Using Terms from Citations for IR: Some First Results

Anna Ritchie[1], Simone Teufel[1], and Stephen Robertson[2]

[1] University of Cambridge, Computer Laboratory, 15 J J Thompson Avenue, Cambridge, CB3 0FD, U.K. ar283,sht25@cl.cam.ac.uk
[2] Microsoft Research Ltd, Roger Needham House, 7 J J Thomson Avenue, Cambridge, CB3 0FB, U.K. ser@microsoft.com

Abstract. We present the results of experiments using terms from citations for scientific literature search. To index a given document, we use terms used by citing documents to describe that document, in combination with terms from the document itself. We find that the combination of terms gives better retrieval performance than standard indexing of the document terms alone and present a brief analysis of our results. This paper marks the first experimental results from a new test collection of scientific papers, created by us in order to study citation-based methods for IR.

1 Introduction

There has been a recent resurgence of interest in using citations between documents. However, while the potential usefulness of the text used in association with citations has been noted in relation to, e.g., text summarization [1, 2], thesaurus construction [3] and other tasks, recent work in IR has focused on statistical citation data, like citation counts and PageRank-style methods, e.g., [4–6]. We test whether term-based IR on scientific papers can be improved with citation information, by using terms from the citing document to additionally describe (i.e., index) the cited document. This 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, including anchor text, has been used to advantage in retrieval tasks [7, 8]. In this comparable situation, Web pages are often poorly self-descriptive [9], while anchor text is often a higher-level description of the pointed-to page [10].

We explore whether using terms from citations to a paper in combination with terms from the paper itself can improve the retrieval performance achieved when only the paper terms are indexed. Some work has been done in this area but no previous experiments have used both citing and cited papers. Previous experiments have indexed cited papers using terms from citing papers but no terms from the cited papers themselves: Bradshaw used terms from a fixed window around citations [11], while Dunlop and van Rijsbergen used the abstracts of citing papers [12].

hyperlink: The Google search engine...

citation: "Dictionaries can be constructed in various ways - see *Watson (1993a, 1995)* for a taxonomy of (general) finite-state automata construction algorithms."

Fig. 1. Similarity between Hyperlinks and Citations

In this paper, we first motivate our use of citations for term-based IR. Then, Section 3 describes our experimental setup; in Section 4, we present and analyse our results, which show that using citation terms can indeed improve retrieval; Section 5 concludes and outlines future work.

2 Motivation and Related Work

There are definite parallels between the Web and scientific literature: "hyperlinks...provide semantic linkages between objects, much in the same manner that citations link documents to other related documents" [13]. However, there are also fundamental differences. An important and widespread factor in the use of hyperlinks is the additional use of their anchor text (i.e., the text enclosed in the ⟨a⟩ tags of the HTML document (see Fig. 1)). It is a well-documented problem that Web pages are often poorly self-descriptive [9]. Anchor text, on the other hand, is often a higher-level description of the pointed-to page. Davison discusses just how well anchor text does this and provides experimental results to back this claim [10]. Thus, beginning with McBryan [7], there is a trend of propagating anchor text along its hyperlink to associate it with the linked page, as well as that in which it is found (as in Fig. 2). Google, for example, includes anchor text as index terms for the linked page [9].

Returning to the analogy between the Web and scientific literature, the anchor text phenomenon is also observed with citations: citations are usually introduced purposefully alongside some descriptive reference to the cited document (see Fig. 1). However, no anchor text exists in scientific papers, unlike in Web pages, where there are HTML tags to delimit the text associated with a link. The question is raised, therefore, of what is the anchor text equivalent for formal citations. Bradshaw calls the concept *referential text*, using it as the basis of his *Reference-Directed Indexing* (RDI), whereby a scientific document is indexed by the text that refers to it in documents that cite it [11], instead of by the text in the document itself, as is typical in IR. The theory behind RDI is that, when citing, authors describe a document in terms similar to a searcher's query for the information it contains. Thus, this referential text should contain good index terms for the document and Bradshaw shows an increase in retrieval precision over a standard vector-space model implementation; 1.66 more relevant documents are retrieved in the top ten in a small evaluation on 32 queries.

