Exploiting Citation Contexts for Physics Retrieval

Anna Dabrowska and Birger Larsen[1]

Aalborg University, Copenhagen, Denmark

Abstract. The text surrounding citations within scientific papers may contain terms that usefully describe cited documents and can benefit retrieval. We present a preliminary study that investigates appending citation contexts from citing documents to cited documents in the iSearch test collection. We examine the effect on information retrieval performance of a range of citation context sizes and their variable weighting. We find that relatively short citation contexts with moderate weights can improve retrieval performance, and demonstrate the feasibility of identifying citation contexts in a large collection of physics papers, paving the way for future research that exploits citation contexts for retrieval.

1 Introduction and Related Work

Bibliographic citations have long aided in the retrieval of scientific information [1]. Commercial citation indexes such as Web of Science and Scopus rely on bibliographic references, without providing the full text of documents. As research publications are increasingly available in full text – through institutional repositories and Open Access policies – new possibilities arise for improving information access by analysing citations within text.

In this work, we investigate the feasibility of extracting citation contexts from citing articles and using them in the retrieval of scientific documents. Specifically, we build on research by Ritchie [2], who worked with a collection of approximately 9,800 papers from the ACL Anthology and investigated the use of citation contexts extracted from citing papers as descriptions of cited papers. She found that appending citation contexts to cited document text could significantly improve retrieval [2]. We conduct similar experiments, with a more principled weighting scheme for contexts, and use a test collection from another academic discipline. We use the iSearch collection, a publicly available test collection with over 400,000 physics documents [3], which is much larger than collections used in previous information retrieval (IR) research with citation contexts [2, 4, 5]. Since citation patterns and norms also vary according to academic discipline, using a collection of physics papers may provide new insights [6]. We supplement the original document collection in iSearch, which consists largely of PDF files, with a set of full-text TEX source files recently acquired from arXiv.org, and lay the groundwork for more in-depth experiments with citation data.

This preliminary study has two main research objectives: (1) Determine whether it is feasible to extract citation contexts from TEX source files for IR purposes. (2) Introduce citation contexts into retrieval and assess (a) how citation contexts of different fixed window sizes impact performance, and (b) the effect of altering the weight assigned to citation contexts, seeking the optimal weight.

2 Tools and Methods

2.1 Data

The iSearch test collection is a document collection containing physics publications, and 65 search tasks (called 'topics') with relevance assessments [3]. Since iSearch is designed to support experiments with integrated search, the collection contains different document types. We use a subset of 434,813 documents in two types: (1) ~160,000 documents with full text extracted from PDFs and metadata harvested from arXiv.org; (2) ~275,000 documents with metadata and abstracts, but without full text. These documents were previously extracted from arXiv.org and include metadata wrapped in XML fields. We also use 64 of the topics, which contain information-need descriptions developed by physics lecturers, PhDs, and MSc students. For each topic, document relevance assessments are made on a 4-point scale (0–3) for up to 200 documents [3].

Additionally, we use the following data: (1) Direct citation data describing the citation network within the collection. Citations were extracted automatically as part of CiteBase [7]. (2) TEX source files that we downloaded from arXiv.org[1]. The source files allow us to exploit TEX typesetting commands, and more easily identify references and citations. Note that the content of these new files may differ slightly from the original files since authors could have updated their work following the original iSearch extraction in 2009. (3) An id concordance file matching iSearch document identifiers (e.g. PF417005) to identifiers from arXiv.org (e.g. astro-ph.9903338). This allows us to locate specific iSearch documents in the arXiv TEX files.

We find that a total of 259,093 unique documents are cited 3.7 million times within the iSearch collection. This indicates a high level of intra-collection citation and provides a much larger dataset than previous research. Combining the direct citation data and document relevance assessments, we find that 3,833 assessed documents (ranked 0–3) are cited at least once in the collection, and 863 of these are deemed relevant to at least one topic (ranked 1–3).

