Creating a Test Collection for Citation-based IR Experiments

Anna Ritchie
University of Cambridge
Computer Laboratory
15 J J Thompson Avenue
Cambridge, CB3 0FD, U.K.
ar283@cl.cam.ac.uk

Simone Teufel
University of Cambridge
Computer Laboratory
15 J J Thompson Avenue
Cambridge, CB3 0FD, U.K.
sht25@cl.cam.ac.uk

Stephen Robertson
Microsoft Research Ltd
Roger Needham House
7 J J Thomson Avenue
Cambridge, CB3 0FB, U.K.
ser@microsoft.com

Abstract

We present an approach to building a test collection of research papers. The approach is based on the Cranfield 2 tests but uses as its vehicle a current conference; research questions and relevance judgements of all cited papers are elicited from conference authors. The resultant test collection is different from TREC's in that it comprises scientific articles rather than newspaper text and, thus, allows for IR experiments that include citation information. The test collection currently consists of 170 queries with relevance judgements; the document collection is the ACL Anthology. We describe properties of our queries and relevance judgements, and demonstrate the use of the test collection in an experimental setup. One potentially problematic property of our collection is that queries have a low number of relevant documents; we discuss ways of alleviating this.

1 Introduction

We present a methodology for creating a test collection of scientific papers that is based on the Cranfield 2 methodology but uses a current conference as the main vehicle for eliciting relevance judgements from users, i.e., the authors.

Building a test collection is a long and expensive process but was necessary as no ready-made test collection existed on which the kinds of experiments with citation information that we envisage could be run. We aim to improve term-based IR on scientific articles with citation information, by using index terms from the citing article to additionally describe the cited document. Exactly how to do this is the research question that our test collection should help to address.

This paper is structured as follows: Section 2 motivates our proposed experiments and, thereby, our test collection. Section 3 discusses the how test collections are built and, in particular, our own. Section 4 briefly describes the practicalities of compiling the document collection and the processing we perform to prepare the documents for our experiments. In Section 5, we show that our test collection can be used with standard IR tools. Finally, Section 6 discusses the problem of the low number of relevant documents judged so far and two ways of alleviating this problem.

2 Motivation

The idea of using terms external to a document, coming from a 'citing' document, has been borrowed from web-based IR. When one paper cites another, a link is made between them and this *link structure* is analogous to that of the web: "hyperlinks ... provide semantic linkages between objects, much in the same manner that citations link documents to other related documents" (Pitkow and Pirolli, 1997). Link structure, particularly anchor text, has been used to advantage in web-based IR. While web pages are often poorly self-descriptive (Brin and Page, 1998) anchor text is often a higher-level description of the pointed-to page. (Davison,

2000) provides a good discussion of how well anchor text does this and provides experimental results in support. Thus, beginning with (McBryan, 1994), there is a trend of propagating anchor text along its hyperlink to associate it with the linked page, as well as the page in which it is found. Google, for example, includes anchor text as index terms for the linked page (Brin and Page, 1998). The TREC Web tracks have also shown that using anchor text improves retrieval effectiveness for some search tasks (Hawking and Craswell, 2005).

This idea has already been applied to citations and scientific articles (Bradshaw, 2003). In Bradshaw's experiment, scientific documents are indexed by the text that refers to them in documents that cite them. However, unlike in experiments with previous collections, we need both the citing and the cited article as full documents in our collection. The question of how to identify citation 'anchor text' and its extent is a matter for research; this requires the full text of the citing article. Previous experiments and test collections have had only limited access to the content of the citing article: Bradshaw had access only to a fixed window of text around the citation, as provided by CiteSeer's 'citation context'; in the GIRT collections (Kluck, 2003), a dozen or so content-bearing information fields (e.g., title, abstract, methodological descriptors) represent each document and the full text is not available. Additionally, in Bradshaw's experiment, no access is given to the text of the *cited* article itself so that the influence of a term-based IR model cannot be studied and so that documents can only be indexed if they have been cited at least once. A test collection containing full text for many citing and cited documents, thus, has advantages from a methodological point of view.

