2005 enterprise track," in *Proceedings of the Fourteenth Text Retrieval Conference (TREC 2005)*, Gaithersburg, Maryland, USA, 2005.

[71] N. Craswell and D. Hawking, "Overview of the TREC 2002 Web track," in *The Eleventh Text Retrieval Conference (TREC-2002)*, NIST Special Publication 500-251, pp. 86–95, Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, 2003.

[72] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey, "An experimental comparison of click position-bias models," in *Proceedings of the international conference on Web search and web data mining*, pp. 87–94, ACM, 2008.

[73] W. B. Croft, "A file organization for cluster-based retrieval," in *Proceedings of the 1st Annual International ACM SIGIR Conference on Information Storage and Retrieval*, pp. 65–82, New York, NY, USA: ACM, 1978.

[74] W. B. Croft, D. Metzler, and T. Strohman, *Search Engines: Information Retrieval in Practice*. Addison Wesley, 1st ed., 2009.

[75] E. Cutrell, D. Robbins, S. Dumais, and R. Sarin, "Fast, flexible filtering with phlat," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 261–270, New York, NY, USA: ACM, 2006.

[76] A. Davies, *A Document Test Collection for Use in Information Retrieval Research*, (Dissertation). Department of Information Studies. University of Sheffield, 1983.

[77] G. Demartini and S. Mizzaro, "A classification of IR effectiveness metrics," in *Advances in Information Retrieval*, Lecture Notes in Computer Science, vol. 3936, pp. 488–491, 2006.

[78] B. K. Dennis, J. J. Brady, and J. A. Dovel, "Index manipulation and abstract retrieval by computer," *Journal of Chemical Documentation*, vol. 2, no. 4, pp. 234–242, 1962.

[79] S. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins, "Stuff I've seen: A system for personal information retrieval and re-use," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 72–79, ACM, 2003.

[80] G. E. Dupret and B. Piwowarski, "A user browsing model to predict search engine click data from past observations," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 331–338, ACM, 2008.

[81] B. Efron and R. Tibshirani, "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy," *Statistical Science*, vol. 1, no. 1, pp. 54–77, 1986.

[82] B. Efron and R. J. Tibshirani, "An introduction to the bootstrap," *Monographs on Statistics and Applied Probability*, vol. 57, pp. 1–177, 1993.

[83] R. A. Fairthorne, "Implications of test procedures," in *Information Retrieval in Action*, pp. 109–113, Cleveland, Ohio, USA: Western Reserve UP, 1963.

[84] E. M. Fels, "Evaluation of the performance of an information-retrieval system by modified Mooers plan," *American Documentation*, vol. 14, no. 1, pp. 28–34, doi:10.1002/asi.5090140105, 1963.

[85] E. A. Fox, *Characterization of two new experimental collections in computer and information science containing textual and bibliographic concepts*, (Computer Science Technical Reports). Cornell University. Retrieved from http://techreports.library.cornell.edu:8081/Dienst/UI/1.0/Display/cul.cs/TR83-561, 1983.

[86] E. A. Fox, *Virginia Disc One*. Blacksburg, VA, USA: Produced by Nimbus Records, 1990.

[87] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White, "Evaluating implicit measures to improve web search," *ACM Transactions on Information Systems (TOIS)*, vol. 23, no. 2, pp. 147–168, 2005.

[88] H. P. Frei and P. Schäuble, "Determining the effectiveness of retrieval algorithms," *Information Processing and Management: An International Journal*, vol. 27, no. 2–3, pp. 153–164, 1991.

[89] N. Fuhr, "Optimum polynomial retrieval functions based on the probability ranking principle," *ACM Transactions on Information Systems*, vol. 7, no. 3, pp. 183–204, 1989.

[90] N. Fuhr, N. Gövert, G. Kazai, and M. Lalmas, "INEX: INitiative for the Evaluation of XML retrieval," in *Proceedings of the SIGIR 2002 Workshop on XML and Information Retrieval*, 2002.

[91] N. Fuhr and G. E. Knorz, "Retrieval test evaluation of a rule based automatic indexing (AIR/PHYS)," in *Proceedings of the 7th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 391–408, UK: British Computer Society Swindon, 1984.

[92] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees, "The TREC spoken document retrieval track: A success story," in *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*, pp. 107–130, Gaithersburg, Maryland, USA, 2000.

[93] F. Gey, R. Larson, M. Sanderson, H. Joho, P. Clough, and V. Petras, "GeoCLEF: The CLEF 2005 cross-language geographic information retrieval track overview," in *Accessing Multilingual Information Repositories*, Lecture Notes in Computer Science, vol. 4022, pp. 908–919, 2006.

[94] G. Gigerenzer, "Mindless statistics," *Journal of Socio-Economics*, vol. 33, no. 5, pp. 587–606, 2004.

[95] W. Goffman, "On relevance as a measure," *Information Storage and Retrieval*, vol. 2, pp. 201–203, 1964.

[96] R. Grishman and B. Sundheim, "Message Understanding Conference-6: A brief history," in *Proceedings of the 16th Conference on Computational Linguistics — vol. 1*, pp. 466–471, Copenhagen, Denmark: Association for Computational Linguistics. Retrieved from http://portal.acm.org/citation.cfm?id=992628.992709, 1996.

[97] C. D. Gull, "Seven years of work on the organization of materials in the special library," *American Documentation*, vol. 7, no. 4, pp. 320–329, doi:10.1002/asi.5090070408, 1956.