However, a number of issues may be raised with RDI. Firstly, it only indexes referential text so a document must be cited at least once (by a document available to the indexer) in order to be indexed. Bradshaw's evaluation excluded any documents that were not cited and does not disclose how many of these there were. Secondly, referential text is extracted using CiteSeer's *citation context* (a window of around one hundred words around the citation). This method is simplistic: the terms that are definitely associated with a citation are variable in number and in distance from the citation, so a fixed window will not accurately capture the citation terms for all citations. In a much earlier study, O'Connor noted the inherent difficulty in identifying which terms belong with a citation [14] and Bradshaw too states the difficulty in extracting good index terms automatically from a citation.

Dunlop investigated a similar technique with a different application in mind (i.e., retrieval of non-textual documents, such as image, sound and video files [12]). Dunlop's retrieval model uses clustering techniques to create a description of a non-textual document from terms in textual documents with links to that document. In order to establish how well descriptions made using the model represent documents, the method was applied to textual documents, indeed, to the CACM test collection, where the links between documents were citations. The experiment compared retrieval performance using the cluster-based descriptions against using the documents themselves; the cluster-based descriptions achieved roughly 70% of the performance from using the document content. Again, however, Dunlop did not measure the performance using the cluster-based descriptions in combination with the document content.

Thus, there is a gap in the research: retrieval performance from using the combination of terms from citing and cited documents has not been measured. How will this compare to using the document terms alone? How will it compare to using terms from the citing documents alone? We could make these comparisons on the CACM collection of abstracts. However, a test collection with the full text of a substantial number of citing and cited papers will allow broader experimentation, e.g., comparisons between using the full cited paper versus the abstract alone. We previously introduced such a test collection [15, 16], since no existing test collection satisfies our requirements. The newswire articles in traditional collections do not contain citations. CACM contains abstracts and the GIRT collection [17], likewise, consists of content-bearing fields, not full documents. The earlier TREC Genomics collections consist of MEDLINE records, containing abstracts but not full papers [18, 19]. In the 2006 track, a new collection of full-text documents was introduced but this was designed for passage retrieval for a QA task, not document retrieval [20]. Our test collection allows us to conduct not only the experiments we report here, but a wider range of ex-

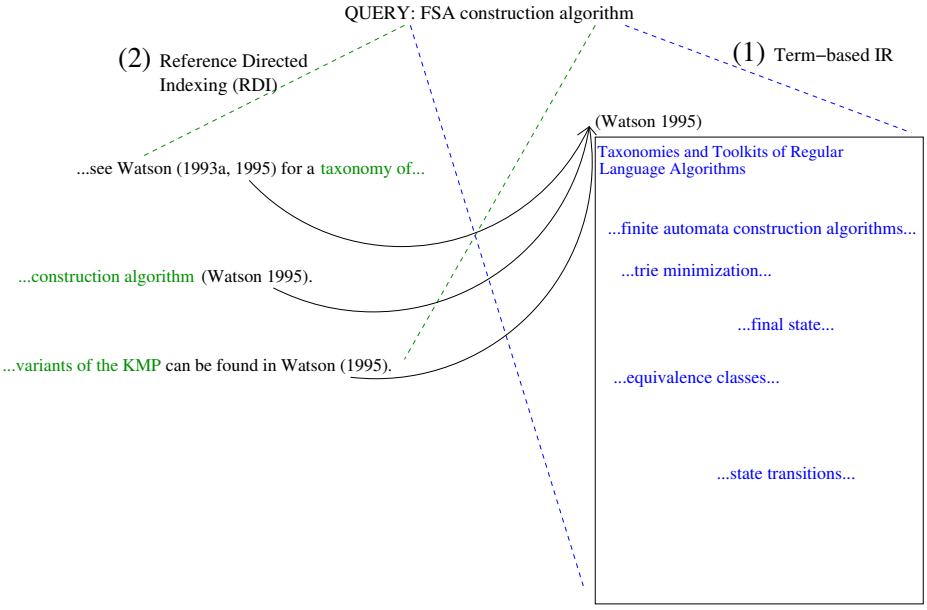

Fig. 2. Use of citation terms in IR: how does (1)+(2) compare to just (1) or (2) alone?

periments using different combinations of information from citing and/or cited documents.