2.1 Choosing a Genre

When choosing a scientific field to study, we looked for one that is practicable for us to compile the document collection (freely available machine-readable documents; as few as possible document styles), while still ensuring good coverage of research topics in an entire field. Had we chosen the medical field or bioinformatics, the prolific number of journals would have been a problem for the practical document preparation.

We also looked for a relatively self-contained field. As we aim to propagate referential text to cited papers as index terms, references from documents in the collection to other documents *within* the collection will be most useful. We call these *internal* references. While it is impossible to find or create a collection of documents with only internal references, we aim for as high a proportion of internal references as possible.

We chose the ACL (Association for Computational Linguistics) Anthology[1], a freely available digital archive of computational linguistics research papers. Computational linguistics is a small, homogenous research field and the Anthology contains the most prominent publications since the beginning of the field in 1960, consists of only 2 journals, 7 conferences and 5 less important publications, such as discontinued conferences and a series of workshops, resulting in only 7000 papers[2].

With the ACL Anthology, we expect a high proportion of internal references within a relatively compact document collection. We empirically measured the proportion of collection-internal references. We found a proportion of internal references to all references of 0.33 (the *in-factor*). We wanted to compare this number to a situation in another, larger field (genetics) but no straightforward comparison is possible, as there are very many genetics journals and quality of journals probably plays a larger role in a bigger field. We tried to simulate a similar collection to the 9 main journals+conferences in the Anthology, by considering 10 journals in genetics with a range of impact factors[3], resulting in an in-factor of 0.17 (dropping to 0.14 if only 5 journals are considered). Thus, our hypothesis that the Anthology is reasonably self-contained, at least in comparison with other possible collections, was confirmed.

The choice of computational linguistics has the added benefit that we are familiar with the domain; we can interpret the subject matter better than we would be able to in the medical domain. This should be of use to us in our eventual experiments.

[1] http://www.aclweb.org/anthology/

[2] This is our estimate, after substracting non-papers such as letters to the editor, tables of contents etc. The Anthology is growing by $\sim$500 papers per year.

[3] Journal impact factor is a measure of the frequency with which its average article is cited and is a measure of the relative importance of journals within a field (Garfield, 1972).

3 Building Test Collections

To turn our document collection into a test collection, a parallel set of search queries and relevance judgements is needed. There are a number of alternative methods for building a test collection. For TREC, humans devise queries specifically for a given set of documents and make relevance judgements on pooled retrieved documents from that set (Harman, 2005). Theirs is an extremely labour-intensive and expensive process and an unrealistic option in the context of our project.

The Cranfield 2 tests (Cleverdon et al., 1966) introduced an alternative method for creating a test collection, specifically for scientific texts. The method was subject to criticism and has not been employed much since. Nevertheless, we believe this method to be worth revisiting for our current situation. In this section, we describe in turn the Cranfield 2 method and our adapted method. We discuss some of the original criticisms and their bearing on our own work, then describe our returns thus far.

3.1 The Cranfield 2 Test Collection

The Cranfield 2 tests (Cleverdon et al., 1966) were a comparative evaluation of indexing language devices. From a base collection of 182 (high speed aerodynamics and aircraft structures) papers, the Cranfield test collection was built by asking the authors to formulate the research question(s) behind their work and to judge how relevant each reference in their paper was to each of their research questions, on a 5-point scale. Referenced documents were obtained and added to the base set. Authors were also asked to list additional relevant papers not cited in their paper. The collection was further expanded in a second stage, using bibliographic coupling to search for similar papers to the referenced ones and employing humans to search the collection for other relevant papers. The resultant collection comprised 1400 documents and 221 queries (Cleverdon, 1997).

The principles behind the Cranfield technique are:

- Queries: Each paper has an underlying research question or questions; these constitute valid search queries.
- Relevant documents: A paper's reference list is a good starting point for finding papers relevant to its research questions.
- Judges: The paper author is the person best qualified to judge relevance.