[98] D. K. Harman, "Evaluation issues in information retrieval," *Information Processing & Management*, vol. 28, no. 4, pp. 439–440, 1992.

[99] D. K. Harman, "Overview of the second text retrieval conference (TREC-2)," in *NIST Special Publication. Presented at the Second Text Retrieval Conference (TREC 2)*, Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, 1993.

[100] D. K. Harman, "Overview of the third text retrieval conference (TREC-3)," in *The Third Text Retrieval Conference (TREC-3), Gaithersburg, MD, USA*, NIST Special Publication, Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, 1994.

[101] D. K. Harman, "Overview of the fourth text retrieval conference (TREC-4)," in *The Forth Text Retrieval Conference (TREC-4), Gaithersburg, MD, USA*, NIST Special Publication, Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, 1995.

[102] D. K. Harman, "Overview of the second text retrieval conference (TREC-2)," *Information Processing and Management*, vol. 31, no. 3, pp. 271–289, 1995.

[103] D. K. Harman, "Overview of the TREC 2002 novelty track," in *Proceedings of the Eleventh Text Retrieval Conference (TREC 2002)*, pp. 46–56, Gaithersburg, Maryland, USA, 2002.

[104] D. K. Harman, "Some Interesting Unsolved Problems in Information Retrieval," Presented at the Center for Language and Speech Processing, Workshop 2002, The Johns Hopkins University 3400 North Charles Street, Barton Hall Baltimore, MD 21218. Retrieved from http://www.clsp.jhu.edu/ws02/preworkshop/lecture_harman.shtml, July 2 2002.

[105] D. K. Harman, "The TREC ad hoc experiments," in *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*, pp. 79–98, MIT Press, 2005.

[106] D. K. Harman and C. Buckley, "The NRRC reliable information access (RIA) workshop," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 528–529, New York, NY, USA: ACM, 2004.

[107] D. K. Harman and G. Candela, "Retrieving records from a gigabyte of text on a minicomputer using statistical ranking," *Journal of the American Society for Information Science*, vol. 41, no. 8, pp. 581–589, 1990.

[108] V. Harmandas, M. Sanderson, and M. D. Dunlop, "Image retrieval by hypertext links," in *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 296–303, New York, NY, USA: ACM, 1997.

[109] T. H. Haveliwala, A. Gionis, D. Klein, and P. Indyk, "Evaluating strategies for similarity search on the web," in *Proceedings of the 11th International Conference on World Wide Web*, pp. 432–442, New York, NY, USA: ACM Press, 2002.

[110] D. Hawking, "Overview of the TREC-9 Web track," in *NIST Special Publication*, pp. 87–102, 2001. Presented at the Ninth Text Retrieval Conference (TREC-9), Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology.

[111] D. Hawking, P. Bailey, and N. Craswell, "ACSys TREC-8 Experiments," in *NIST Special Publication*, pp. 307–316, 2000. Presented at the Eighth Text Retrieval Conference (TREC-8), Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology.

[112] D. Hawking, N. Craswell, and P. Thistlewaite, "Overview of TREC-7 very large collection track," in *The Seventh Text Retrieval Conference (TREC-7)*, pp. 91–104, NIST Special Publication, 1998. Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology.

[113] D. Hawking, F. Crimmins, and N. Craswell, "How valuable is external link evidence when searching enterprise Webs?," in *Proceedings of the 15th Australasian database conference*, vol. 27, pp. 77–84, Darlinghurst, Australia: Australian Computer Society, Inc, 2004.

[114] D. Hawking and S. E. Robertson, "On collection size and retrieval effectiveness," *Information Retrieval*, vol. 6, no. 1, pp. 99–105, 2003.

[115] D. Hawking and P. Thistlewaite, "Overview of TREC-6 very large collection track," in *The Sixth Text Retrieval Conference (TREC-6)*, pp. 93–106, NIST Special Publication, 1997. Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology.

[116] M. A. Hearst, *Search User Interfaces*. Cambridge University Press, 1st ed., 2009.

[117] M. A. Hearst and C. Plaunt, "Subtopic structuring for full-length document access," in *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 59–68, New York, NY, USA: ACM, 1993.

[118] W. Hersh, C. Buckley, T. J. Leone, and D. Hickam, "OHSUMED: An interactive retrieval evaluation and new large test collection for research," in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 192–201, New York, NY, USA: Springer-Verlag New York, Inc, 1994.

[119] W. Hersh, A. M. Cohen, P. Roberts, and H. K. Rekapalli, "TREC 2006 genomics track overview," in *The Fifteenth Text Retrieval Conference*, pp. 52–78, Gaithersburg, Maryland, USA, 2006.

[120] W. Hersh, M. K. Crabtree, D. H. Hickam, L. Sacherek, C. P. Friedman, P. Tidmarsh, and C. Mosbaek et al., "Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions," *Journal of American Medical Informatics Association*, vol. 9, 2002.

[121] W. Hersh and P. Over, "TREC-9 Interactive Track Report," in *proceedings of the Ninth Text REtrieval Conference (TREC-9)*, pp. 41–50, Gaithersburg, Maryland: NTIS, 2000.

[122] W. Hersh, A. Turpin, S. Price, B. Chan, D. Kramer, L. Sacherek, and D. Olson, "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*, pp. 17–24, New York, NY, USA: ACM Press, 2000.

