Creating a Test Collection: Relevance Judgements of Cited & Non-cited Papers

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

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

Abstract

We investigate the effect of different sources of relevant documents in the creation of a test collection in the scientific domain. Based on the Cranfield 2 design, paper authors are asked to judge their cited papers for relevance in the first stage. In a second stage, documents outside the reference list are judged. In this paper, we use the test collection with standard IR engines to compare the information contained in the judgements of the first vs second stage. Using different correlation studies, we found that the judgements of the cited papers do not predict those from the non-cited papers, which means that the combination of sources results in a higher quality collection.

1 Introduction

Building a test collection is a long and expensive process but is sometimes necessary when no ready-made collection with the right properties exists. We aim to improve term-based IR on scientific papers with citation information, by using terms from the citing document to additionally describe (i.e., index) the cited document. We needed a test collection with full text for many citing and cited documents. A high proportion of citations from documents in the collection to other collection documents will be most useful; we built our test collection around the ACL Anthology[1], since we empirically found Computational Linguistics to be a relatively self-contained field.

The idea of using terms external to a document for indexing, coming from a 'citing' document, is also used in web IR. Citations are quite like hyperlinks and link structure, particularly anchor text, has been used to advantage in retrieval tasks (McBryan, 1994; Hawking and Craswell, 2005). 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). Some work has been done in this area, e.g., (Bradshaw, 2003; Dunlop and van Rijsbergen, 1993). However, previous experiments and test collections have had only limited access to the content of the citing and/or cited documents: (Bradshaw, 2003) found index terms in Citeseer *citation contexts* rather than full texts, (Dunlop and van Rijsbergen, 1993) experimented on the CACM collection of abstracts and the GIRT collection (Kluck, 2003), likewise, consists of content-bearing fields, not full documents. The original TREC Genomics collection[2] consists of MEDLINE records, containing abstracts but not full papers (Hersh and Bhupatiraju, 2003). Our test collection must contain full text for many

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

[2] In the 2006 track, a new collection of full-text documents was introduced but this was not available when our work began (Hersh *et al.*, 2006). Its suitability as a test collection for citation-related work, e.g., the proportion of internal citations, has not yet been established.

citing and cited documents. It should, thus, help to address the research question of how to use citations between documents for IR.

To turn a 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). This is too labour-intensive for our project, particularly as we use scientific papers as data, where deciding on relevance would take even more time than for newspaper text. We, instead, adapted the methodology from the Cranfield 2 tests (Cleverdon *et al.*, 1966), which is specific to scientific texts.

The Cranfield test collection was built by asking 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. From a base collection of 182 (high speed aerodynamics and aircraft structures) papers, referenced documents were obtained and added. 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(s); 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.

The *source-document principle* (i.e., using queries created from documents in the collection) attracted criticism: the fact that the queries were formulated after the cited papers had been read may have influenced the wording of the queries and, thus, led to a bias towards one particular indexing language (Vickery, 1967). While this may be true, it is far more a problem for Cranfield 2 (which investigated indexing devices per se) than for us, as the indexing language will be kept constant in our experiments. For our purposes, we assume that the source-document principle is sound.

We adapted the Cranfield method to fit a fixed, existing document collection. We designed our methodology around an upcoming (ACL Anthology) 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. 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. Personalised 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.

This resulted in a test collection of 196 queries; however, we commented that the low number of judged relevant documents is potentially problematic (Ritchie *et al.*, 2006). In line with Cranfield,

Class	Description and *Example*
Typo	Corrected spelling or typographical error in the research question, as returned by the author. *Handling biograpical questions with implicature in a question answering system.* → *Handling biographical questions with implicature in a question answering system.*
Filler	Removed part(s) of the research question that did not contribute to its meaning, e.g., contentless 'filler' phrases or repetitions of existing content. *We present a novel mechanism for improving reference resolution by using the output of a relation tagger to rescore coreference hypotheses.* → *improving reference resolution by using the output of a relation tagger to rescore coreference hypotheses.*
Anaphor	Resolved anaphoric references in the research question to ideas introduced in earlier questions from the same author. *How can the best alignment according to the model be found?* → *How can the best word-alignment according to the weighted linear model be found?*
Context	Added terms from earlier research questions to provide apparently missing context. *Identifying an appropriate domain* → *Identifying an appropriate domain - natural language generation*

Table 1: Classes of Query Reformulation

therefore, we expanded our test collection to add judgements for non-cited papers. In §2, we present our methodology for this expansion, which we call Phase Two. We briefly survey the relevance data accumulated via our methods. In §3, we describe using our test collection with standard IR tools, comparing results before and after the judgement set is expanded. §4 concludes and outlines future work.