2.2 Context Extraction

For this initial study, we limit ourselves to a scenario similar to re-ranking. We seek to improve retrieval scores by appending citation contexts to a portion of cited documents in the iSearch collection. We choose to add citation contexts to the top 1,000 documents retrieved by a set of preliminary retrieval runs. We

[1] arXiv.org has made source files from the entire preprint archive available for bulk download from Amazon S3, see: http://arxiv.org/help/bulk_data_s3

conduct these preliminary runs on the original iSearch files (XML), as described in Sect. 3. Then, we extract citation contexts from documents in the collection that cite the retrieved documents, and append the contexts to the retrieved documents. This allows us to experiment with using citation contexts in further retrieval runs. Since we have TEX files corresponding to iSearch documents, we can attempt to extract contexts from every citing document. Otherwise, we would be limited to the 160,000 full text documents in the original iSearch files. TEX files also allow us to take advantage of typesetting commands in order to both identify reference items pointing to cited documents and locate in-text citations. Standard reference items are formatted to include a marker in the command: `\bibitem{marker}`. The marker also is used in the `\cite{marker}` command to automatically cite reference items within the text.

The TEX files from arXiv.org are located in separate directories, each one corresponding to a particular iSearch document. Some documents are split into multiple TEX files. To simplify extraction, we use directories containing only a single TEX file. Two methods are employed for locating reference items within TEX and BBL files[2]. The first method uses arXiv ids, and the second uses cluster keys. Since arXiv ids are unique identifiers often occurring in references, we search all citing document reference lists for items with arXiv ids matching documents from the preliminary retrieval, i.e. the cited documents. The second method uses regular expression cluster keys to identify references to the cited documents. A cluster key is an abbreviated series of letters and numbers composed from a document's metadata, and used to identify references to the document [8]. Due to variability in citation styles and formatting, we construct three cluster keys with varying degrees of leniency to identify each cited document. In order to find in-text citations we rely on the convenience of the `\cite{marker}` command. We search for in-text citations to the preliminary documents and extract windows of up to 100 words before and 100 words after the citation marker. We narrow our scope to citations mentioning a single reference item, and exclude instances where multiple citations are made in a single block. This is done to increase the likelihood that extracted contexts pertain to the cited document.

We attach citation contexts to the cited documents from our preliminary retrieval with fixed windows of up to 25, 50, 75 and 100 words on each side of the citations, following Ritchie [2]. The underlying assumption is that terms surrounding a citation are likely to pertain to the cited document, and the further away that words occur from the citation, the less likely they are to be relevant. For each citation context extracted from a citing document, we create the four window sizes and append them within XML fields. In the case of multiple citation contexts from one or more citing documents, the contexts are merged. By appending the citation contexts as new fields to each XML file, we are able to weight the original document text with respect to citation contexts during retrieval.

[2] BBL files are separate BibTex files containing reference items.

2.3 Weighting Citation Contexts

We weight the original documents and the citation fields for retrieval using a linear combination of evidence from the original document text and the citation contexts: Let d be the original document field and c be the citation context field that we append to the document. We weight their respective impact on the final ranking R using simple linear combination (Eq. 1):

$$R = d \times (1 - \alpha) + c \times \alpha \tag{1}$$

where $0 \leq \alpha \leq 1$ is a parameter controlling the effect of d over c. $a < 0.5$ boosts d over c and vice versa.

The linear combination in Eq. 1 is implemented using Indri's `#weight` operator, see Table 1 for an example. We vary α between 0.0 and 1.0 in increments of 0.1, e.g. $\alpha = 0.3$ means that the original document text is assigned a weight of 0.7 and the appended citation context a weight of 0.3.

Table 1. A query syntax example with two fields: orig_doc = the original document text, 25word_citContext = appended 25 word citation context, $\alpha = 0.3$.