[123] J. E. Holmstrom, "Section III. Opening plenary session," in *The Royal Society Scientific Information Conference, 21 June–2 July 1948: Report and papers submitted*, London: Royal Society, 1948.

[124] S. B. Huffman and M. Hochster, "How well does result relevance predict session satisfaction?," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 567–574, New York, NY, USA: ACM Press, 2007.

[125] D. Hull, "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*, pp. 329–338, New York, NY, USA: ACM, 1993.

[126] S. Huuskonen and P. Vakkari, "Students' search process and outcome in Medline in writing an essay for a class on evidence-based medicine," *Journal of Documentation*, vol. 64, no. 2, pp. 287–303, 2008.

[127] E. Ide, "New experiments in relevance feedback," in *Report ISR-14 to the National Science Foundation*, Cornell University, Department of Computer Science, 1968.

[128] P. Ingwersen and K. Järvelin, *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer, 2005.

[129] M. Iwayama, A. Fujii, N. Kando, and A. Takano, "Overview of patent retrieval task at NTCIR-3," in *Proceedings of the third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, 2003.

[130] B. J. Jansen and A. Spink, "How are we searching the World Wide Web? A comparison of nine search engine transaction logs," *Information Processing and Management*, vol. 42, no. 1, pp. 248–263, 2006.

[131] B. J. Jansen, A. Spink, and S. Koshman, "Web searcher interaction with the Dogpile.com metasearch engine," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 5, pp. 744–744, 2007.

[132] K. Järvelin and J. Kekäläinen, "IR evaluation methods for retrieving highly relevant documents," in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 41–48, New York, NY, USA: ACM, 2000.

[133] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 422–446, 2002.

[134] E. Jensen, *Repeatable Evaluation of Information Retrieval Effectiveness In Dynamic Environments*. Illinois Institute of Technology. Retrieved from http://ir.iit.edu/∼ej/jensen_phd_thesis.pdf, May 2006.

[135] E. C. Jensen, S. M. Beitzel, A. Chowdhury, and O. Frieder, "Repeatable evaluation of search services in dynamic environments," *ACM Transactions on Information Systems (TOIS)*, vol. 26, no. 1, p. 1, doi:10.1145/1292591.1292592, 2007.

[136] T. Joachims, "Evaluating retrieval performance using clickthrough data," in *Proceedings of the SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval*, pp. 12–15, 2002.

[137] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 133–142, New York, NY, USA: ACM Press, 2002.

[138] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, "Accurately interpreting clickthrough data as implicit feedback," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 154–161, New York, NY, USA: ACM, 2005.

[139] T. Joachims and F. Radlinski, "Search engines that learn from implicit feedback," *Computer*, pp. 34–40, 2007.

[140] J. Kamps, J. Pehcevski, G. Kazai, M. Lalmas, and S. E. Robertson, "INEX 2007 evaluation measures," pp. 24–33, Retrieved from http://dx.doi.org/10.1007/978-3-540-85902-4_2, 2008.

[141] N. Kando, "Evaluation of information access technologies at the NTCIR workshop," in *Comparative Evaluation of Multilingual Information Access Systems. 4th Workshop of the Cross-Language Evaluation Forum, CLEF*, pp. 29–43, Springer, 2003.

[142] N. Kando, K. Kuriyama, T. Nozue, K. Eguchi, H. Kato, and S. Hidaka, "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, 1999.

[143] P. Kantor and E. M. Voorhees, "The TREC-5 Confusion Track," in *The Fifth Text Retrieval Conference (TREC-5)*, NIST Special Publication.

Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, 1996.

[144] P. B. Kantor and E. M. Voorhees, "The TREC-5 Confusion Track: Comparing Retrieval Methods for Scanned Text," vol. 2, no. 2, pp. 165–176, 2000.

[145] R. V. Katter, "The influence of scale form on relevance judgments," *Information Storage and Retrieval*, vol. 4, no. 1, pp. 1–11, 1968.

[146] J. Katzer, M. J. McGill, J. A. Tessier, W. Frakes, and P. DasGupta, "A study of the overlap among document representations," *Information Technology: Research and Development*, vol. 1, no. 4, pp. 261–274, 1982.

[147] J. Katzer, J. A. Tessier, W. Frakes, and P. DasGupta, *A Study of the Impact of Representations in Information Retrieval Systems*. Syracuse, New York: School of Information Studies, Syracuse University, 1981.

[148] G. Kazai and M. Lalmas, "INEX 2005 evaluation measures," in *Advances in XML Information Retrieval and Evaluation*, Lecture Notes in Computer Science, vol. 3977, pp. 16–29, 2006.

[149] G. Kazai, M. Lalmas, and A. P. de Vries, "The overlap problem in content-oriented XML retrieval evaluation," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 72–79, New York, NY, USA: ACM, 2004.

[150] G. Kazai, N. Milic-Frayling, and J. Costello, "Towards methods for the collective gathering and quality control of relevance assessments," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 452–459, ACM, 2009.

[151] E. M. Keen, "Evaluation parameters," in *Report ISR-13 to the National Science Foundation*, Cornell University, Department of Computer Science, 1967.

[152] E. M. Keen, "Presenting results of experimental retrieval comparisons," *Information Processing and Management: An International Journal*, vol. 28, no. 4, pp. 491–502, 1992.

[153] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1–2, pp. 81–93, 1938.

[154] A. Kent, *Encyclopedia of Library and Information Science*. CRC Press, 2002.

