# Bibliometric-Enhanced arXiv: A Data Set for Paper-Based and Citation-Based Tasks

Tarek Saier and Michael Färber

Department of Computer Science, University of Freiburg, Germany
tarek.saier@uranus.uni-freiburg.de
michael.faerber@cs.uni-freiburg.de

**Abstract.** In recent years, several research paper-based tasks, such as paper recommendation, and citation-based tasks, such as citation recommendation and citation context-based document summarization, have been proposed. The evaluations of approaches to such tasks and their applicability in real-world scenarios heavily depend on the used data set. However, existing data sets are limited in several regards. In this paper, we propose a new data set based on all publications from all scientific fields available on arXiv.org. Apart from providing the papers' plain text, in-text citations were annotated via global identifiers. As far as possible, cited publications were linked to the Microsoft Academic Graph. Our data set consists of over one million documents and 29.2 million citation contexts. The data set, which is made freely available for research purposes, not only can enhance the future evaluation of research paper-based and citation context-based approaches but also serve as a basis for novel ideas to analyze papers.

**Keywords:** scholarly data, citations, arXiv.org, digital libraries, data set

## 1 Introduction

A variety of tasks exploit scientific paper collections to help researchers in their work. For instance, research paper recommender systems have been developed [1]. Related are systems that operate on a more fine-grained level within the full text, such as the textual contexts in which citations are mentioned (i.e., citation contexts). Based on citation contexts, things like the citation function [2], the citation polarity [3] and the citation importance can be determined. Furthermore, citation contexts are necessary for context-dependent citation recommendation [4,5], as well as for citation-based document summarization tasks, such as citation-based automatic survey generation [6] and automatic related work section generation [7].

The evaluations of approaches developed for all these tasks as well as the actual applicability and usefulness of the developed systems in real-world scenarios heavily depend on the used data set. This data set is typically a collection of papers provided in full text or a set of already extracted citation

contexts, consisting, for instance, of 1-3 sentences each. Existing data sets, however, do not fulfill all of the following criteria (see Sec. 2 for more details):

1. *Size.* The data set can be comparably small (below 100,000 documents) which makes it difficult to use it for training and testing supervised machine learning approaches;
2. *Cleanliness.* The papers' full texts or citation contexts are often very noisy due to the conversion from PDF to plain text and due to encoding issues;
3. *Global citation annotations.* No links from the citations in the text to the structured representations of the cited publications across documents are provided;
4. *Data set interlinkage.* Data sets often do not provide identifiers of the citing and cited documents from other data sets (e.g., DBLP or the Microsoft Academic Graph);
5. *Cross-domain coverage.* Often, only a single scientific discipline is considered.

In this paper, we propose a new data set for paper-based tasks as well as citation-based tasks, based on all publications available on `http://arXiv.org`. It consists of over one million full text documents (about 269 million sentences) and links to 2.7 million unique papers via 29.2 million citation contexts (having 15.9 million unique references).[1] Thus, we argue that it is considerably large, fulfilling item (1). By using the LaTeX source files and by developing a highly accurate transformation method that converts LaTeX to plain text, we can resolve issue (2). Besides the pure papers' content, in-text citations are annotated directly in the text via global identifiers, contributing to aspect (3). As far as possible, cited publications are linked to the Microsoft Academic Graph (MAG)[2] [8] (cf. aspect (4)). This enables us to use the arXiv paper content in combination with the MAG data, which contains metadata of 213 million publications as of February 2019, along with metadata about researchers, venues, and fields of study. Our data set also fulfills constraint (5) as all disciplines covered in arXiv are considered. This enables researchers to analyze papers from several disciplines and to compare approaches across disciplines.

Our data set is freely available at `http://doi.org/10.5281/zenodo.2609187` and the implementation for creating it at `https://github.com/IllDepence/unarXive`. Not only can the data set be used as a new large data set for evaluating paper-based and citation-based approaches but also as a basis for novel ways of paper analytics within bibliometrics and scientometrics. For instance, based on the citation contexts and the cited papers' metadata using the MAG, one can analyze whether there exist biases in the citing behavior of researchers.