3 Experimental Setup

3.1 Data and Tools

Our test collection is centred around the ACL Anthology[3] digital archive of Computational Linguistics (CL) research papers. The document collection is a ∼9800 document subset of the archive; roughly, all documents published in 2005 or earlier, with non-papers (e.g., letters to the editor, tables of contents) removed. Our query set consists of 82 research questions from CL papers, with an average of 11.4 judged relevant documents per query [16], such as:

– *Does anaphora resolution improve summarization (based on latent semantic analysis) performance?*
– *Can knowledge-lean methods be used to discourse chunk a sentence?*

The test collection was built using a methodology based on the Cranfield 2 design [21]. The principle behind the method is that research papers are written in response to research questions, i.e, information needs, and that the references in a paper are a list of documents that are relevant to that need. Thus, papers

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

are a source of queries (the research questions) and relevant documents (the references). For our queries, the authors of accepted papers for two upcoming CL conferences were asked, by email, for the research question or questions underlying their papers. By asking for queries from recent conference authors, we aimed for a query set that is a realistic model of searches that representative users of the document collection would make; genuine information needs from many different people with many different research interests from across the CL domain. They were also asked to make relevance judgements for their references. Due to the relative self-containedness of the CL domain[4], we expected a significant proportion of the relevance judgements gathered in this way to be for documents in the ACL Anthology and, thus, useful as test collection data.

Based on some analytical experiments [15], there were too few relevance judgements at this stage; we executed a second stage to obtain more judgements for our queries. The Anthology is too large to make complete judgements feasible. Therefore, we used the pooling method to identify potentially relevant documents in the Anthology for each of our queries, for the query authors to judge. First, for each query, we manually searched the Anthology using its Google search facility. We then ran the queries through three standard retrieval models, as implemented in the Lemur Toolkit[5]: Okapi BM25, KL-divergence (both with relevance feedback using the existing relevant documents) and Cosine similarity. We pooled the results from the manual and automatic searches, including all manual search results and adding non-duplicates from each of the automatic rankings in turn until fifteen documents were in the list. Our pool was very shallow compared to TREC-style pools; our method relies on volunteer judges and therefore we needed to keep the effort asked of each judge to a minimum. The list of potentially relevant documents was sent to the query author, with an invitation to judge them and materials to aid the relevance decision.

For the experiments in this paper, we index our documents using Lemur, specifically Indri [22], its integrated language-model based component, using stopping and stemming. Our queries are likewise stopped and stemmed. We use the Cosine, Okapi, KL-divergence and Indri retrieval models with standard parameters to test our method. We also performed retrieval runs using KL[6] with relevance feedback (KL FB). In each run, 100 documents were retrieved per query; this is already far greater than the number of relevant documents for any query. For evaluation, we use the TREC evaluation software, trec_eval[7].

[4] We empirically measured the proportion of collection-internal references in Anthology papers to be 0.33.

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

[6] We do not report results using Okapi with relevance feedback: the Lemur documentation notes a suspected bug in the Okapi feedback implementation.

[7] http://trec.nist.gov/trec_eval/trec_eval.8.1.tar.gz

3.2 Citation Method

We firstly carry out some pre-processing of the documents: we use methods based on regular expressions to annotate the reference list, to identify citations in the running text and to associate these with items in the reference list. This is a non-trivial task, for which high precision methods have been developed independently [23]. Our approach is more simplistic but nevertheless performs well: from a small study of ten journal papers, we found and correctly matched 388 out of 461 citations with their corresponding reference (84.2%). Errors mostly occur due to noise from the PDF to XML conversion.

Next, we use the reference information to identify which references are to documents in the ACL Anthology; we extract terms from the citations associated with these references to a database. Specifically, we use the tokeniser from a statistical natural language parser [24] to segment the text into sentences, then extract all terms from the sentence that contains the citation. Identifying which terms are associated with a citation is an interesting problem, which we have discussed elsewhere [25]; our method is only one of many possibilities and we do not claim that it is optimal. Our database contains terms from over 23,000 citations to over 3300 papers. Finally, we add these terms to an XML representation of the document before indexing. We build one index from the XML documents alone and another from the documents plus citation terms. In order to investigate the effect of weighting citation terms differently relative to document terms, we build separate indexes where the citation terms are added in duplicate to the XML document, to achieve the desired weight. The method is resource-hungry, however, and we investigate only a small range of weights in this way.