3.2 Our Anthology Test Collection

We altered the Cranfield design to fit to a fixed, existing document collection. We designed our methodology around an upcoming conference and approached the paper authors at around the time of the conference, to maximize their willingness to participate and to minimise possible changes in their perception of relevance since they wrote the paper. Due to the relatively high in-factor of the collection, we expected a significant proportion of the relevance judgements gathered in this way to be about Anthology documents and, thus, useful as evaluation data.

Hence, the authors of accepted papers for ACL-2005 and HLT-EMNLP-2005 were asked, by email, for their research questions and relevance judgements for their references. We defined a 4-point relevance scale, c.f. Table 1, since we felt that the distinctions between the Cranfield grades were not clear enough to warrant 5. Our guidelines also included examples of referencing situations that might fit each category. Personalized materials for participation were sent, including a reproduction of their paper's reference list in their response form. This meant that invitations could only be sent once the paper had been made available online.

We further deviated from the Cranfield methodology by deciding not to ask the authors to try to list additional references that could have been included in their reference list. An author's willingness to name such references will differ more from author to author than their naming of original references, as referencing is part of a standardized writing process. By asking for this data, the consistency of the data across papers will be degraded and the status of any additional references will be unclear. Furthermore, feedback from an informal pilot study conducted on ten paper authors confirmed that some authors found this task particularly difficult.

Each co-author of the papers was invited individually to participate, rather than inviting the first author alone. This increased the number of invitations that needed to be prepared and sent (by a factor of around 2.5) but also increased the likelihood of getting a return for a given paper. Furthermore, data from multiple co-authors of the same paper can be used to

Grade	Description
4	The reference is crucially relevant to the problem. Knowledge of the contents of the referred work will be fundamental to the reader's understanding of your paper. Often, such relevant references are afforded a substantial amount of text in a paper e.g., a thorough summary.
3	The reference is relevant to the problem. It may be helpful for the reader to know the contents of the referred work, but not crucial. The reference could not have been substituted or dropped without making significant additions to the text. A few sentences may be associated with the reference.
2	The reference is somewhat (perhaps indirectly) relevant to the problem. Following up the reference probably would not improve the reader's understanding of your paper. Alternative references may have been equally appropriate (e.g., the reference was chosen as a representative example from a number of similar references or included in a list of similar references). Or the reference could have been dropped without damaging the informativeness of your paper. Minimal text will be associated with the reference.
1	The reference is irrelevant to this particular problem.

Table 1: Relevance Scale

measure co-author agreement on the relevance task. This is an interesting research question, as it is not at all clear how much even close collaborators would agree on relevance, but we do not address this here.

We plan to expand the collection in a second stage, in line with the Cranfield 2 design. We will reapproach contributing authors after obtaining retrieval results on our collection (e.g., with a standard IR engine) and ask them to make additional relevance judgements on these papers.

3.3 Criticisms of Cranfield 2

Both Cranfield 1 (Cleverdon, 1960) and 2 were subject to various criticisms; (Spärck Jones, 1981) gives an excellent account of the tests and their criticisms. The majority were criticisms of the test collection paradigm itself and are not pertinent here. However, the *source-document principle* (i.e., the use of queries created from documents in the collection) attracted particular criticisms. The fundamental concern was that the way in which the queries were created led to "an unnaturally close relation" between the terms in the queries and those used to index the documents in the colection (Vickery, 1967); any such relationship might have created a bias towards a particular indexing language, distorting the comparisons that were the goal of the project.

In Cranfield 1, system success was measured by retrieval of source documents alone, criticized for being an over-simplification and a distortion of 'real-life' searching. The evaluation procedure was changed for Cranfield 2 so that source documents were excluded from searches and, instead, retrieval of other relevant documents was used to measure success. This removed the problem that, usually, when a user searches, there is no source document for their query. Despite this, Vickery notes that there were "still verbal links between sought document and question" in the new method: each query author was asked to judge the relevance of the source document's references and "the questions ... were formulated *after* the cited papers had been read and has possibly influenced the wording of his question".