2 Expanding Our Test Collection

Whereas the Cranfield expansion also involved adding more documents to the collection, the purpose of our Phase Two was solely to obtain more relevance judgements for the queries from Phase One. Our methodology was as follows.

First, we inspected the research questions returned in Phase One and noted that some were unsuitable as search queries. Mostly, these were artefacts of the method by which the queries were created: we did not explicitly ask the authors for independent search queries. Thus, where an author had returned multiple research questions, the later questions sometimes contained anaphoric references to earlier ones or did not include terms describing the background context of the research (that had been introduced in an earlier question). In addition, some questions contained spelling or typographical errors and some were formulated elaborately or verbosely, with many terms that did not contribute to the underlying meaning, e.g., contentless 'filler' phrases or repetitions of existing content. While a good retrieval system should be robust to query imperfections, this is outside the domain of our research. Therefore, we minimally reformulated 34 of the 196 research questions, to turn them into error-free, standalone queries, while keeping them as close to the author's original research question as possible. Authors were asked to approve our reformulations (i.e., confirm that the reformulated query corresponded to their intentions) or to correct the query, for resubmission to the pooling process. Table 1 describes the four classes of query reformulation. We note that some number of the Cranfield queries were similarly reformulated (Cleverdon *et al.*, 1966).

For each query, we next constructed a list of potentially relevant documents in the Anthology. We

first 'manually' searched the entire Anthology using the Google Search facility on the Anthology website. We started with the author's complete research question (or our reformulation) as the search query then used successive query refinements or alternatives. These query changes were made depending on the relevance of search results, i.e., relevance according to our intuitions about the query meaning and guided, where necessary, by the author's Phase One judgements. Our manual searches were not strictly manual in the same sense of the Cranfield manual searches: we did use an automated search tool rather than search through papers by hand. We use the term 'manual' to indicate the significant human involvement in the searches.

We then ran the queries through three 'standard' IR models, implemented in Lemur[3], with standard parameters:

1. Okapi BM25 with relevance feedback
2. KL-divergence LM with relevance feedback and document model smoothing
3. Cosine similarity

We pooled the manual and automatic search results, including all manual search results and adding one from each of the automatic retrieved lists (removing duplicates) to make a list of fifteen documents. If there were fifteen or more manual search results, only manual results (and all of these) were included, as these were felt to be more 'trustworthy', having already been judged as likely to be relevant. Some lists were, thus, longer than fifteen documents.

The list of potentially relevant documents was then included in personalised materials and sent to the query author for judgement. The materials included instructions and a response form in both plaintext and PDF, including the URL for a webpage with identifying details about the papers for relevance judgement (i.e., title and authors) and links to the papers in PDF, to aid the relevance decision.

We decided to ask for binary relevance judgements for this second round. Firstly, the relevance scale used in Phase One was designed for the specific task of grading the relevance of referenced papers in relation to the research question underlying the source paper; the grades were described in terms of how important the information in that reference would be to someone reading the paper. Judging the relevance of papers from outside the reference list is a slightly different task, therefore, and would have required a translation of the relevance scale. It was not clear that an exactly equivalent set of grades could have been formulated, such that a Phase One grade 4 was equivalent to a Phase Two grade 4 etc. Furthermore, it was already unclear whether we would be able to make use of the graded relevance judgements from Phase One, since most of the standard evaluation measures use binary relevance, without the added complication of having a new set of graded judgements that weren't straightforwardly interchangable.