The paper is structured as follows: After outlining related data sets in Sec. 2, we describe in Sec. 3 how we created our data set. In Sec. 4, we present an

[1] Note that references are links to cited documents on the document level, while citations are links to cited documents within the text.

[2] See `https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/` and `http://ma-graph.org`.

Table 1: Overview of existing data sets.

| Data set | #Papers | Cit. context | Disciplines | Full text | Ref. IDs |
|---|---|---|---|---|---|
| arXiv CS [9] | 90k | 1 sentence | CS | yes | DBLP |
| CiteSeerX [10]/RefSeer [11] | 1M | 400 chars | (unrestricted) | no | no |
| PubMed Central OAS[3] | 2.3M | extractable | Biomed./Life Sci. | yes | mixed |
| Scholarly Dataset 2 [12] | 100k | extractable | CS | yes | no |
| ACL-ARC [13] | 11k | extractable | CS/Comp. Ling. | yes | no |
| ACL-AAN [14] | 18k | extractable | CS/Comp. Ling. | yes | no |

evaluation of our reference resolution approach. Sec. 5 is dedicated to statistics and key figures of our data set. We conclude in Sec. 6 with a summary and an outlook.

## 2 Existing Data Sets

In the past, we already published a data set with annotated arXiv papers' content [9]. However, our new data set is superior to this initial version in the following regards:

1. The new data set is considerably larger (1M instead of 90k documents).
2. The new data set provides a similar level of cleanliness regarding the papers' full texts and citation contexts to the old data set.
3. A new method for resolving identical cited documents to the same global identifiers has been developed. Contrary to the old method, the new method has been evaluated and performs very well (see Sec. 4).
4. While the old data set links documents solely to DBLP, which covers computer science papers, the new data set links (cited) documents to the Microsoft Academic Graph, which covers all scientific disciplines.
5. While the old data set is restricted to computer science, the new data set covers all domains of arXiv (see Sec. 5 and Fig. 3).

Table 1 gives an overview of further related data sets. CiteSeerX can be regarded as the most frequently used evaluation data set for citation-based tasks. For our investigation, we use the snapshot of the entire CiteSeerX data set as of October 2013, published in 2015 by [11]. This data set consists of 1,017,457 papers, together with 10,760,318 automatically extracted citation contexts. This data set has the following drawbacks [15,9]: The provided meta-information about cited publications is often not accurate. Citing and cited documents are not interlinked to other data sets. Moreover, the citation contexts can contain noise from non-ASCII characters, formulas, section titles, missed references and/or other "unrelated" references, and do not begin with a complete word.

The PubMed Central Open Access Subset is another large data set that has been used for citation-based tasks [16,17]. Contained publications are

[3] See `https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/`.

already processed and available in XML format. While the data set overall is comparatively clean, heterogeneous annotation of citations within the text and mixed usage of global reference identifiers (PubMed, MEDLINE, DOI, ...) make it difficult to retrieve high quality citation interlinkings of documents from the data set[4] [17].

Beside the abovementioned, there are other collections of scientific publications. Among them are the ACL Anthology corpus [13] and Scholarly Dataset 2 [12]. Note that these data sets only contain the publications themselves, typically in PDF format. Therefore, using such data sets for paper-based or citation-based approaches is troublesome, since one must preprocess the data (i.e., (1) extract the content without introducing too much noise, (2) build global identifiers for cited papers, and (3) annotate citations with those identifiers). Last but not least, data sets for evaluating paper recommendation tasks, such as CiteULike[5] or Mendeley,[6] only provide metadata about publications or are not freely available for research purposes.

## 3 Data Set Creation

Scientific publications are usually distributed in formats targeted at human consumption (e.g. PDF) or, in cases like arXiv, also as source files *for* the aforementioned (e.g. LaTeX sources for generating PDFs). Citation-based tasks, such as context-dependent citation recommendation, in contrast, require automated processing of the publications' textual contents as well as the documents' interlinking through citations. The creation of a data set for such tasks therefore encompasses two main steps: extraction of plain text and resolution of references. In the following we will describe how we approached these two steps using arXiv publications' LaTeX sources and the Microsoft Academic Graph.