While adapting the Cranfield 2 method to our needs, we have tried to address some of the criticisms, e.g., that authors' relevance judgements change over time. Nevertheless, we still have source-document queries and must consider the associated criticisms. Firstly, our test collection is not intended for comparisons of indexing languages. Rather, we aim to compare the effect of adding extra index terms to a base indexing of the documents. The source documents will have no influence on the base indexing of a document above that of the other documents. The additional index terms, coming from citations to that document, will generally be 'chosen' by someone other than the query author, with no knowledge of the query terms[4]. Also, our documents will be indexed fully automatically, further diminishing the scope of any subconscious human influence.

Thus, we believe that the suspect relationship between queries and indexing is negligible in the con-

[4]The exception to this is self-citation. This (very indirectly) allows the query author to influence the indexing but it seems highly improbable that an author would be thinking about their query whilst citing a previous work.

text of our work, as opposed to the Cranfield tests, and that the source-document principle is sound.

3.4 Returns and Analysis

Out of around 500 invitations sent to conference authors, 85 resulted in research questions with relevance judgements being returned; 235 queries in total. Example queries are:

- *Do standard probabilistic parsing techniques, developed for English, fare well for French and does lexicalistion help improve parsing results?*
- *Analyze the lexical differences between genders engaging in telephone conversations.*

Of the 235 queries, 18 were from authors whose co-authors had also returned data and were discarded (for retrieval purposes); we treat co-author data on the same paper as 'the same' and keep only the first authors'. 47 queries had no relevant Anthology-internal references and were discarded. Another 15 had only relevant Anthology references not yet included in the archive[5]; we keep these for the time being. This leaves 170 unique queries with at least 1 relevant Anthology reference and an average of 3.8 relevant Anthology references each. The average in-factor across queries is 0.42 (similar to our previously estimated Anthology in-factor)[6].

Our average number of judged relevant documents per query is lower than for Cranfield, which had an average of 7.2 (Spärck Jones et al., 2000). However, this is the final number for the Cranfield collection, arrived at after the second stage of relevance judging, which we have not yet carried out. Nevertheless, we must anticipate a potentially low number of relevant documents per query, particularly in comparison to, e.g., the TREC ad hoc track (Voorhees and Harman, 1999), with 86.8 judged relevant documents per query.

4 Document Collection and Processing

The Anthology documents are distributed in PDF, a format designed to visually render printable documents, not to preserve editable text. So the PDF collection must be converted into a fully textual format.

[5]HLT-NAACL-2004 papers, e.g., are listed as 'in process'.

[6]We cannot directly compare this to Cranfield's in-factor as we do not have access to the documents.

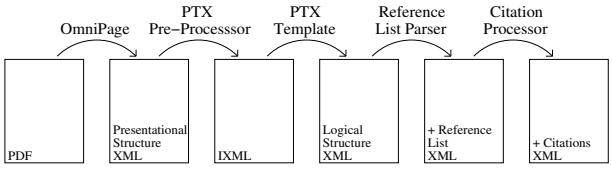

Figure 1: Document Processing Pipeline

A pipeline of processing stages has been developed in the framework of a wider project, illustrated in Figure 1.

Firstly, OmniPage Pro 14[7], a commercial PDF processing software package, scans the PDFs and produces an XML encoding of character-level page layout information. AI algorithms for heuristically extracting character information (similar to OCR) are necessary since many of the PDFs were created from scanned paper-copies and others do not contain character information in an accessible format.

The OmniPage output describes a paper as text blocks with typesetting information such as font and positional information. A pre-processor (Lewin et al., 2005) filters and summarizes the OmniPage output into Intermediate XML (IXML), as well as correcting certain characteristic errors from that stage. A journal-specific template converts the IXML to a logical XML-based document structure (Teufel and Elhadad, 2002), by exploiting low-level, presentational, journal-specific information such as font size and positioning of text blocks.

Subsequent stages incrementally add more detailed information to the logical representation. The paper's reference list is annotated in more detail, marking up individual references, author names, titles and years of publication. Finally, a citation processor identifies and marks up citations in the document body and their constituent parts, e.g., author names and years.