[155] A. Kent, M. M. Berry, F. U. Luehrs Jr, and J. W. Perry, "Machine literature searching VIII. Operational criteria for designing information retrieval systems," *American Documentation*, vol. 6, no. 2, pp. 93–101, doi:10.1002/asi.5090060209, 1955.

[156] K. Kuriyama, N. Kando, T. Nozue, and K. Eguchi, "Pooling for a large-scale test collection: An analysis of the search results from the first NTCIR workshop," *Information Retrieval*, vol. 5, no. 1, pp. 41–59, 2002.

[157] M. Lalmas and A. Tombros, "Evaluating XML retrieval effectiveness at INEX," *SIGIR Forum*, vol. 41, no. 1, pp. 40–57, doi:10.1145/1273221.1273225, 2007.

[158] F. W. Lancaster, *Evaluation of the MEDLARS Demand Search Service*. (No. PB-178-660) (p. 278). Springfield, VA 22151: Clearinghouse for Federal Scientific and Technical Information, 1968.

[159] F. W. Lancaster, *Information Retrieval Systems Characteristics, Testing, and Evaluation*. John Wiley & Sons, Inc, 1968.

[160] R. Ledwith, "On the difficulties of applying the results of information retrieval research to aid in the searching of large scientific databases," *Information Processing & Management*, vol. 28, no. 4, pp. 451–455, doi:10.1016/0306-4573(92)90003-I, 1992.

[161] C. Léger, J. P. Romano, and D. N. Politis, "Bootstrap technology and applications," *Technometrics*, vol. 34, no. 4, pp. 378–398, 1992.

[162] M. E. Lesk, "SIG — The significance programs for testing the evaluation output," in *Report ISR-12 to the National Science Foundation*, Cornell University, Department of Computer Science, 1966.

[163] M. E. Lesk, D. Cutting, J. Pedersen, T. Noreault, and M. Koll, "Real life information retrieval (panel): Commercial search engines," in *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 333, New York, NY, USA: ACM, 1997.

[164] M. E. Lesk and G. Salton, "Relevance assessments and retrieval system evaluation*1," *Information Storage and Retrieval*, vol. 4, no. 4, pp. 343–359, doi:10.1016/0020-0271(68)90029-6, 1968.

[165] D. Lewis, "The TREC-5 filtering track," in *The Fifth Text Retrieval Conference (TREC-5)*, pp. 75–96, Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, 1996.

[166] J. Lin and B. Katz, "Building a reusable test collection for question answering," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 7, pp. 851–861, 2006.

[167] H. Liu, R. Song, J. Y. Nie, and J. R. Wen, "Building a test collection for evaluating search result diversity: A preliminary study," in *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pp. 31–32, 2009.

[168] T. Y. Liu, J. Xu, T. Qin, W. Xiong, and H. Li, "Letor: Benchmark dataset for research on learning to rank for information retrieval," in *Proceedings of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval*, pp. 3–10, 2007.

[169] B. Magnini, A. Vallin, C. Ayache, G. Erbach, A. Peñas, M. De Rijke, and P. Rocha et al., "Overview of the CLEF 2004 multilingual question answering track," *Lecture notes in Computer Science*, vol. 3491, p. 371, 2005.

[170] P. Majumder, M. Mitra, D. Pal, A. Bandyopadhyay, S. Maiti, S. Mitra, and A. Sen et al., "Text collections for FIRE," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 699–700, ACM, 2008.

[171] R. Manmatha, T. Rath, and F. Feng, "Modeling score distributions for combining the outputs of search engines," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 267–275, New York, NY, USA: ACM Press, 2001.

[172] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[173] M. Maron, J. Kuhns, and L. Ray, *Probabilistic Indexing: A Statistical Technique for Document Identification and Retrieval* (Technical Memorandum No. 3) (p. 91), *Data Systems Project Office*. Los Angeles, California: Thompson Ramo Wooldridge Inc, 1959.

[174] D. Meister and D. Sullivan, "Evaluation of User Reactions to a Prototype On-line Information Retrieval System," Prepared under Contract No.

NASw-1369 by Bunker-Ramo Corporation, Canoga Park, CA. (No. NASA CR-918). NASA, 1967.

[175] M. Melucci, "On rank correlation in information retrieval evaluation," *ACM SIGIR Forum*, vol. 41, no. 1, pp. 18–33, 2007.

[176] T. Minka and S. E. Robertson, "Selection bias in the LETOR datasets," in *SIGIR Workshop on Learning to Rank for Information Retrieval*, pp. 48–51, 2008.

[177] S. Mizzaro and S. Robertson, "Hits hits TREC: exploring IR evaluation results with network analysis," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 479–486, New York, NY, USA: ACM, 2007.

[178] A. Moffat, W. Webber, and J. Zobel, "Strategic system comparisons via targeted relevance judgments," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 375–382, Amsterdam, The Netherlands, 2007.

[179] A. Moffat and J. Zobel, "Rank-biased precision for measurement of retrieval effectiveness," *ACM Transactions on Information Systems*, vol. 27, no. 1, p. Article No. 2, 2008.

[180] C. N. Mooers, *The Intensive Sample Test for the Objective Evaluation of the Performance of Information Retrieval System* (No. ZTB-132) (p. 20). Cambridge, Massachusetts: Zator Corporation, 1959.

[181] C. N. Mooers, "The next twenty years in information retrieval: Some goals and predictions," in *Papers Presented at the Western Joint Computer Conference*, pp. 81–86, ACM, 1959.