Baseline query: `#combine(manipulation nano spheres)`

Example experimental query: `#weight(0.7 #combine(manipulation.(orig_doc) nano.(orig_doc) spheres.(orig_doc)) 0.3 #combine(manipulation.(25word_citContext) nano.(25word_citContext) spheres.(25word_citContext)))`

3 Experiments

3.1 Experimental setup

Retrieval is conducted with the Indri search engine[3]. Indexing is done without stopping and with Krovetz stemming [9]. For ranking, we use Language Modeling with Dirichlet smoothing and tune the μ parameter with values between 0 and 5,000 in increments of 500 following [10]. For each of 64 topics we retrieve 1,000 documents, topic 5 is excluded because it only retrieves 286 documents. We report scores for Mean Average Precision (MAP) and Normalized Discount Cumulative Gain (nDCG). MAP scores are computed by considering documents with a relevance assessment of 1–3 as equally relevant. Tuning is done separately for MAP and nDCG scores, resulting in a separate retrieval for each. In our experiments we measure upper-bound performance, tuning the μ parameter for the highest MAP and nDCG scores respectively.

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

3.2 Findings: Extracted Citation Contexts

Both the nDCG and MAP preliminary runs ranked 1,000 documents over 64 queries, but there was a large overlap in the documents retrieved. We identified 52,586 unique documents from the combination of both result sets. According to the direct citation data, 48% (25,356) of the documents are cited at least once. Documents from the preliminary retrieval provided a useful range of citations, with an average citation count (over cited documents) of 13.2. A few documents are cited very often, the highest citation count is 4,904 to one document.

We appended at least one citation context to a large portion of the documents, 76% (19,248) of cited documents from the preliminary retrieval. We found that 1,577 relevant documents (ranked 1–3) were retrieved, 399 of which had at least one citation context appended.

Using the direct citation data, we identified 134,077 unique documents citing retrieved documents. We successfully found references in 88% (118,357) of citing documents. Since citing documents often cite more than one document, these files accounted for 87% of the direct citations (292,988 out of 335,356). Using the arXiv ids and cluster keys, markers were extracted from citing documents for 75% of the direct citations (249,881). Finally, citation contexts were identified and extracted for 39% (131,285) of the citations, while 118,596 markers were not linked to an in-text citation.

3.3 Findings: Using Citation Contexts in Retrieval

Table 2 and Fig. 2 display the retrieval results. Overall, the retrieval scores were relatively low, but in line with other experiments using iSearch (e.g. [10, 11]). We treat the $\alpha = 0.0$ runs as our baselines, because all weight is given to the original document text in retrieval. The MAP baseline was at 0.0951 and the nDCG baseline was 0.3082. Low baseline scores may be a result of shallow pools of assessed documents for each topic in iSearch.

The citation context runs improved slightly over the baseline scores. Overall, moderate weight given to the citation contexts, α between 0.1 and 0.3, tended to score higher than document text alone. Assigning α weights of 0.5 or above to the contexts led to decreased performance for both MAP and nDCG scores across all window sizes. With regard to window size, scores were improved by the addition of citation contexts in 25, 50, 75, and 100-word windows. The best performance was achieved with 25-word windows, but differences were small. A Student's paired t-test found no statistically significant differences between experiment and baseline MAP and nDCG scores for $p \leq 0.05$.

MAP scores were highest for the 25-word window, it outperformed others with a 4.8% increase over the baseline. Weighting affected scores differently across window sizes. Our results pointed to a higher document weighting performing better in large windows, and higher context weighting performing better with smaller windows. The 25 and 50-word windows show the best MAP performance at α of 0.2, and both the 75 and 100-word windows at α of 0.1. This indicates that larger context sizes may introduce more irrelevant terms.

Table 2. Overview of retrieval results: MAP and nDCG scores with percent difference from baselines. MAP baseline = 0.0951, nDCG baseline = 0.3082.