### 3.1 Used Data Sets

The following two resources are the basis of the data set creation process.

*arXiv* hosts over 1.4 million submissions from August 1991 onward.[7] They are available not only as PDF, but (in most cases) also as LaTeX source files. The discipline most prominently represented is physics, followed by mathematics, with computer science seeing a continued increase in percentage of submissions

[4] To be more precise, the heterogeneity makes the usage of the data set *as is* unfeasible. Resolving references retrospectively would be an option but comparatively challenging in the case of PubMed because of the frequent usage of special notation in publication titles; see also: `http://www.sciplore.org/files/citrec/CITREC_Parser_Documentation.pdf`.

[5] See `http://citeulike.org/`.

[6] See `https://data.mendeley.com/`.

[7] See `https://arxiv.org/stats/monthly_submissions`.

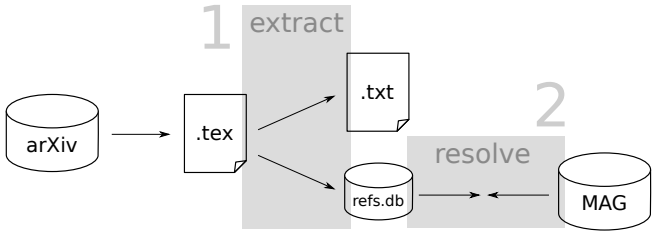


Fig. 1: Schematic representation of the data set generation process.

ranking third (see Fig. 4). The availability of LaTeX sources makes arXiv submissions particularly well suited for extracting high quality plain text and accurate citation information. So much so, that it has been used to generate ground truths for the evaluation of PDF-to-text conversion tools [18].

*Microsoft Academic Graph* is a very large, automatically generated data set on publications, related entities (authors, venues, etc.) and their interconnections through citation. While (comparably noisy) citation contexts are available to some degree, full text documents are not. The size of the MAG makes it a good target for matching reference strings[8] against it, especially given that arXiv spans several fields of study.

### 3.2 Pipeline Overview

To create the data set, we start out with arXiv sources (see Fig. 1). From these we generate, per publication, a plain text file with the document's textual contents and a set of database entries reflecting the document's reference section.[9] In a second step, we then iterate through all reference strings in the database and match them against paper metadata records in the MAG. The result of this process are MAG paper records associated with one or more reference strings, which in turn are associated with citation contexts in the plain text files. In other words, we end up with cited documents described by their MAG metadata and a distributed description of the document, consisting of citation contexts over many citing documents.

### 3.3 LaTeX Parsing

In the following we will describe the tools considered for parsing LaTeX, the challenges we faced in general and with regard to arXiv sources in particular, and our resulting approach.

[8] I.e., the entries in the reference section of a publication. See Lst. 1 for examples.
[9] Association between reference strings and in-text citation locations are preserved by placing citation markers in the text.

Table 2: Comparison of tools for parsing LaTeX.

| Tool | Output | Robust | Usable w/o modification |
|---|---|---|---|
| plastex[10] | DOM | no | yes |
| TexSoup[11] | document tree | no | yes |
| opendetex[12]/detex[13] | plain text | no | yes |
| GrabCite[9] | plain text + resolved references | yes | no |
| LaTeXML[14] | XML | yes | yes |
| Tralics[15] | XML | yes | yes |

**Tools** We took several tools for a direct conversion from LaTeX to plain text or to intermediate formats into consideration and evaluated them. Table 2 gives an overview of our results. Half of the tools failed to produce any output for a large amount of arXiv submissions we used as test input and were therefore deemed not robust enough. *GrabCite* [9] is able to parse 78.5% of arXiv CS submissions but integrates resolving references (see Sec. 3.4) against DBLP into the parsing process and therefore would require significant modification to fit our new system architecture. *LaTeXML* and *Tralics* are both robust and can be used as LaTeX conversion tools as is. Based on subsequent tests we observed that *LaTeXML* needs on average 7.7 seconds (3.3 if formula environments are heuristically removed beforehand) to parse an arXiv submission while *Tralics* needs 0.09. Because the quality of their output seemed comparable we chose to use *Tralics*.