[182] M. Morita and Y. Shinoda, "Information filtering based on user behavior analysis and best match text retrieval," in *Proceedings of the 17th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 272–281, New York, NY, USA: Springer-Verlag New York, Inc, 1994.

[183] R. Nuray and F. Can, "Automatic ranking of information retrieval systems using data fusion," *Information Processing and Management*, vol. 42, no. 3, pp. 595–614, 2006.

[184] D. W. Oard, D. Soergel, D. Doermann, X. Huang, G. C. Murray, J. Wang, and B. Ramabhadran et al., "Building an information retrieval test collection for spontaneous conversational speech," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 41–48, ACM, 2004.

[185] I. Ounis, M. De Rijke, C. Macdonald, G. Mishne, and I. Soboroff, "Overview of the TREC-2006 blog track," in *Proceedings of the Fifteenth Text Retrieval Conference (TREC 2006)*, pp. 17–31, Gaithersburg, Maryland, USA, 2006.

[186] P. Over, "TREC-6 interactive report," in *Proceedings of the sixth Text Retrieval Conference (TREC-6)*, NIST Special Publication, vol. 500, pp. 73–82, Gaithersburg, Maryland, USA, 1997.

[187] P. Over, "The TREC interactive track: An annotated bibliography," *Information Processing and Management*, vol. 37, no. 3, pp. 369–381, 2001.

[188] P. Over, T. Ianeva, W. Kraaij, A. F. Smeaton, and S. Valencia, "TRECVID 2006-An overview," in *Proceedings of the TREC Video Retrieval Evaluation Notebook Papers*, 2006.

[189] W. R. Pearson, "Comparison of methods for searching protein sequence databases," *Protein Science: A Publication of the Protein Society*, vol. 4, no. 6, p. 1145, 1995.

[190] B. Piwowarski, G. Dupret, and R. Jones, "Mining user web search activity with layered bayesian networks or how to capture a click in its context," in *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pp. 162–171, New York, NY, USA: ACM, 2009.

[191] S. M. Pollock, "Measures for the comparison of information retrieval systems," *American Documentation*, vol. 19, no. 4, pp. 387–397, doi:10.1002/asi.5090190406, 1968.

[192] F. Radlinski, M. Kurup, and T. Joachims, "How does clickthrough data reflect retrieval quality?," in *Proceeding of the 17th ACM Conference on Information and Knowledge Management*, pp. 43–52, 2008.

[193] A. M. Rees and D. G. Schultz, "A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching," Final Report to the National Science Foundation. Volume II, Appendices. Clearinghouse for Federal Scientific and Technical Information, Springfield, VA. 22151 (PB-176-079), MF $0.65, HC $3.00), October 1967.

[194] A. Ritchie, S. Teufel, and S. Robertson, "Creating a test collection for citation-based IR experiments," in *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pp. 391–398, Association for Computational Linguistics Morristown, NJ, USA, 2006.

[195] S. E. Robertson, "The parametric description of retrieval tests: Part II: Overall measures," *Journal of Documentation*, vol. 25, no. 2, pp. 93–107, 1969.

[196] S. E. Robertson, "On sample sizes for non-matched-pair IR experiments," *Information Processing and Management: An International Journal*, vol. 26, no. 6, pp. 739–753, 1990.

[197] S. E. Robertson, "Salton award lecture on theoretical argument in information retrieval," *SIGIR Forum*, vol. 34, no. 1, pp. 1–10, doi:10.1145/373593.373597, 2000.

[198] S. E. Robertson, "On GMAP: And other transformations," in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pp. 78–83, New York, NY, USA: ACM Press, 2006.

[199] S. E. Robertson, "On the history of evaluation in IR," *Journal of Information Science*, vol. 34, no. 4, pp. 439–456, doi:10.1177/0165551507086989, 2008.

[200] S. E. Robertson and D. A. Hull, "The TREC-9 filtering track final report," in *Proceedings of the Ninth Text REtrieval Conference (TREC-2001)*, pp. 25–40, Gaithersburg, Maryland, USA: NTIS, 2001.

[201] S. E. Robertson and H. Zaragoza, "On rank-based effectiveness measures and optimization," *Information Retrieval*, vol. 10, no. 3, pp. 321–339, 2007.

[202] G. Roda, J. Tait, F. Piroi, and V. Zenz, "CLEF-IP 2009: Retrieval experiments in the intellectual property domain," in *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, 2009.

[203] M. E. Rorvig, "The simple scalability of documents," *Journal of the American Society for Information Science*, vol. 41, no. 8, pp. 590–598, doi: 10.1002/(SICI)1097-4571(199012)41:8<590::AID-ASI5>3.0.CO;2-T, 1990.

[204] D. Rose and C. Stevens, "V-twin: A lightweight engine for interactive use," in *Proceedings of the Fifth Text Retrieval Conference (TREC-5)*, pp. 279–290, 1996.

[205] T. Sakai, "New performance metrics based on multigrade relevance: Their application to question answering," in *NTCIR-4 Proceedings*, 2004.

[206] T. Sakai, "Evaluating evaluation metrics based on the bootstrap," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 525–532, Seattle, Washington, USA: ACM Press New York, NY, USA, 2006. doi:10.1145/1148170.1148261.

[207] T. Sakai, "Alternatives to bpref," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 71–78, ACM, 2007.