5 Preliminary Experimentation

We expect that our test collection, built for our citation experiments, will be of wider value and we intend to make it publicly available. As a sanity check on our data so far, we carried out some preliminary experimentation, using standard IR tools: the Lemur Toolkit[8], specifically Indri (Strohman et al., 2005),

[7]http://www.scansoft.com/omnipage/

[8]http://www.lemurproject.org/

its integrated language-model based search engine, and the TREC evaluation software, trec_eval[9].

5.1 Experimental Set-up

We indexed around 4200 Anthology documents. This is the total number of documents that have, at the time of writing, been processed by our pipeline (24 years of CL journal, 25 years of ACL proceedings, 14 years of assorted workshops), plus another ∼90 documents for which we have relevance judgements that are not currently available through the Anthology website but should be incorporated into the archive in the future. The indexed documents do not yet contain annotation of the reference list or citations in text. 19 of our 170 queries have no relevant references in the indexed documents and were not included in these experiments. Thus, Figure 2 shows the distribution of queries over number of relevant Anthology references, for a total of 151 queries.

Our Indri index was built using default parameters with no optional processing, e.g., stopping or stemming, resulting in a total of 20117410 terms, 218977 unique terms and 2263 'frequent'[10] terms.

We then prepared an Indri-style query file from the conference research questions. The Indri query language is designed to handle highly complex queries but, for our very basic purposes, we created simple bag-of-words queries by stripping all punctuation from the natural language questions and using Indri's `#combine` operator over all the terms. This means Indri ranks documents in accordance with query likelihood. Again, no stopping or stemming was applied.

Next, the query file was run against the Anthology index using IndriRunQuery with default parameters and, thus, retrieving 1000 documents for each query.

Finally, for evaluation, we converted the Indri's ranked document lists to TREC-style top_results file and the conference relevance judgements compiled into a TREC-style qrels file, including only judgements corresponding to references within the indexed documents. These files were then input to trec_eval, to calculate precision and recall metrics.

[9]http://trec.nist.gov/trec_eval/trec_eval.8.0.tar.gz

[10]Terms that occur in over 1000 documents.

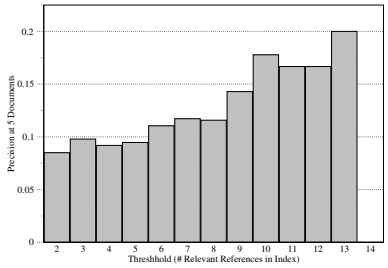

Figure 3: Effect of Thresholding on P at 5 Docs

5.2 Results and Discussion

Out of 489 relevant documents, 329 were retrieved within 1000 (per query) documents. The mean average precision (MAP) was 0.1014 over the 151 queries. This is the precision calculated at each relevant document retrieved (0.0, if that document is not retrieved), averaged over all relevant documents for all queries, i.e., non-interpolated. R-precision, the precision after R (the number of relevant documents for a query) documents are returned, was 0.0965. The average precision at 5 documents was 0.0728.

We investigated the effect of excluding queries with lower than a threshold number of judged relevant documents. Figure 3 shows that precision at 5 documents increases as greater threshold values are applied. Similar trends were observed with other evaluation measures, e.g., MAP and R-precision increased to 0.2018 and 0.1528, respectively, when only queries with 13 or more relevant documents were run, though such stringent thresholding does result in very few queries. Nevertheless, these trends do suggest that the present low number of relevant documents has an adverse effect on retrieval results and is a potential problem for our test collection.

We also investigated the effect of including only authors' main queries, as another potential way of objectively constructing a 'higher quality' query set. Although, this decreased the average in-factor of relevant references, it did, in fact, increase the average absolute number of relevant references in the index. Thus, MAP increased to 0.1165, precision at 5 documents to 0.1016 and R-precision to 0.1201.