**Challenges** Apart from the general difficulty of parsing LaTeX due to its feature richness and people's free-spirited use of it, we especially note difficulty in dealing with extra packages not included in submissions' sources.[16] While *Tralics*, for example, is supposed to deal with *natbib* citations,[17] normalization of such citations lead to a decrease of citation markers not being able to be matched to an entry in the document's reference section from 30% to 5% in a sample of 565,613 citations we tested.

**Resulting Approach** Our LaTeX parsing solution consists of two steps. First, we flatten each arXiv submission's sources to a single LaTeX file using

[10] See `https://github.com/tiarno/plastex`.
[11] See `https://github.com/alvinwan/texsoup`.
[12] See `https://github.com/pkubowicz/opendetex`.
[13] See `https://www.freebsd.org/cgi/man.cgi?query=detex`.
[14] See `https://github.com/brucemiller/LaTeXML`.
[15] See `https://www-sop.inria.fr/marelle/tralics/`.
[16] The arXiv guidelines specifically suggest the omission of such (see `https://arxiv.org/help/submit_tex#wegotem`).
[17] See `https://www-sop.inria.fr/marelle/tralics/packages.html#natbib`.

Listing 1: Examples of reference strings.

```
(1) V. N. Senoguz and Q. Shafi, arXiv:hep-ph/0412102
(2) V.N. Senoguz and Q. Shafi, Phys. Rev. D 71 (2005) 043514.
(3) V. N. Şenoğuz and Q. Shafi, ''Reheat temperature in supersymmetric
    hybrid inflation models,'' Phys. Rev. D 71, 043514 (2005)
    [hep-ph/0412102].
(4) V.Sauli, JHEP 02, 001 (2003).
(5) Aaij, Roel, et al. "Search for the $B^{0}_{s} \to \eta^{\prime}\phi$
    decay" Journal of High Energy Physics 2017.5 (2017): 158.
(6) According to the numerous discussions with my colleagues <removed>
    and <removed> an experimental verification of our theoretical
    predictions is feasible.
```

*latexpand*[18,19] and normalize `\cite` commands to prevent parsing problems later on. In the second step, we then generate an XML representation of the LaTeX document using *Tralics*, replace formulas, figures, tables and intra-document references with replacement tokens and extract the plain text. Furthermore, each entry in the document's reference section is assigned a unique identifier, its text is stored in a database, and corresponding citation markers are placed in the plain text.

### 3.4 Reference Resolution

Resolving references to globally consistent identifiers (e.g. detecting that the reference strings (1), (2), and (3) in Listing 1 all reference the same document) is a challenging and still unsolved task [19]. Given it is, by itself, the most distinctive part of a publication, we base our reference resolution on the title of the cited work and use other pieces of information (e.g., the authors' names) only in secondary steps. In the following, we will describe the challenges we faced, matching arXiv submissions' reference strings against MAG paper records and how we approached the task.

**Challenges** Reference resolution can be challenging when reference strings contain only minimal amounts of information, when formulas are used in titles or when they refer to non publications (e.g., Listing 1, (4)–(6)). Another concrete problem we encountered was noise in the MAG, as 13,143 reference strings like `K. Kondo, hep-th/0303251.` or `T. Heinzl, hep-th/9812190.` matched MAG paper `2811252340` with the title *"hep-th."*.

**Resulting Approach** Our reference resolution procedure can be broken down in two steps: title identification and matching. If possible, title identification is

[18] See `https://ctan.org/pkg/latexpand`.

[19] We also tested flatex (`https://ctan.org/pkg/flatex`) and flap (`https://github.com/fchauvel/flap`) but got the best results with latexpand.