A switch to binary judgements, however, raises a similar question: how do we know that the threshold between relevant and irrelevant in Phase Two corresponds to the same threshold for Phase One? Indeed, how do we know that the threshold corresponds to the boundary between any two of the Phase One grades? In short, we do not know. However, graded judgements have

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

Statistic	All Phase One	T_1	T_{1+2}	Cranfield 2	TREC 8	TREC Robust
# Queries	196	82	82	221	150	50
Mean # Judgements Per Query (Rel)	4.5	4.8	11.4	7.0	94	131.2
Mean # Judgements Per Query (Irrel)	3.3	3.4	12.3	4.1	1642	624.74
# Documents	9084	9084	9084	1400	528000	1033000

Table 2: Test Collection Comparison

been collapsed in previous studies and shown to give stable evaluation results (Voorhees, 1998). Additionally, in our case, the binary and graded judgements are made by the same person so we might conjecture that their judgement thresholds are more consistent. Therefore, we changed to binary judgements, in the hope that this would also make the task easier for the authors and encourage a higher response rate.

2.1 Returns and Analysis

Around 500 invitations were sent in Phase One. 85 completed response forms were returned, giving 235 queries with relevance judgements. We discarded queries from co-authors whose first author had also responded and queries with no relevant Anthology-internal references, leaving 196 queries, henceforth the All Phase One set.

74 invitations were sent in Phase Two and 44 forms were returned; 82 queries[4] . 22 of these had been reformulated and all were approved by the author except two. In both cases, the author submitted an alternative reformulation for pooling and a new list (including the previous manual search results) was sent back for judgement. Both authors judged the (non-duplicate) documents in the new list.

Table 2 compares our test collection, before and after Phase Two, to some other test collections. T_{1+2} is the complete test collection, i.e., the set of queries for which we have both Phase One and Two judgements and all those judgements. T_1 represents the T_{1+2} collection prior to Phase Two, i.e., the same queries but with only Phase One judgements. After Phase Two, the average number of judged relevant documents per query is 11.4, higher than for Cranfield, which had an average of 7.0 (Cleverdon *et al.*, 1966). It is still low in comparison to, e.g., the TREC ad hoc track, with an average of 94 judged relevant documents per query (Voorhees and Harman, 1999).

However, the scientific aspect of the collection makes it very different in nature from TREC, with its newswire articles and purpose-made queries. Intuitively, because most scientific queries are very specific, we do not expect a large number of relevant documents per query. A more appropriate modern comparison might be with TREC Robust (Voorhees, 2005), whose queries are selected precisely for being 'hard'. Furthermore, the document collection is also small in comparison to TREC and this possibly influences the absolute number of relevant documents per query. We have 1.14 judged relevant documents per thousand documents, compared with 0.18 for TREC 8 ad hoc and 0.13 for TREC Robust[5].

[4]In fact, judgements were returned for 83 queries, including one discarded query with no relevant Anthology Phase One judgements, mistakenly processed in Phase Two.

[5]Counted from http://trec.nist.gov/data/t14_robust.html.

Document Source	Judged			Uniquely Found			Uniquely Found (No Manual)		
	Total	Rel	Irrel	Total	Rel	Irrel	Total	Rel	Irrel
All	23.71	11.39	12.32	21.13	10.74	10.39	9.79	3.70	6.10
→ Phase One	8.20	4.82	3.38	6.76	3.52	3.23	5.56	2.44	3.12
→ Phase Two	15.51	6.57	8.94	14.38	7.22	7.16	4.23	1.26	2.98
→ → Manual	10.99	5.22	5.62	10.41	6.06	4.35	-	-	-
→ → Automatic	4.04	1.10	2.94	3.96	1.16	2.80	4.23	1.26	2.98

Table 3: Average # of Judged Documents By Source (T_{1+2} Queries)

Document Source	ACL Anthology Judged			Cranfield 2 Judged		
	Total	Rel	Irrel	Total	Rel	Irrel
All	23.71	11.39	12.32	11.1	7.0	4.1
→ Phase One	8.20	4.82	3.38	7.1	4.5	2.6
→ Phase Two	15.51	6.57	8.94	4.0	2.5	1.5
→ → Manual	10.99	5.22	5.62	3.3	2.1	1.2
→ → Automatic	4.04	1.10	2.94	0.7	0.4	0.3

Table 4: Average # of Judged Documents By Source (Comparison with Cranfield)

Table 3 shows how many of the judged documents were found by the various methods, for our 82 queries. On average, 23.71 documents per query were judged throughout the entire procedure and 11.39 of these were judged relevant. 4.8 relevant documents were judged during Phase One (i.e., 'found' in the reference list of the query source document). Of the additional relevant documents judged during Phase Two, 5.22 were found by manual searching, compared to 1.10 by the automatic searches.