There are alternatives to this weighting method: the Indri query language allows terms to be weighted according to which part of the document they occur in. However, this method can only be used with the Indri retrieval model; we cannot use weighted queries to investigate the effects of citation term weighting on all models' performance. Furthermore, the two methods are not equivalent, in terms of document scoring and ranking, for multiple reasons. Firstly, the weighted query method calculates scores using term statistics across individual fields, rather than across whole documents, as in the case of unweighted queries. Thus, the ranking produced by a weighted query where the fields are weighted equally and the ranking produced by its unweighted counterpart on the same index will not necessarily be the same. Secondly, in the term duplication method, the statistics for a given term will be different in each index, as it is altered by the citation 'weight': there will be an additional occurrence of that term in the index for every duplicate citation term that is added. This is not the case in the weighted query method, where each citation term is added exactly

Retrieval Model (W)	MAP			P(5)			R-precision			GMAP		bpref		
	W/out	With	p	W/out	With	p	W/out	With	p	W/out	With	W/out	With	p
Okapi (1)	.083	.084	.582	.110	.120	.251	.094	.098	.313	.004	.004	.218	.227	.118
(2)		.084	.786		.127	.070		.103	.133		.005		**.234**	**.018**
(3)		.085	.636		.127	.070		.108	.053		.005		**.234**	**.016**
(4)		.084	.794		**.129**	**.045**		.104	.228		.004		.230	.130
Cosine (1)	.140	.143	.454	.185	.188	.567	.141	.146	.223	.041	.046	**.313**	**.328**	**.001**
(2)		.146	.148		.190	.418		**.156**	**.002**		.048		**.333**	**.001**
(3)		.143	.528		.185	1.000		**.156**	**.028**		.048		**.335**	**.001**
(4)		.146	.326		.190	.596		**.155**	**.044**		.049		**.338**	**.001**
Indri (1)	.158	**.172**	**.000**	.254	**.285**	**.001**	.188	**.199**	**.025**	.056	.072	.366	**.379**	**.014**
(2)		**.176**	**.000**		**.298**	**.000**		**.210**	**.000**		.077		**.383**	**.005**
(3)		**.180**	**.000**		**.305**	**.000**		**.213**	**.000**		.080		**.387**	**.006**
(4)		**.182**	**.000**		**.302**	**.000**		**.220**	**.000**		.082		**.385**	**.019**
KL (1)	.166	**.174**	**.026**	.256	**.273**	**.019**	.192	**.206**	**.028**	.065	.074	.373	.379	.420
(2)		**.180**	**.003**		.271	.159		**.213**	**.003**		.077		.387	.095
(3)		**.183**	**.001**		.278	.072		**.215**	**.006**		.080		.389	.059
(4)		**.184**	**.004**		.283	.070		**.216**	**.006**		.082		**.393**	**.028**
KL FB (1)	.238	**.250**	**.004**	.332	.349	.090	.251	**.264**	**.015**	.157	.177	.483	.493	.199
(2)		**.259**	**.000**		.346	.259		**.268**	**.009**		.189		**.504**	**.020**
(3)		**.263**	**.000**		.349	.195		**.279**	**.000**		.195		**.511**	**.008**
(4)		**.267**	**.000**		.354	.095		**.282**	**.000**		.199		**.515**	**.006**

Table 1. Retrieval Performance With versus Without Citations (W = weight of citation terms. Differences in **bold** are significant for $p \leq 0.05$ and those **underlined** for $p \leq 0.01$)

once to the index. The differences between these weighting methods opens the door for comparative experimentation between them; we intend to investigate this in the future.

4 Results and Analysis

Table 1 summarizes the results. In each row, we compare the performance of a given retrieval model on the index without citation terms to its performance on one index with citation terms (i.e., with a particular citation term weight). We consider the values of a range of standard performance measures and t-test for statistical significance of with- versus without-citation performance; differences highlighted in bold are significant for $p \leq 0.05$.

Performance is uniformly higher with citations than without, for all models, for all measures, with the exception of two Okapi runs where GMAP is unchanged. The general trend is for performance to increase as citation terms are weighted more highly. Notably, the performance increases on all Indri runs for all measures are statistically significant. All MAP and R-precision increases are significant for Indri, KL and KL FB. Cosine and Okapi show the smallest and least significant performance increases. The results for Okapi, in particular, do not appear to follow the trend of increasing performance with increasing

citation term weight. One possible explanation for this is that the weights investigated here are too small to have any significant effect on Okapi's performance; in the comparable situation of Web IR, optimal Okapi performance has been achieved by weighting anchor text as much as 35 times higher than Web page body text [26]. This may also be the case for the Cosine model. Similarly, the narrow range of citation weights does not show a plateau in the performance of any of the other models. Further investigation is required to discover the optimal weighting of citation terms to document terms.