Listing 2: Excerpts from (top to bottom) a plain text file, corresponding data base entries in refs.db, entries in the MAG and extracted citation context CSV.

```
It has over 79 million images stored at the resolution of FORMULA . Each
image is labeled with one of the 75,062 non-abstract nouns in English, as
listed in the Wordnet{{cite:9ad20b7d-87d1-47f5-aeed-10a1cf89a2e2}}{{cite:
298db7f5-9ebb-4e98-9ecf-0bdda28a42cb}} lexical database.
```

| [uuid] | [in_doc] | [mag_id] | [reference_string] |
|---|---|---|---|
| 9ad20b7d-87d1 -47f5-aeed-.. | 1412.3684 | 2081580037 | George A. Miller (1995). WordNe t: A Lexical Database for Eng.. |
| 298db7f5-9ebb -4e98-9ecf-.. | 1412.3684 | 2038721957 | Christiane Fellbaum (1998), ""W ordNet: An Electronic Lexical.. |

| [paperid] | [originaltitle] | [publisher] | ... |
|---|---|---|---|
| 2038721957 | WordNet : an electronic lexical database | MIT Press | ... |
| 2081580037 | WordNet: a lexical database for English | ACM | ... |

```
2038721957|2081580037|1412.3684|It has over 79 million images stored at
    the resolution of FORMULA . Each image is labeled with one of the
    75,062 non-abstract nouns in English, as listed in the Wordnet CIT
    MAINCIT lexical database. It has been noted that many of the labels
    are not reliable CIT .
```

done by arXiv ID or DOI (where we retrieve the title from an arXiv metadata dump or via crossref.org[20]); otherwise we use Neural ParsCit [20]. The identified title is matched against the normalized titles of all publications in the MAG. Resulting candidates are considered, if at least one of the author's names is present in the reference string. If multiple candidates remain, we judge by the citation count given in the MAG.

### 3.5 Result format

Listing 2 shows some example content from the data set. In addition to the plain text files and references database we also extract the citation contexts of all successfully resolved references for ease of use (see bottom of Listing 2). We choose a citation context length of 3 sentences—the sentence containing the citation as well as the one before and after. For each citation, we store `cited doc MAG ID`, `MAG IDs of adjacent citations`, `citing doc arXiv ID` and `text` in a CSV file. Citations are deemed adjacent, if they are part of a citation group or are at most 5 characters apart (e.g. "*[27,42]*", *"[27], [42]"* or *"[27] and [42]"*). Sentence tokenization is performed with NLTK's pre-trained PunktSentenceTokenizer.

[20] See `https://www.crossref.org/`.

Table 3: Confidence intervals for a sample size of 300 with 297 positive results as given by Wilson score interval and Jeffreys interval [21].

| Confidence level | Method | Lower limit | Upper limit |
|---|---|---|---|
| 0.99 | Wilson | 0.9613 | 0.9975 |
| | Jeffreys | 0.9666 | 0.9983 |
| 0.95 | Wilson | 0.9710 | 0.9966 |
| | Jeffreys | 0.9736 | 0.9972 |

Table 4: Mismatched documents

| # | | Document |
|---|---|---|
| 1 | matched | *"The Maunder Minimum"* (John A. Eddy; 1976) |
| | correct | *"The Maunder Minimum: A reappraisal"* (John A. Eddy; 1983) |
| 2 | matched | *"Support Vector Machines"* (Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani; 2013) |
| | correct | *"1-norm Support Vector Machines"* (Ji Zhu, Saharon Rosset, Robert Tibshirani, Trevor J. Hastie; 2003) |
| 3 | matched | *"The Putative Liquid-Liquid Transition is a Liquid-Solid Transition in Atomistic Models of Water"* (David Chandler, David Limmer; 2013) |
| | correct | *"The putative liquid-liquid transition is a liquid-solid transition in atomistic models of water. II"* (David T. Limmer, David Chandler; 2011) |

## 4 Evaluation of Reference Resolution

To evaluate the quality of our reference resolution results, we take a random sample of 300 matched reference strings and manually check if the correct record in the MAG was identified by our method.[21] Given the 300 items, we obtained 3 errors, giving us an accuracy estimate of 96% at the worst, as shown in Table 3. Table 4 shows the three incorrectly identified documents. In all three cases the misidentified document's title is contained in the correct document's title, and there is a large or complete author overlap between correct and actual match. This shows that authors sometimes title follow-up work very similarly, which leads to hard to distinguish cases.