Words	α	MAP	nDCG
25	0.5	0.0803 (-15.6)	0.2949 (-4.3%)
25	0.4	0.0895 (-5.9)	0.3065 (-0.6%)
25	0.3	0.0989 (4.0)	0.3150 (2.2%)
25	0.2	0.0997 (4.8)	0.3120 (1.2%)
25	0.1	0.0989 (4.0)	0.3110 (0.9%)
25	0.0	0.0951 (0.0)	0.3082 (0.0%)
50	0.5	0.0799 (-16.0)	0.2891 (-6.2%)
50	0.4	0.0863 (-9.3)	0.3005 (-2.5%)
50	0.3	0.0958 (0.7)	0.3104 (0.7%)
50	0.2	0.0994 (4.5)	0.3127 (1.5%)
50	0.1	0.0993 (4.4)	0.3124 (1.4%)
50	0.0	0.0951 (0.0)	0.3082 (0.0%)
75	0.5	0.0797 (-16.2)	0.2883 (-6.5%)
75	0.4	0.0875 (-8.0)	0.3005 (-2.5%)
75	0.3	0.0960 (0.9)	0.3097 (0.5%)
75	0.2	0.0991 (4.2)	0.3115 (1.1%)
75	0.1	0.0995 (4.6)	0.3126 (1.4%)
75	0.0	0.0951 (0.0)	0.3082 (0.0%)
100	0.5	0.0768 (-19.2)	0.2853 (-7.4%)
100	0.4	0.0868 (-8.7)	0.2971 (-3.6%)
100	0.3	0.0952 (0.1)	0.3090 (0.3%)
100	0.2	0.0988 (3.9)	0.3106 (0.8%)
100	0.1	0.0994 (4.5)	0.3117 (1.1%)
100	0.0	0.0951 (0.0)	0.3082 (0.0%)

The highest nDCG scores also resulted from using the two smaller windows. Again, the 25-word window performed best, with a 2.2% increase over the baseline. This increase could be due to more relevant documents appearing closer to the top of the retrieval ranking with smaller citation windows. The nDCG scores exhibited a pattern similar to MAP scores in the effect of weighting. The 25-word window performed best at an α of 0.3, the 50 at 0.2, and both the 75 and 100-word windows at an α of 0.1. This pattern deserves further investigation.

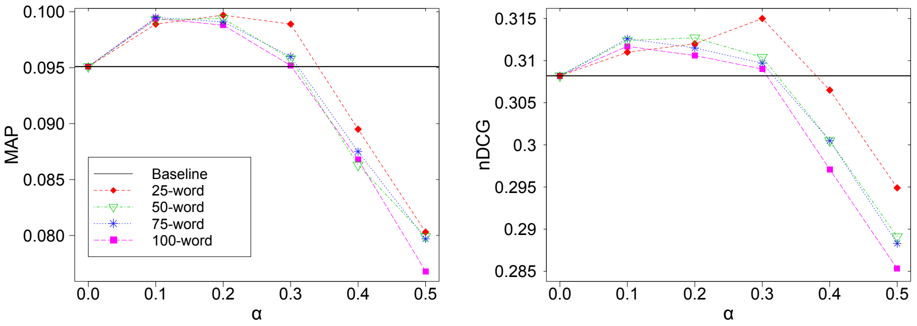

Fig. 1. Plots of MAP and nDCG scores as a function of the α parameter determining weights assigned to citation contexts. Horizontal lines represent the baseline scores.

4 Discussion and Conclusions

Our study takes advantage of the iSearch test collection in a new way. We use the internal citation network and additional arXiv.org files to show that citation contexts can be used with iSearch. We also provide some initial results in weighting citation contexts of different sizes with a language modeling approach to retrieval. The changes in retrieval scores are relatively small, and a reason may be that few relevant document had citation contexts appended (399 out of 1,577 relevant documents retrieved by our baseline). The reranking-like scenario we chose for this initial study appears insufficient to show significant improvement. However, results point to citation contexts positively impacting the retrieval of physics documents.