[208] T. Sakai, "Evaluating information retrieval metrics based on bootstrap hypothesis tests," *Information and Media Technologies*, vol. 2, no. 4, pp. 1062–1079, 2007.

[209] T. Sakai and N. Kando, "On information retrieval metrics designed for evaluation with incomplete relevance assessments," *Information Retrieval*, vol. 11, no. 5, pp. 447–470, doi:10.1007/s10791-008-9059-7, 2008.

[210] G. Salton, *Automatic Information Organization and Retrieval*. McGraw Hill Text, 1968.

[211] G. Salton, *The Smart Retrieval System. Experiments in Automatic Document Processing*. Prentice Hall, 1971.

[212] G. Salton, J. Allan, and C. Buckley, "Approaches to passage retrieval in full text information systems," in *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 49–58, New York, NY, USA: ACM, 1993.

[213] G. Salton, E. A. Fox, and H. Wu, "Extended Boolean information retrieval," *Communications of the ACM*, vol. 26, no. 11, pp. 1022–1036, doi:10.1145/182.358466, 1983.

[214] G. Salton and M. E. Lesk, "Computer evaluation of indexing and text processing," *Journal of the ACM (JACM)*, vol. 15, no. 1, pp. 8–36, 1968.

[215] G. Salton and C. T. Yu, "On the construction of effective vocabularies for information retrieval," in *Proceedings of the 1973 Meeting on Programming Languages and Information Retrieval Table of Contents*, pp. 48–60, New York, NY, USA: ACM, 1973.

[216] M. Sanderson, "Accurate user directed summarization from existing tools," in *Proceedings of the Seventh International Conference on Information and Knowledge Management*, pp. 45–51, New York, NY, USA: ACM, 1998.

[217] M. Sanderson and H. Joho, "Forming test collections with no system pooling," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 33–40, New York, NY, USA: ACM, 2004.

[218] M. Sanderson and C. J. Rijsbergen, "NRT: News retrieval tool," *Electronic Publishing*, vol. 4, no. 4, pp. 205–217, 1991.

[219] M. Sanderson, T. Sakai, and N. Kando EVIA 2007: The First International Workshop on Evaluating Information Access, 2007.

[220] M. Sanderson and I. Soboroff, "Problems with Kendall's tau," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 839–840, New York, NY, USA: ACM, 2007.

[221] M. Sanderson, J. Tang, T. Arni, and P. Clough, "What else is there? Search diversity examined," in *Proceedings of the 31st European Conference on IR Research on Advances in Information Retrieval*, pp. 562–569, Springer, 2009.

[222] M. Sanderson and J. Zobel, "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*, pp. 162–169, New York, NY, USA: ACM, 2005.

[223] T. Saracevic, *An Inquiry into Testing of Information Retrieval Systems: Part II: Analysis of Results*. 1968. (No. CSL:TR-FINAL-II). Comparative Systems Laboratory: Final Technical Report. Center for Documentation and Communication Research, School of Library Science, Case Western Reserve University.

[224] T. Saracevic, "RELEVANCE: A review of and a framework for the thinking on the notion in information science," *Journal of the American Society for Information Science*, vol. 26, no. 6, pp. 143–165, 1975.

[225] T. Saracevic, "Evaluation of evaluation in information retrieval," in *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 138–146, New York, NY, USA: ACM, 1995.

[226] J. Savoy, "Statistical inference in retrieval effectiveness evaluation," *Information Processing and Management*, vol. 33, no. 4, pp. 495–512, 1997.

[227] Y. Shang and L. Li, "Precision evaluation of search engines," *World Wide Web*, vol. 5, no. 2, pp. 159–173, doi:10.1023/A:1019679624079, 2002.

[228] P. Sheridan, M. Wechsler, and P. Schäuble, "Cross-language speech retrieval: Establishing a baseline performance," in *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 99–108, New York, NY, USA: ACM, 1997.

[229] M. Shokouhi and J. Zobel, "Robust result merging using sample-based score estimates," *ACM Transactions on Information Systems (TOIS)*, vol. 27, no. 3, pp. 1–29, 2009.

[230] A. Smeaton and R. Wilkinson, "Spanish and Chinese document retrieval in TREC-5," in *Proceedings of the Fifth Text Retrieval Conference (TREC-5)*, pp. 57–64, Gaithersburg, Maryland, USA, 1997.

[231] A. F. Smeaton, W. Kraaij, and P. Over, "TRECVID-An overview," in *Proceedings of the TRECVID 2003 Conference*, Gaithersburg, Maryland, USA. Retrieved from http://www-nlpir.nist.gov/projects/tvpubs/tvpapers03/tv3overview.pdf, 2003.

[232] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pp. 321–330, New York, NY, USA: ACM, 2006.

[233] A. F. Smeaton, P. Over, and R. Taban, "The TREC-2001 video track report," in *Proceedings of the Tenth Text REtrieval Conference (TREC-2001)*, pp. 52–60, 2001.

[234] C. L. Smith and P. B. Kantor, "User adaptation: Good results from poor systems," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 147–154, New York, NY, USA: ACM, 2008.

[235] M. D. Smucker, J. Allan, and B. Carterette, "A comparison of statistical significance tests for information retrieval evaluation," in *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, pp. 623–632, New York, NY, USA: ACM, 2007.

[236] I. Soboroff, "On evaluating web search with very few relevant documents," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 530–531, 2004.