[21] Details can be found at `https://github.com/IllDepence/unarXive/tree/master/doc/matching_evaluation`.

# 5 Statistics and Key Figures

## 5.1 Creation Process

We used an arXiv source dump containing all submissions up until the end of 2018 (1,492,923 documents). 114,827 of these were only available in PDF format, leaving 1,378,096 sources. Our pipeline output 1,283,584 (93.1%) plain text files, 1,139,790 (82.7%) of which contained citation markers. The number of reference strings identified is 39,694,083, for which 63,633,427 citation markers were placed within the plain text files. This first part of the process took 67 hours to run, unparallalized on a 8 core Intel Core i7-7700 3.60GHz machine with 60 GB of memory.

Of the 39,694,083 reference strings, we were able to match 16,926,159 (42.64%). For 31.32% of the reference strings we could neither find an arXiv ID or DOI, nor was Neural ParsCit able to identify a title. For the remaining 26.04% a title was identified but could not be matched with the MAG. Of the matched 16.9 million items' titles, 52.60% were identified via Neural ParsCit, 28.31% by DOI and 19.09% by arXiv ID. Of the identified DOIs 32.9% were found as is while 67.1% were heuristically determined[22]. The matching process took 119 hours, run in 10 parallel processes on a 64 core Intel Xeon Gold 6130 2.10GHz machine with 500 GB of memory.

Looking only at the numbers for arXiv submissions from 2018 (i.e. recent content), we note that the percentage of pipeline output goes up from 93.1 to 95.9% (82.7 to 87.8% only counting plain text files containing citation markers) and the reference resolution percentage increases from 42.64 to 59.39%.

## 5.2 Resulting Data Set

Our data set consists of *2,746,288 cited papers, 1,043,126 citing papers, 15,954,664 references and 29,203,190 citation contexts.*[23]

Figure 2 shows the number of citing documents for all cited documents. There is one cited document with over 10,000 citing documents, another 8 with more than 5,000 and another 14 with more than 3,000. 1,485,074 (54.07%) of the cited documents are cited at least two times, 646,509 (23.54%) at least five times. The mean number of citing documents per cited document is 5.81 (SD 28.51). Figure 3 shows the number of citation contexts per entry in a document's reference section. 10,537,235 (66.04%) entries have only one citation context, the maximum is 278, the mean 1.83 (SD 2.00). This means a cited document is described by on average $1.83 \times 5.81 \approx 10.63$ citation contexts.

Figure 4 depicts the flow of citations by field of study for all 15.9 million matched references. Fields of study with very small numbers of references are combined to *other* for legibility reasons. For the citing document's side, these are economics, electrical engineering and systems science,

[22] This was possible because the DOIs of articles in journals of the American Physical Society follow predictable patterns.

[23] References with no citation markers (due to parsing errors) are not counted here.

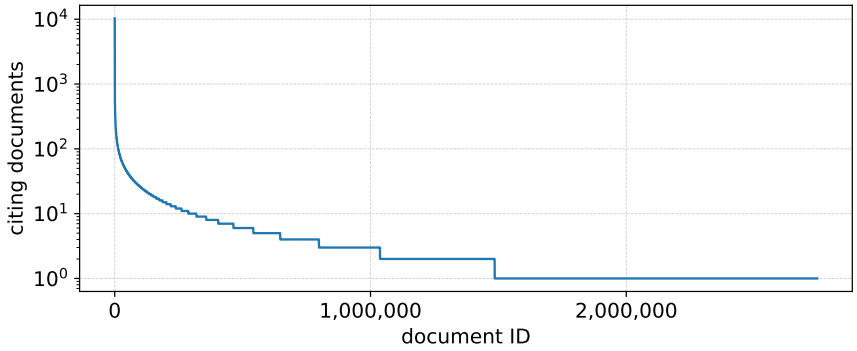


Fig. 2: Number of citing documents per cited document.