It would be unfair to conclude from these statistics that the automatic searches were less effective than the manual ones at finding relevant documents: they reflect which single method first resulted in the documents being found, irrespective of 'later' methods that might also have found them, and manual search results were prioritised over automatic ones when compiling the judgement lists. These numbers, thus, do not support a direct comparison of the effectiveness of manual vs automatic searches.

However, the Uniquely Found columns give the numbers of judged documents that were *only* found by one method. On average, 6.06 documents per query were uniquely found by manual searching, compared to 1.16 by automatic searching. The rightmost columns further consider the situation if we had not performed the manual searches and, instead, created the list of fifteen documents for relevance judgement from the automatic lists alone. In this case, Phase Two would have (uniquely) found only 1.26 of the judged relevant documents per query.

This does not take into account potentially relevant documents found by this method that were never judged (in favour of manual documents) but the ratio of relevant to irrelevant automatic documents (compared to manual) makes it doubtful that as many relevant documents would have been found this way. More relevant documents may also have been found by increasing the number of documents sent to the author for judgement but at the expense of increasing the difficulty of the task and potentially decreasing overall returns. This seems to vindicate our decision to expend the significant effort involved in the manual searches (around 80 person-hours).

A rough comparison can be made with Cranfield. In Table 4, Phase One denotes the respective reference list judgement stages and Phase Two the overall expansion stages, comprised of Manual and Automatic searching (where the Cranfield automatic searches were based on bibliographic coupling). While the numbers of relevant documents found in Phase One are very similar, our Phase Two contributed notably more relevant documents. In particular, our manual searches found over three relevant documents per query more than the equivalent Cranfield searches.

3 Experiments

In order to compare the effect on (perceived) retrieval performance when using the extended judgement set, we carried out some experimentation, using standard IR tools: the Lemur Toolkit and the TREC evaluation software, trec_eval[6].

3.1 Experimental Set-up

We indexed 9084 Anthology documents, with stopping and stemming. This is the total number of documents that we had processed from PDF to XML. We also removed certain classes of document, e.g., letters to the editor, book reviews and non-English papers. The resultant index has 37,758,643 terms, 325,693 unique terms and 2320 'frequent'[7] terms.

We ran our 82 queries against the index using various retrieval models as implemented in Lemur: the cosine similarity (Cosine), Okapi BM25 (Okapi), KL-divergence LM-based (KL) and Indri LM/inference network (Indri) models. In each run, 100 documents were retrieved per query. The output from each retrieval run was evaluated twice using trec_eval; first, using only the Phase One judgement (J_1) TREC-style qrels and, next, using both Phase One and Two judgements (J_{1+2}) qrels.

We also performed retrieval runs using relevance feedback with Okapi and KL, allowing 20 feedback terms. Runs using J_1 for feedback and then J_{1+2} were carried out. Note that using the same (or some of the same) judged documents for feedback as for evaluation is an unrealistic experiment. However, it does allow us to investigate the effect of adding more judgements for feedback.

3.2 Results and Discussion

Table 5 summarizes the retrieval results. Each row gives some standard performance measures for one retrieval (model) run; mean average precision (MAP), precision at 5 documents (P(5)), R-precision (R-P), geometric mean average precision (GMAP) and bpref. The values of these measures when evaluated using the Phase One only and Phase One and Two judgement sets are given in the J_1 and J_{1+2} subcolumns, respectively.

The test collection, as it stood after Phase One, had a very low number of judged relevant docu-

[6] http://trec.nist.gov/trec_eval/trec_eval.8.1.tar.gz
[7] Terms that occur in over 1000 documents.