These numbers look poor in comparison to the performance of IR systems at TREC but, importantly, they are not intended as performance results. Their purpose is to demonstrate that such numbers *can* be produced using the data we have collected,

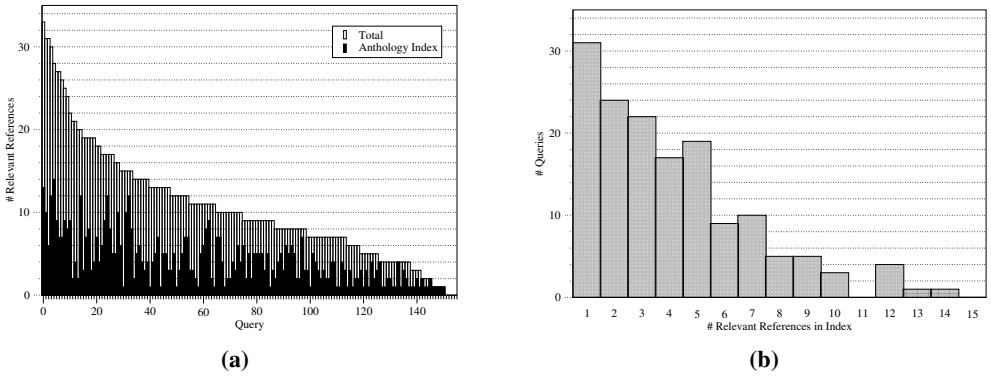

Figure 2: **(a)** Relevant References Per Query and **(b)** Distribution of Queries over Number of Relevant References

rather than to evaluate the performance of some new retrieval system or strategy.

A second point for consideration follows directly from the first: our experiments were carried out on a new test collection and "different test collections have different intrinsic difficulty" (Buckley and Voorhees, 2004). Thus, it is meaningless to compare statistics from this data (from a different domain) to those from the TREC collections, where queries and relevance judgements were collected in a different way, and where there are very many relevant documents.

Thirdly, our experiments used only the most basic techniques and the results could undoubtedly be improved by, e.g., applying a simple stop-list. Nevertheless, this notion of intrinsic difficulty means that it may be the case that evaluations carried out on this collection will produce characteristically low precision values.

Low numbers do not necessarily preclude our data's usefulness as a test collection, whose purpose is to facilitate comparative evaluations. (Voorhees, 1998) states that "To be viable as a laboratory tool, a [test] collection must reliably rank different retrieval variants according to their true effectiveness" and defends the Cranfield paradigm (from criticisms based on relevance subjectivity) by demonstrating that the relative performance of retrieval runs is stable despite differences in relevance judgements. The underlying principle is that it is not the absolute precision values that matter but the ability to compare these values for different retrieval techniques or systems, to investigate their relative benefits. A test collection with low precision values will still allow this.

It is known that all evaluation measures are unstable for very small numbers of relevant documents (Buckley and Voorhees, 2000) and there are issues arising from incomplete relevance information in a test collection (Buckley and Voorhees, 2004). This makes the second stage of our test collection compilation even more indispensable (asking subjects to judge retrieved documents), as this will increase the number of judged relevant documents, as well as bridging the completeness gap.

There are further possibilities of how the problem could be countered. We could exclude queries with lower than a threshold number of relevant documents (after the second stage). Given the respectable number of queries we have, we might be able to afford this luxury. We could add relevant documents from outside the Anthology to our collection. This is least preferable methodologically: using the Anthology has the advantage that it has a real identity and was created for real reasons outside our experiments. Furthermore, the collection 'covers a field', i.e., it includes all important publications and only those. By adding external documents to the collection, it would lose both these properties.

6 Conclusions and Future Work

We have presented an approach to building a test collection from an existing collection of research papers and described the application of our method to the ACL Anthology. We have collected 170 queries with relevance data, centered around the ACL-2005 and HLT-EMNLP-2005 conferences. We

have sanity-checked the usability of our data by running the queries through a retrieval system and evaluating the results using standard software. The collection currently has a low number of judged relevant documents and further experimentation is needed to determine if this poses a real problem.

We plan a second stage of collecting relevance judgements, in line with the original Cranfield design, whereby authors who have contributed queries will be asked to judge the relevance of documents in retrieval rankings from standard IR models and, ideally, from our eventual citation-based experiments.