[237] I. Soboroff, "Dynamic test collections: Measuring search effectiveness on the live web," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 276–283, New York, NY, USA: ACM, 2006.

[238] I. Soboroff, C. Nicholas, and P. Cahan, "Ranking retrieval systems without relevance judgments," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 66–73, New Orleans, Louisiana, United States: ACM. doi:10.1145/383952.383961, 2001.

[239] I. Soboroff and S. E. Robertson, "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*, pp. 243–250, New York, NY, USA: ACM, 2003.

[240] E. Sormunen, "Liberal relevance criteria of TREC-: Counting on negligible documents?," in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 324–330, New York, NY, USA: ACM, 2002.

[241] K. Spärck Jones, "Automatic indexing," *Journal of Documentation*, vol. 30, no. 4, pp. 393–432, 1974.

[242] K. Spärck Jones, *Information Retrieval Experiment*. Butterworth-Heinemann Ltd, 1981.

[243] K. Spärck Jones, "Letter to the editor," *Information Processing & Management*, vol. 39, no. 1, pp. 156–159, doi:10.1016/S0306-4573(02)00026-2, 2003.

[244] K. Spärck Jones and R. G. Bates, *Report on a design study for the 'ideal' information retrieval test collection* (British Library Research and Development Report No. 5428). Computer Laboratory, University of Cambridge, 1977.

[245] K. Spärck Jones and C. J. van Rijsbergen, *Report on the need for and the provision of an 'ideal' information retrieval test collection* (British Library Research and Development Report No. 5266) (p. 43). Computer Laboratory, University of Cambridge, 1975.

[246] K. Spärck Jones and C. J. van Rijsbergen, "Information retrieval test collections," *Journal of Documentation*, vol. 32, no. 1, pp. 59–75, doi:10.1108/eb026616, 1976.

[247] L. T. Su, "Evaluation measures for interactive information retrieval," *Information Processing & Management*, vol. 28, no. 4, pp. 503–516, 1992.

[248] L. T. Su, "The relevance of recall and precision in user evaluation," *Journal of the American Society for Information Science*, vol. 45, no. 3, pp. 207–217, 1994.

[249] J. A. Swets, "Information retrieval systems," *Science*, vol. 141, no. 3577, pp. 245–250, 1963.

[250] J. A. Swets, "Effectiveness of information retrieval methods," *American Documentation*, vol. 20, no. 1, pp. 72–89, doi:10.1002/asi.4630200110, 1969.

[251] R. Tagliacozzo, "Estimating the satisfaction of information users," *Bulletin of the Medical Library Association*, vol. 65, no. 2, pp. 243–249, 1977.

[252] J. M. Tague and M. J. Nelson, "Simulation of user judgments in bibliographic retrieval systems," in *Proceedings of the 4th Annual International ACM SIGIR Conference on Information Storage and Retrieval: Theoretical Issues in Information Retrieval*, pp. 66–71, New York, NY, USA: ACM, 1981.

[253] J. M. Tague-Sutcliffe, "Some perspectives on the evaluation of information retrieval systems," *Journal of the American Society for Information Science*, vol. 47, no. 1, pp. 1–3, doi:10.1002/(SICI)1097-4571(199601)47:1<1::AID-ASI1>3.0.CO;2-3, 1996.

[254] J. M. Tague-Sutcliffe and J. Blustein, "A statistical analysis of the TREC-3 data," in *The Third Text Retrieval Conference (TREC-3), Gaithersburg, MD, USA*, pp. 385–398, NIST Special Publication, 1994. Department of Commerce, National Institute of Standards and Technology.

[255] J. A. Thom and F. Scholer, "A comparison of evaluation measures given how users perform on search tasks," *Presented at the Proceedings of the Twelfth Australasian Document Computing Symposium*, pp. 56–63, 2007.

[256] P. Thomas and D. Hawking, "Evaluation by comparing result sets in context," in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pp. 94–101, New York, NY, USA: ACM Press, 2006.

[257] R. Thorne, "The efficiency of subject catalogues and the cost of information searches," *Journal of Documentation*, vol. 11, pp. 130–148, 1955.

[258] A. Turpin and W. Hersh, "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*, pp. 225–231, New York, NY, USA: ACM, 2001.

[259] A. Turpin and W. Hersh, "User interface effects in past batch versus user experiments," in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 431–432, New York, NY, USA: ACM, 2002.

[260] A. Turpin and F. Scholer, "User performance versus precision measures for simple search tasks," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 11–18, New York, NY, USA: ACM, 2006.

[261] C. J. van Rijsbergen, "Foundation of evaluation," *Journal of Documentation*, vol. 30, no. 4, pp. 365–373, 1974.

[262] C. J. van Rijsbergen, *Information Retrieval*. Butterworth-Heinemann Ltd, 2nd ed., 1979.

[263] P. K. T. Vaswani and J. B. Cameron, *The National Physical Laboratory Experiments in Statistical Word Associations and Their Use in Document Indexing And Retrieval.* National Physical Laboratory Computer Science Division-Publications; COM.SCI.42 (p. 171). National Physical Lab., Teddington (Great Britain), 1970.

[264] J. Verhoeff, W. Goffman, and J. Belzer, "Inefficiency of the use of Boolean functions for information retrieval systems," *Communications of the ACM*, vol. 4, no. 12, pp. 557–558, 1961.

[265] B. C. Vickery, *On Retrieval System Theory.* Butterworths, 2nd ed., 1965.