Retrieval Model	MAP		P(5)		R-P		GMAP		bpref	
	$\mathbf{J_1}$	$\mathbf{J_{1+2}}$	$\mathbf{J_1}$	$\mathbf{J_{1+2}}$	$\mathbf{J_1}$	$\mathbf{J_{1+2}}$	$\mathbf{J_1}$	$\mathbf{J_{1+2}}$	$\mathbf{J_1}$	$\mathbf{J_{1+2}}$
Indri	0.1182	0.1592	0.1049	0.2610	0.1322	0.1823	0.0090	0.0370	0.3570	0.3107
Cosine	0.0704	0.1275	0.0537	0.2024	0.0700	0.1431	0.0039	0.0282	0.3314	0.2654
Okapi	0.0717	0.0830	0.0439	0.1146	0.0577	0.0957	0.0008	0.0024	0.2565	0.2000
KL	0.1112	0.1654	0.1049	0.2683	0.1194	0.1857	0.0088	0.0399	0.3725	0.3209
Okapi FB (J_1)	0.6816	0.3425	0.4634	0.5171	0.6447	0.3589	0.3863	0.1833	0.8371	0.4526
Okapi FB (J_{1+2})	0.3367	0.4121	0.2634	0.5171	0.3083	0.4230	0.1078	0.1985	0.7081	0.6064
KL FB (J_1)	0.2362	0.2303	0.1610	0.3244	0.2430	0.2554	0.0580	0.1040	0.5627	0.3939
KL FB (J_{1+2})	0.1519	0.2300	0.1317	0.3220	0.1451	0.2450	0.0278	0.1023	0.4842	0.3934

Table 5: Evaluation Results using J_1 versus J_{1+2}

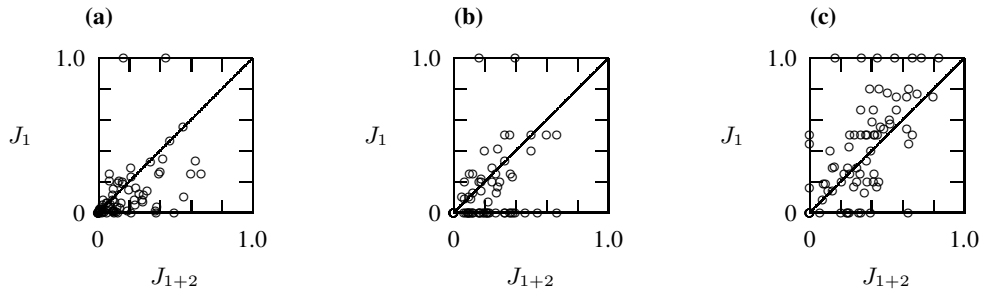

Figure 1: **(a)** MAP, **(b)** R-P and **(c)** bpref for KL, Evaluated with J_1 versus J_{1+2}

ments per query; we conjectured that such incomplete relevance information had an adverse effect on perceived retrieval performance (Ritchie *et al.*, 2006). These results confirm that adding more relevance information does generally increase the absolute values of the performance measures. The exception to this trend is bpref, where we observe a drop in performance using J_{1+2} compared to J_1. This is expected: bpref (Buckley and Voorhees, 2004) is a measure that is more robust to incomplete relevance data, since it is calculated using only judged documents and does not assume that unjudged documents are irrelevant. By design, absolute bpref values should not differ significantly when using partial judgement sets. In practice (according to their analysis on TREC data), this is true until the level of completeness drops below a certain point ($\sim$40% for TREC 8), when average bpref values begin to rise. This suggests that J_1 may be too small a fraction of J_{1+2} to be representative.

Alternatively, and more fundamentally worryingly, the J_1 judgements may not be representative by nature, i.e., because their source is the author's reference list. Cited documents may well have a particular relationship with the query that other relevant documents do not have. Regardless of the reason, though, the drop in bpref shows that J_1 is not a representative sample of relevant documents. In other words, the Phase Two judgements are necessary to more accurately evaluate performance.

Figure 1 shows how the MAP, P(5) and bpref values (from the basic KL run) are affected per query when Phase Two judgements are added for evaluation. The other runs exhibited the same trends, except Okapi FB: the Lemur documentation notes a suspected bug in the implementation of Okapi feedback because "performance is not as expected". Indeed, the Okapi FB J_1 run in particular, behaved very anomalously. We intend to investigate this.