To try to better understand the overall results, we studied in detail the retrieval rankings for a few queries and observed the effects of adding citation terms at the individual document level. We selected queries with the most marked and/or anomalous performance changes. Judged relevant and irrelevant documents are denoted by R and I, respectively; ? denotes an unjudged document.

Query #34 *Given perspective, opinion, and private state words, can a computer infer the hierarchy among the different perspectives?*
{`perspective opinion private state word compute infer hierarchy perspective`}

This query exhibits a drop in Okapi performance while the majority of the models' performance increases or stays the same. Okapi's MAP drops from 0.7037 to 0.5913, while bpref, R-precision and P(5) are all unchanged. The pertinent ranking changes are summarised as follows:

Doc ID	Rel	Rank	Cits	Query Terms in Doc+Cits
W03-04_LONG	R	1→1	4	opinion 18+1, private 5+0, perspective 17+0, compute 1+0
C04-1018	R	2→3	4	opinion 23+0, private 86+0, perspective 4+0, state 93+0, word 5+0, compute 1+0
P99-1017	I	3→4	4	private 8+0, perspective 2+0, infer 1+0, state 1+0, compute 1+0, hierarchy 1+0
W05-0308	?	4→2	6	opinion 15+2, private 72+0, perspective 1+1, infer 1+0, state 79+0, word 10+0, compute 2+0
W03-0404	?	5→5	8	opinion 8+3, private 10+0, perspective 2+0, state 12+0, word 48+0, compute 3+0, hierarchy 1+0
C90-2069	R	27→28	0	private 50+0, perspective 1+0, state 51+0, compute 4+0

The query has only three judged relevant documents in total (versus 16 judged irrelevant), all of which are retrieved both with and without citation terms. The relevant documents are retrieved at ranks 1, 2 and 27 without citations and at ranks 1,3 and 28 with citations, respectively. This accounts for the drop in MAP. The new document at rank 2 is an unjudged document with six citations added to it, resulting in two additional occurrences of the query term *opinion* and one of *perspective* in the with-citations index. This overtakes the relevant document previously at rank 2, which gains only one *opinion* from its four citations. Because the document is unjudged, bpref is not affected. Its

title is *'Annotating Attributions and Private States'*, suggesting it might indeed be relevant to the query, in which case MAP would increase (from 0.725 to 0.786) not decrease. The relevant document at rank 1 has no citations in the database and, thus, no new terms in the with-citations index. However, it has a high occurrence of the query terms *opinion*, *private* and *state*, as well as some occurrences of other query terms, and retains a high enough score to remain at rank 1. This document would not be retrieved if only citation terms (and not document terms) were used for indexing. The one judged irrelevant document in the top 5 is moved from rank 3 to 4 since its citations add no query terms.

Therefore, it appears that the citations are, in fact, contributing useful terms and helping relevant documents be retrieved. In the case of Okapi, however, the positive effects are not shown by the evaluation measures, for several reasons: a) the movements in the rankings are small because there are few citations for the judged documents, b) there are very few judged relevant documents for the query and c) some of the (potentially positive) effect of the citation terms affects unjudged documents, which affects MAP negatively. Given the incompleteness of the judgements, it is likely that measures which do not assume completeness, such as bpref, will be more reliable indicators of retrieval performance.

Query #15 *How can we handle the problem of automatic identification of sources of opinions?*

{`handle problem automatic identify source opinion`}

This query shows one of the largest increases in performance: the values of all measures increase (or stay the same) for all models. Considering Indri in detail, MAP increases from 0.4492 to 0.5361, R-precision from 0.5263 to 0.5789 and bpref from 0.6645 to 0.8224. The number of relevant documents retrieved increases from thirteen to sixteen, out of a possible nineteen in total. These newly retrieved documents largely account for the overall performance increases, summarised as follows:

Doc ID	Rel	Rank	Cits	Query Terms in Doc+Cits
H05-1044	R	_→74	7	opinion 12+7, identify 12+4, source 0+3, automatic 4+2
P02-1053	R	_→81	40	opinion 6+5, handle 1+0, identify 1+6, source 1+0, automatic 2+7, problem 2+1
W02-1011	R	_→87	31	opinion 7+1, identify 1+43, source 3+0, automatic 7+3, problem 14+2