Information extraction could be improved in several ways to secure better coverage. First, citations were sought only to documents in our preliminary retrieval, and only for documents with a single TEX file. Citation contexts could be appended to all cited documents in the collection, and documents with multiple TEX files included. Second, the regular expression cluster keys used for this experiment may have omitted references to some cited documents due to identification of multiple reference items with the strictest key, mistakes in the references, references that lacked markers for identifying items in text, or references that were otherwise non-standard. Third, there may be errors in the direct citation data from CiteBase. Fourth, when searching for in-text citations, we also narrowed our scope significantly by seeking only citations to individual documents. Many documents are cited within a group. Fifth, the large decrease between finding reference items and identifying in-text citations can also be attributed to non-standard citation styles, or documents appearing within the reference list without being cited in the text[4].

Nonetheless, the scale of our citation extraction for the iSearch collection is larger than in previous work [2, 4, 5]. Our study primarily determines the viability of extracting citations in the manner pursued. With this validation, we are currently working to more thoroughly extract contexts within the whole collection and improve coverage for each cited document. Additionally, future retrieval will be conducted with 3-fold cross-validation in order to reduce potential overfitting caused by tuning for the highest MAP and nDCG scores.

When appending citation contexts to cited documents, we adopted the simple rationale that words describing a cited paper occur close to citations [2]. We are aware that fixed sizes around a citation may not capture all terms relevant to a citation, and are likely to capture irrelevant ones as well. In future work, we will experiment with appending contexts of different sizes before and after citations. We will also consider linguistic features when determining the size of citation contexts to include in retrieval experiments. At a basic level, this includes sentence demarcation. Further considerations can be made for automatically determining more refined context sizes in order to reduce overlap between multiple contexts and potential noise. In this experiment we included unaltered

[4] This happens when authors publish bibliographies, rather than reference lists.

contexts, but the text contains formulae and formatting commands that can be removed with additional processing. Our study points to more promising results with advanced computational techniques and additional cleaning.

The results also indicate that optimal citation context weighting relative to document weighting may vary according to context size. Context weighting is worth further pursuit and additional data can be incorporated into experiments. Perhaps insight can be gained from varying weights with respect to publication and citation dates, or the location of citations within document sections. Tools for automatic sentiment detection and discourse analysis may also inspire graded weighting schemes [12].

An additional contribution of this study is in opening the iSearch collection to a range of future experiments using data from citation links and contexts. Working further with structured files in a controlled setting will provide the initial scaffolding to help determine which techniques are worth pursuing more broadly.

References

1. Garfield, E., 1963. Science Citation Index. *Science Citation Index 1961* (1), v-xvi.
2. Ritchie, A., 2009. Citation context analysis for information retrieval. *University of Cambridge Computer Laboratory Technical Report 744.*
3. Lykke, M., Larsen, B., Lund, H., & Ingwersen, P., 2010. Developing a test collection for the evaluation of integrated search. In *ECIR*, 627–630.
4. O'Connor, J., 1982. Citing statements: Computer recognition and use to improve retrieval. *Information Processing and Management 18*(3), 125–131.
5. Bradshaw, S., 2003. Reference directed indexing: Redeeming relevance for subject search in citation indexes. In *ECDL*, 499–510.
6. Thompson, P., & Tribble, C., 2001. Looking at citations: Using corpora in English for academic purposes. *Language Learning and Technology, 5*(3), 91–105.
7. Brody, T., 2003. Citebase search: Autonomous citation database for e-print archives. In *SINN*, 10 pages.
8. Glanzel, W., 2003. *Bibliometrics as a research field: A course on theory and application of bibliometric indicators.* (Course handouts).
9. Krovetz, R., 1993. Viewing morphology as an inference process. In *SIGIR*, 191–202.
10. Lioma, C., Kothari, A., & Schutze H., 2011. Sense discrimination for physics retrieval. In *SIGIR*, 1101–1102.
11. Zhao, H., & Hu, X., 2014. Language model document priors based on citation and co-citation analysis. In *BIR*, 29–36.
12. Teufel, S., 2014. Scientific argumentation detection as limited-domain intention recognition. In *Workshop for Frontiers and Connections between Argumentation Theory and Natural Language.* To Appear.