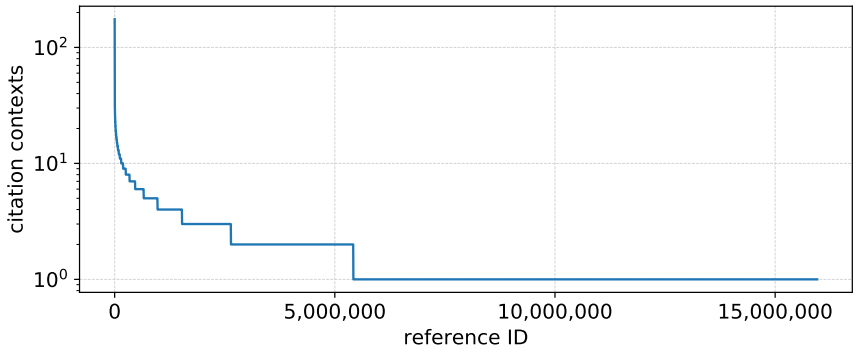


Fig. 3: Number of citation contexts per reference.

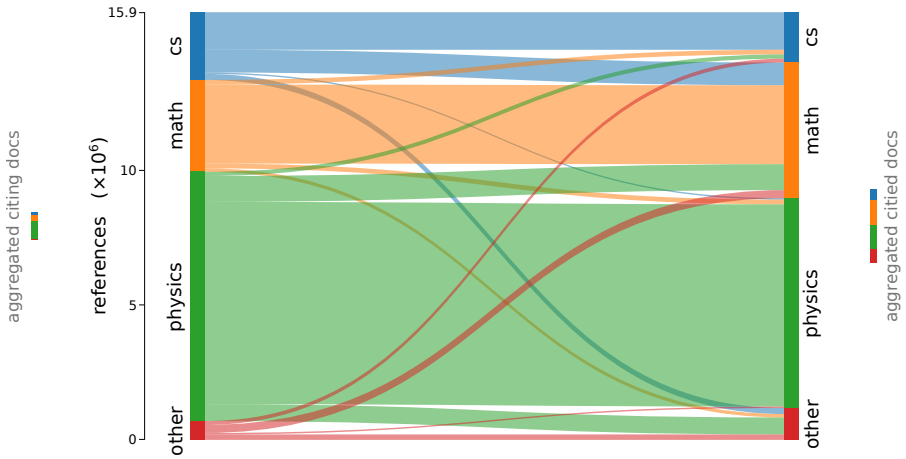


Fig. 4: Citation flow by field of study for 15.9 million references. The number of citing and cited documents per field of study are plotted on the sides.

quantitative biology, quantitative finance and statistics. Combined on the cited document's side are chemistry, biology, engineering, materials science, economics, geology, psychology, medicine, business, geography, sociology, political science, philosophy, environmental science and art. To no surprise, publications in each field are cited the most from within the field itself. Notable is, however, that the incoming citations in mathematics are the most varied (physics and computer science combined make up 35% of the citations).

## 6 Conclusion

Evaluating and applying approaches of research paper-based and citation-based tasks typically requires large, high-quality, citation-annotated, interlinked data sets. In this paper, we proposed a new data set with over one million papers' fulltexts, 29.2 million annotated citations, and 29.2 million extracted citation contexts (of three sentences each), ready to be used by researchers and practitioners. We provide the data set and the implementation of creating the data set based on arXiv source files online for further usage.

For the future, we plan to use the data set for a variety of tasks. Among others, we will develop a citation recommendation system based on all arXiv papers. Furthermore, we plan to analyze citations across scientific disciplines, and to use the differences in the citing behavior for enhanced citation recommendation.

**Acknowledgements.** This research has been supported by the Research Innovation Fund of the University of Freiburg (#2100189801).