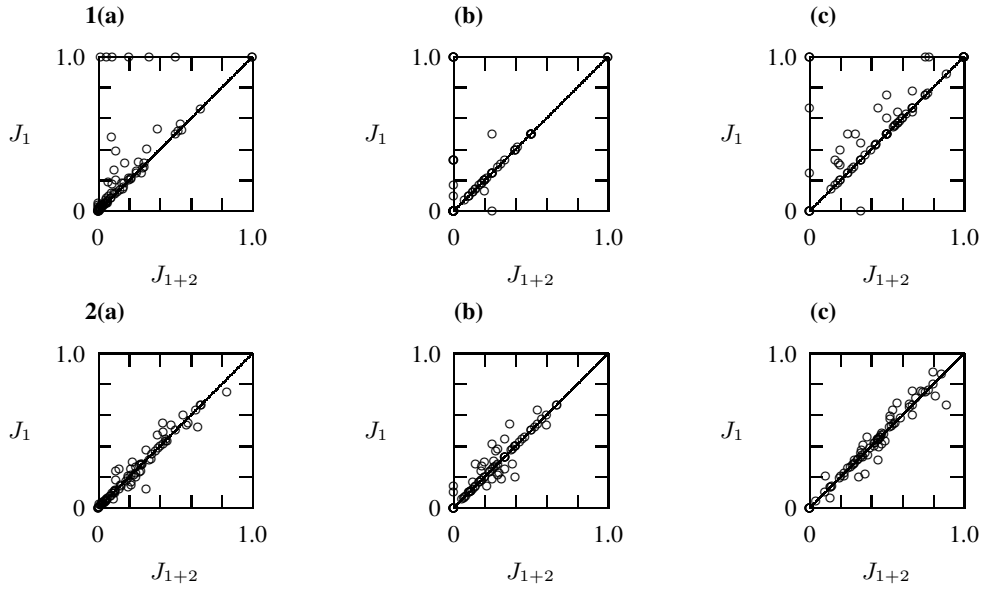

Figure 2: **(a)** MAP, **(b)** R-P and **(c)** bpref for KL with J_1 versus J_{1+2} Feedback, Evaluated with **(1)** J_1 and **(2)** J_{1+2}

It is not the case that performance is uniformly increased across queries (or decreased, in the case of bpref). There are some queries whose MAP and R-P values fall from 1.0 when evaluated with J_{1+2}. However, the queries concerned had only one judged relevant document in J_1 (and this was successfully retrieved) and additional relevant documents in J_{1+2} (not all of which were retrieved). Such cases are good examples of where the smaller judgement set almost certainly cannot reliably indicate performance.

There are also queries where performance was 0 using J_1 and is increased using J_{1+2}, indicating that the query's J_1 relevant documents were not retrieved but that other relevant documents judged during Phase Two were retrieved. The average results in Table 5 confirm that most queries experience a positive change in MAP and R-P (i.e., lie below the diagonal) and a negative change in bpref (i.e., lie above it). In general, though, the points are quite scattered; the J_1 judgements do not predict performance using J_{1+2} very well on a per query basis.

Figure 2 shows how performance on each query is affected when the extended judgement set is used for feedback. The values plotted are from the KL with FB runs evaluated using J_1 on the top row and J_{1+2} on the bottom. We observe that, for certain queries, J_1 feedback evaluated with J_1 gives excellent performance values, that drop substantially when J_{1+2} is used for feedback. Good performance given the same judged documents for feedback and evaluation is unsurprising, as noted earlier. The drop in performance when different documents are used for feedback can be accounted for by the fact that a different query model has been generated, that does not perfectly match the documents used for evaluation.

The same trend might therefore be expected using J_{1+2} evaluation, i.e., the model generated by J_{1+2} feedback would encapsulate documents in J_{1+2} (but not in J_1) and boost their retrieval. However, the J_{1+2} plots have no outliers; the points are clustered neatly round the diagonal. In fact, Table 5 shows a (probably insignificant) decrease in average performance across queries, e.g.,

MAP changes from 0.2303 to 0.2300. In other words, J_{1+2} feedback seems to learn a good, general model for each query, that performs equally well on J_1 and J_{1+2} documents, whereas using J_1 for feedback, for some queries, produces a model that overfits the data. This is another argument in favour of Phase Two: for some queries, there are too few judgements in J_1 to learn a sufficiently general model from feedback.