Nevertheless, our test collection is likely to suffer from incomplete relevance information. The bpref measure (Buckley and Voorhees, 2004) gauges retrieval effectiveness solely on the basis of judged documents and is more stable to differing levels of completeness than measures such as MAP, R-precision or precision at fixed document cutoffs. Thus, bpref may offer a solution to the incompleteness problem and we intend to investigate its potential use in our future evaluations.

When finished, we hope our test collection will be a generally useful IR resource. In particular, we expect the collection to be useful for experimentation with citation information, for which there is currently no existing test collection with the properties that ours offers.

Acknowledgements Thanks to the reviewers for their useful comments and to Karen Spärck Jones for many instructive discussions.

References

Shannon Bradshaw. 2003. Reference directed indexing: Redeeming relevance for subject search in citation indexes. In *Research and Advanced Technology for Digital Libraries (ECDL)*, pages 499–510.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117.

Chris Buckley and Ellen Voorhees. 2000. Evaluating evaluation measure stability. In *Research and Development in Information Retrieval (SIGIR)*.

Chris Buckley and Ellen Voorhees. 2004. Retrieval evaluation with incomplete information. In *Research and development in information retrieval (SIGIR)*.

Cyril Cleverdon, Jack Mills, and Michael Keen. 1966. Factors determining the performance of indexing sytems, volume 1. design. Technical report, ASLIB Cranfield Project.

Cyril Cleverdon. 1960. Report on the first stage of an investigation into the comparative efficiency of indexing systems. Technical report, ASLIB Cranfield Project.

Cyril Cleverdon. 1997. The Cranfield tests on index language devices. In *Readings in information retrieval*, pages 47–59. Morgan Kaufmann Publishers Inc.

Brian D. Davison. 2000. Topical locality in the web. In *Research and Development in Information Retrieval (SIGIR)*, pages 272–279.

Eugene Garfield. 1972. Citation analysis as a tool in journal evaluation. *Science*, 178 (4060):471–479.

Donna Harman. 2005. The TREC test collections. In Ellen Voorhees and Donna Harman, editors, *TREC Experiment and Evaluation in Information Retrieval*, chapter 2. MIT Press.

David Hawking and Nick Craswell. 2005. The very large collection and web tracks. In Ellen Voorhees and Donna Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, chapter 9. MIT Press.

Michael Kluck. 2003. The GIRT data in the evaluation of CLIR systems - from 1997 until 2003. In *CLEF*, pages 376–390.

Ian Lewin, Bill Hollingsworth, and Dan Tidhar. 2005. Retrieving hierarchical text structure from typeset scientific articles - a prerequisite for e-science text mining. In *UK e-Science All Hands Meeting*.

Oliver McBryan. 1994. GENVL and WWWW: Tools for taming the web. In *World Wide Web Conference*.

James Pitkow and Peter Pirolli. 1997. Life, death, and lawfulness on the electronic frontier. In *Human Factors in Computing Systems*.

Karen Spärck Jones, Steve Walker, and Stephen Robertson. 2000. A probabilistic model of information retrieval: development and comparative experiments - parts 1 and 2. *Information Processing and Management*, 36(6):779–840.

Karen Spärck Jones. 1981. The Cranfield tests. In Karen Spärck Jones, editor, *Information Retrieval Experiment*, chapter 13, pages 256–284. Butterworths.

Trevor Strohman, Donald Metzler, Howard Turtle, and W. Bruce Croft. 2005. Indri: a language-model based search engine for complex queries. Technical report, University of Massachusetts.

Simone Teufel and Noemie Elhadad. 2002. Collection and linguistic processing of a large-scale corpus of medical articles. In *Language Resources and Evaluation Conference (LREC)*.

B. C. Vickery. 1967. Reviews of CLEVERDON, C. W., MILLS, J. and KEEN, E. M. the Cranfield 2 report. *Journal of Documentation*, 22:247–249.

Ellen Voorhees and Donna Harman. 1999. Overview of the eighth Text REtrieval Conference (TREC 8). In *Text REtrieval Conference (TREC)*.

Ellen Voorhees. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Research and Development in Information Retrieval (SIGIR)*, pages 315–323.