[266] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 319–326, New York, NY, USA: ACM Press, 2004.

[267] E. M. Voorhees, "On expanding query vectors with lexically related words," in *The Second Text Retrieval Conference (TREC 2)*, NIST Special Publication 500-215, pp. 223–231, Department of Commerce, National Institute of Standards and Technology, 1993.

[268] E. M. Voorhees, "Variations in relevance judgments and the measurement of retrieval effectiveness," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in information retrieval*, pp. 315–323, New York, NY, USA: ACM Press, 1998.

[269] E. M. Voorhees, "The TREC-8 question answering track report," in *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*, pp. 77–82, Gaithersburg, Maryland, USA, 1999.

[270] E. M. Voorhees, "Variations in relevance judgments and the measurement of retrieval effectiveness," *Information Processing and Management*, vol. 36, no. 5, pp. 697–716, 2000.

[271] E. M. Voorhees, "Evaluation by highly relevant documents," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 74–82, New Orleans, Louisiana, United States. New York, NY, USA: ACM Press, 2001.

[272] E. M. Voorhees, "Overview of the TREC 2003 question answering track," in *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003),* vol. 142, 2003.

[273] E. M. Voorhees, "Overview of the TREC 2004 robust retrieval track," in *The Thirteenth Text Retrieval Conference (TREC 2004)*, NIST Special Publication. Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, 2005.

[274] E. M. Voorhees, "On test collections for adaptive information retrieval," *Information Processing and Management*, 2008.

[275] E. M. Voorhees, "Topic set size redux," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 806–807, ACM, 2009.

[276] E. M. Voorhees and C. Buckley, "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*, pp. 316–323, New York, NY, USA: ACM, 2002.

[277] E. M. Voorhees and D. K. Harman, "Overview of the seventh text retrieval conference," in *Proceedings of the Seventh Text Retrieval Conference (TREC-7)*, pp. 1–24, NIST Special Publication, 1998.

[278] E. M. Voorhees and D. K. Harman, "Overview of the eighth text retrieval conference (TREC-8)," in *The Eighth Text Retrieval Conference (TREC-8)*, pp. 1–24, NIST Special Publication, 1999. Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology.

[279] E. M. Voorhees and D. K. Harman, "Overview of TREC 2001," in *NIST Special Publication 500-250*, pp. 1–15, Presented at the Tenth Text Retrieval Conference (TREC 2001), Government Printing Office, 2001.

[280] E. M. Voorhees and D. K. Harman, *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press, illustrated ed., 2005.

[281] C. Wade and J. Allan, *Passage Retrieval and Evaluation* (CIIR Technical Report No. IR-396). Amherst, MA, USA: University of Massachusetts, Amherst Center for Intelligent Information Retrieval, 2005.

[282] W. Webber, A. Moffat, and J. Zobel, "Score standardization for inter-collection comparison of retrieval systems," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 51–58, New York, NY, USA: ACM, 2008.

[283] W. Webber, A. Moffat, and J. Zobel, "Statistical power in retrieval experimentation," in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pp. 571–580, ACM, 2008.

[284] R. W. White and D. Morris, "Investigating the querying and browsing behavior of advanced search engine users," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 255–262, New York, NY, USA: ACM Press, 2007.

[285] D. Williamson, R. Williamson, and M. E. Lesk, "The Cornell implementation of the SMART system," in *The SMART Retrieval System: Experiments in Automatic Document Processing*, (G. Salton, ed.), p. 12, Englewood Cliffs, New Jersey: Prentice-Hall, Inc, 1971.

[286] I. H. Witten, A. Moffat, and T. C. Bell, *Managing Gigabytes*. Morgan Kaufmann, 1999.

[287] S. Wu and F. Crestani, "Methods for ranking information retrieval systems without relevance judgments," in *Proceedings of the 2003 ACM symposium on Applied computing*, pp. 811–816, New York, NY, USA: ACM, 2003.

[288] S. Xu, S. Bao, B. Fei, Z. Su, and Y. Yu, "Exploring folksonomy for personalized search," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 155–162, New York, NY, USA: ACM, 2008.

[289] E. Yilmaz and J. A. Aslam, "Estimating average precision with incomplete and imperfect judgments," in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pp. 102–111, New York, NY, USA: ACM Press, 2006.

[290] E. Yilmaz, J. A. Aslam, and S. E. Robertson, "A new rank correlation coefficient for information retrieval," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 587–594, New York, NY, USA: ACM, 2008.

[291] E. Yilmaz and S. E. Robertson, "On the choice of effectiveness measures for learning to rank," in *Learning to Rank for Information Retrieval. Workshop in Conjunction with the ACM SIGIR Conference on Information Retrieval*, Boston, MA, USA: ACM Press New York, NY, USA, 2009.

[292] T. Zeller Jr, "AOL Moves to Increase Privacy on Search Queries," *The New York Times*, Retrieved from http://www.nytimes.com/2006/08/22/technology/22aol.html, August 22 2006.

[293] C. X. Zhai, W. W. Cohen, and J. Lafferty, "Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 10–17, New York, NY, USA: ACM Press, 2003.

[294] Z. Zheng, K. Chen, G. Sun, and H. Zha, "A regression framework for learning ranking functions using relative relevance judgments," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 287–294, ACM, 2007.

[295] J. Zobel, "How reliable are the results of large-scale information retrieval experiments?," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 307–314, New York, NY, USA: ACM Press, 1998.

Index