## References

1. Beel, J., Gipp, B., Langer, S., Breitinger, C.: Research-paper recommender systems: a literature survey. Int. J. on Digital Libraries **17**(4) (2016) 305–338
2. Teufel, S., Siddharthan, A., Tidhar, D.: Automatic classification of citation function. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. EMNLP'07 (2006) 103–110
3. Ghosh, S., Das, D., Chakraborty, T.: Determining Sentiment in Citation Text and Analyzing Its Impact on the Proposed Ranking Index. In: Proceedings of the 17th International Conference on Computational Linguistics and Intelligent Text Processing. CICLing'16 (2016) 292–306
4. He, Q., Pei, J., Kifer, D., Mitra, P., Giles, C.L.: Context-aware Citation Recommendation. In: Proceedings of the 19th International Conference on World Wide Web. WWW'10 (2010) 421–430
5. Ebesu, T., Fang, Y.: Neural Citation Network for Context-Aware Citation Recommendation. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. Tokyo, Japan (2017) 1093–1096
6. Mohammad, S., Dorr, B.J., Egan, M., Awadallah, A.H., Muthukrishnan, P., Qazvinian, V., Radev, D.R., Zajic, D.M.: Using Citations to Generate surveys

of Scientific Paradigms. In: Proceedings of the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. NAACL-HLT'09 (2009) 584–592
7. Chen, J., Zhuge, H.: Automatic generation of related work through summarizing citations. Concurrency and Computation: Practice and Experience **31**(3) (2019)
8. Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.P., Wang, K.: An Overview of Microsoft Academic Service (MAS) and Applications. In: Proceedings of the 24th International Conference on World Wide Web. WWW'15 (2015) 243–246
9. Färber, M., Thiemann, A., Jatowt, A.: A High-Quality Gold Standard for Citation-based Tasks. In: Proceedings of the 11th International Conference on Language Resources and Evaluation. LREC'18 (2018)
10. Caragea, C., Wu, J., Ciobanu, A.M., Williams, K., Ramírez, J.P.F., Chen, H., Wu, Z., Giles, C.L.: CiteSeer x : A Scholarly Big Dataset. In: Proceedings of the 36th European Conference on IR Research. ECIR'14 (2014) 311–322
11. Huang, W., Wu, Z., Chen, L., Mitra, P., Giles, C.L.: A Neural Probabilistic Model for Context Based Citation Recommendation. AAAI'15 (2015) 2404–2410
12. Sugiyama, K., Kan, M.: A Comprehensive Evaluation of Scholarly Paper Recommendation Using Potential Citation Papers. International Journal on Digital Libraries **16**(2) (2015) 91–109
13. Bird, S., Dale, R., Dorr, B.J., Gibson, B.R., Joseph, M.T., Kan, M., Lee, D., Powley, B., Radev, D.R., Tan, Y.F.: The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation. LREC'08 (2008)
14. Radev, D.R., Muthukrishnan, P., Qazvinian, V., Abu-Jbara, A.: The ACL anthology network corpus. Language Resources and Evaluation **47**(4) (2013) 919–944
15. Roy, D., Ray, K., Mitra, M.: From a Scholarly Big Dataset to a Test Collection for Bibliographic Citation Recommendation. SBD'16 (2016)
16. Duma, D., Klein, E., Liakata, M., Ravenscroft, J., Clare, A.: Rhetorical Classification of Anchor Text for Citation Recommendation. D-Lib Magazine **22** (2016)
17. Gipp, B., Meuschke, N., Lipinski, M.: CITREC: An Evaluation Framework for Citation-Based Similarity Measures based on TREC Genomics and PubMed Central. In: Proceedings of the iConference 2015. (2015)
18. Bast, H., Korzen, C.: A Benchmark and Evaluation for Text Extraction from PDF. In: Proceedings of the 2017 ACM/IEEE Joint Conference on Digital Libraries. JCDL'17 (2017) 99–108
19. Nasar, Z., Jaffry, S.W., Malik, M.K.: Information extraction from scientific articles: a survey. Scientometrics **117**(3) (Dec 2018) 1931–1990
20. Prasad, A., Kaur, M., Kan, M.Y.: Neural ParsCit: A Deep Learning Based Reference String Parser. International Journal on Digital Libraries **19** (2018) 323–337
21. Brown, L.D., Cai, T.T., DasGupta, A.: Interval estimation for a binomial proportion. Statistical Science **16**(2) (2001) 101–133