Now considering Okapi, MAP increases from 0.0999 to 0.1028, bpref from 0.4934 to 0.5263 and eleven relevant documents are retrieved with citations, versus ten without. The remaining measures remain unchanged. Again, the overall performance increase is mainly due to the newly retrieved document:

Doc ID	Rel	Rank	Cits	Query Terms in Doc+Cits
W03-0404	R	_→93	8	opinion 8+3, identify 15+2, source 2+0, automatic 12+2, problem 2+0

Thus, Okapi's performance on this query does improve, following the trend of the other models. However, its increase is somewhat smaller than that of the other models. If this is generally the case, for queries with less marked performance increases than this example query, in addition to queries such as #34 where Okapi's performance drops slightly, then it is unsurprising that the overall measured differences in performance with and without citations are statistically insignificant.

Query #46 *What is the state-of-the-art on semantic role labeling using real syntax?*
{`state art semantic role label real syntax`}

This query shows a general increase in performance across measures and models, with the exception that bpref drops for KL, KL FB and Cosine and stays the same for Okapi and Indri. We consider the KL rankings in detail, where the decrease is the most marked (0.6207 to 0.1034). This is accounted for by the fact that the one judged irrelevant document for this query is retrieved at rank 10 with citations, whereas it was previously unranked without citations, overtaking judged relevant documents:

Doc ID	Rel	Rank	Cits	Query Terms in Doc+Cits
C04-1100	I	_ →10	9	state 0+1, semantic 0+7, role 0+3, label 0+1

Similarly, this document was retrieved at rank 64 by Cosine, where it was previously unranked. Neither Okapi nor Indri retrieved the document, with or without citations, so their bpref values do not change.

This is an example where the citation terms, again, have a definite effect on the retrieved documents but this time result in an irrelevant document, as well as relevant documents, being ranked higher. The citation terms that cause this do, indeed, match the query terms. However, closer inspection of the citing sentences that the terms were taken from reveals that they do not match the particular sense of the terms in the query, e.g., one of the occurrences of the term *role* comes from the phrase *to explore the role of semantic structures in question answering*. Indeed, the document is titled *'Question Answering Based on Semantic Structures'* and is not about semantic role labeling, the topic of the query. This is an inherent danger in term-based IR and not a product of our citation method: such semantic mismatches can occur with document terms as well as citation terms.

5 Conclusions and Future Work

We have conducted some first experiments using the combination of terms from citations to a document and terms from the document itself to index a given document. Performance, as gauged by a range of standard evaluation measures,

generally increases when citation terms are used in addition to document terms. Furthermore, performance generally increases as citations are weighted higher relative to document terms. The Okapi and Cosine retrieval models, as implemented in the Lemur toolkit, do not appear to follow this trend. It may be a characteristic of these models, however, that the citation terms need to be weighted much higher relative to document terms than we have investigated here: for Okapi, weights of up to 35 have been found to be optimal for weighting anchor text relative to Web page body text, in the comparable situation of Web IR. Likewise, our results do not allow us to surmise an optimal citation term weight for the Indri and KL retrieval models. We intend to investigate a wider range of weights in future work. We also intend to investigate alternative methods of weighting the citation terms.

These are the first reported experimental results from a new test collection of scientific papers, created by us in order to more fully investigate citation-based methods for IR. The relevance judgements are known to be incomplete; we have noted the definite effects of this incompleteness on perceived retrieval performance, according to measures such as MAP. It is likely that measures which do not assume complete judgements, such as bpref, will be more reliable indicators of retrieval performance. Each of the retrieval models investigated here shows a statistically significant increase in bpref, when citation terms are added, for some citation term weight.

In conclusion, our experiments indicate that indexing citation terms in addition to document terms improves retrieval performance on scientific papers, compared to indexing the document terms alone. It remains to be seen what the optimal weighting of citation terms relative to document terms is and what the best way of implementing this weighting might be. It will also be interesting to investigate alternative methods of extracting citation terms, e.g., how using citing sentences compares to using fixed windows or more linguistically motivated techniques. Finally, we intend to compare against the performance when only citation terms are indexed. This method, however, restricts the documents that can be retrieved to those with at least one citation; we anticipate that using both document and citation terms will be most effective.