4 Conclusions and Future Work

We have presented a methodology for building a test collection based on the Cranfield 2 method, being a combination of relevance judgements from two sources: cited and non-cited papers. We have experimentally demonstrated the advantages of a combination of the two. In the expansion phase, potentially relevant non-cited papers were found through both manual and automatic searching. The expansion was costly in many ways. Planning, preparation and execution all took time and effort, not least the significant effort involved in manually searching the document collection for relevant documents. We are satisfied that manual searching contributed enough unique relevant documents to make it a worthwhile addition to automatic pooling. Nevertheless, since not every query author participated in the second stage, the resultant test collection has far fewer queries than previously (82 vs. 196), though there are more relevance judgements per query (11.4 vs. 4.5). We have empirically investigated the possibility that the quality of the test collection was sufficient (for IR experiments) without expansion. Our results suggest that the original judgements are not representative enough of all relevant documents to be able to accurately gauge performance.

There may simply be too few judged relevant documents in the original set for them to be a representative sample. Alternatively, it may be a side effect of the way in which they were selected for judgement; each document judged in the first stage of the test collection came from the reference list of the query source document. Documents cited in a paper have a particular relationship with that paper and are relevant to it in a particular way, though other documents may be relevant in a different way. Perhaps these documents alone cannot represent all other relevant documents.

This is a particularly interesting issue for us, since our intended research centres around the use of citations for retrieval. Are judged documents from the query source document's reference list easier to retrieve using citation-based methods? Do such methods have an unfair advantage when evaluated using a test collection with those judgements? We intend to investigate this issue and having the new relevance judgements will allow us to do so. Furthermore, if the reference list judgements *do* introduce a significant bias towards citation-based methods, having the additional judgements allows us the possibility of discarding the reference list judgements.

In conclusion, we believe that the time and effort involved in expanding our test collection was well spent. In our opinion, the collection is significantly, even necessarily, improved as a result of the expansion, despite having fewer queries. When finished, we hope our test collection will be a generally useful IR resource. Indeed, it has already solicited several enquiries as to its availability. 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 our reviewers for many useful comments. The first author gratefully acknowledges the support of Microsoft Research through the European PhD Scholarship Programme.

References

Shannon Bradshaw. (2003). Reference directed indexing: Redeeming relevance for subject search in citation indexes. In *Proceedings of the European Conference on 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*, 30:107–117.

Chris Buckley and Ellen M. Voorhees. (2004). Retrieval evaluation with incomplete information. In *Proceedings of Research and Development in Information Retrieval (SIGIR)*, pages 25–32.

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. (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 *Proceedings of Research and Development in Information Retrieval (SIGIR)*, pages 272–279.

Mark D. Dunlop and C. J. van Rijsbergen. (1993). Hypermedia and free text retrieval. *Information Processing and Management*, 29(3):287–298.

Donna K. Harman. (2005). The TREC test collections. In Ellen M. Voorhees and Donna K. 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 M. Voorhees and Donna K. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, chapter 9. MIT Press.

William Hersh and Ravi Teja Bhupatiraju. (2003). Trec genomics track overview. In *Proceedings of the Text REtrieval Conference (TREC)*, pages 14–23.

William Hersh, Aaron M. Cohen, Phoebe Roberts, and Hari Krishna Rekapilli. (2006). Trec 2006 genomics track overview. In *Proceedings of the Text REtrieval Conference (TREC)*.

Michael Kluck. (2003). The GIRT data in the evaluation of CLIR systems - from 1997 until 2003. In *Proceedings of Cross-Language Evaluation Forum (CLEF)* , pages 376–390.

Oliver McBryan. (1994). GENVL and WWWW: Tools for taming the web. In *Proceedings of the World Wide Web Conference (WWW)*.

Anna Ritchie, Simone Teufel, and Stephen Robertson. (2006). Creating a test collection for citation-based IR experiments. In *Proceedings of Human Language Technology conference and the North American Chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL)*.

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

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

Ellen Voorhees. (2005). Overview of the TREC 2005 robust retrieval track. In *Proceedings of the Text REtrieval Conference (TREC)*.