Acknowledgements The first author gratefully acknowledges the support of Microsoft Research through the European PhD Scholarship Programme.

References

1. Teufel, S., Siddharthan, A., Tidhar, D.: Automatic classification of citation function. In: Proceedings of Empirical Methods in Natural Language Processing. (2006) 103–110
2. Schwartz, A.S., Hearst, M.: Summarizing key concepts using citation sentences. In: Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology. (2006) 134–135
3. Schneider, J.: Verification of bibliometric methods' applicability for thesaurus construction. PhD thesis, Royal School of Library and Information Science (2004)

4. Strohman, T., Croft, W.B., Jensen, D.: Recommending citations for academic papers. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR). (2007) 705–706
5. Fujii, A.: Enhancing patent retrieval by citation analysis. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR). (2007) 793–794
6. Meij, E., de Rijke, M.: Using prior information derived from citations in literature search. In: Proceedings of the International Conference on Recherche d'Information Assiste par Ordinateur (RIAO). (2007)
7. McBryan, O.: GENVL and WWWW: Tools for taming the web. In: Proceedings of the World Wide Web Conference (WWW). (1994)
8. Hawking, D., Craswell, N.: The very large collection and web tracks. In Voorhees, E.M., Harman, D.K., eds.: TREC: Experiment and Evaluation in Information Retrieval. MIT Press (2005)
9. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems **30** (1998) 107–117
10. Davison, B.D.: Topical locality in the web. In: Proceedings of Research and Development in Information Retrieval (SIGIR). (2000) 272–279
11. Bradshaw, S.: 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). (2003) 499–510
12. Dunlop, M.D., van Rijsbergen, C.J.: Hypermedia and free text retrieval. Information Processing and Management **29**(3) (1993) 287–298
13. Pitkow, J., Pirolli, P.: Life, death, and lawfulness on the electronic frontier. In: Proceedings of the Conference on Human Factors in Computing Systems. (1997)
14. O'Connor, J.: Citing statements: Computer recognition and use to improve retrieval. Information Processing and Management **18**(3) (1982) 125–131
15. Ritchie, A., Teufel, S., Robertson, S.: 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 (HLT-NAACL). (2006)
16. Ritchie, A., Robertson, S., Teufel, S.: Creating a test collection: Relevance judgements of cited & non-cited papers. In: Proceedings of the International Conference on Recherche d'Information Assiste par Ordinateur (RIAO). (2007)
17. Kluck, M.: The GIRT data in the evaluation of CLIR systems - from 1997 until 2003. In: Proceedings of Cross-Language Evaluation Forum (CLEF) . (2003) 376–390
18. Hersh, W., Bhupatiraju, R.T.: Trec genomics track overview. In: Proceedings of the Text REtrieval Conference (TREC). (2003) 14–23
19. Hersh, W., Bhupatiraju, R.T., Ross, L., Johnson, P., Cohen, A.M., Kraemer, D.F.: Trec 2004 genomics track overview. In: Proceedings of the Text REtrieval Conference (TREC). (2004)
20. Hersh, W., Cohen, A.M., Roberts, P., Rekapilli, H.K.: Trec 2006 genomics track overview. In: Proceedings of the Text REtrieval Conference (TREC). (2006)
21. Cleverdon, C., Mills, J., Keen, M.: Factors determining the performance of indexing sytems, volume 1. design. Technical report, ASLIB Cranfield Project (1966)
22. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: a language-model based search engine for complex queries. Technical report, University of Massachusetts (2005)
23. Powley, B., Dale, R.: Evidence-based information extraction for high accuracy citation and author name identification. In: Proceedings of the International Conference on Recherche d'Information Assiste par Ordinateur (RIAO). (2007)
24. Briscoe, E., Carroll, J.: Robust accurate statistical annotation of general text. In: Proceedings of the Conference on Language Resources and Evaluation (LREC). (2002) 1499–1504
25. Ritchie, A., Teufel, S., Robertson, S.: How to find better index terms through citations. In: Proceedings of COLING/ACL Workshop on How Can Computational Linguistics Improve Information Retrieval? (2006)
26. Zaragoza, H., Craswell, N., Taylor, M., Saria, S., Robertson, S.: Microsoft Cambridge at TREC-13: Web and HARD tracks. In: Proceedings of the Thirteenth Text REtrieval Conference (TREC). (2004)