Scholarly Big Data Quality Assessment: A Case Study of Document Linking and Conflation with S2ORC

Jian Wu, Ryan Hiltabrand, Dominik Soós
Old Dominion University
Norfolk, Virginia, United States

C. Lee Giles
Pennsylvania State University
University Park, Pennsylvania, United States

Abstract

Recently, the Allen Institute for Artificial Intelligence released the Semantic Scholar Open Research Corpus (S2ORC), one of the largest open-access scholarly big datasets with more than 130 million scholarly paper records. S2ORC contains a significant portion of automatically generated metadata. The metadata quality could impact downstream tasks such as citation analysis, citation prediction, and link analysis. In this project, we assess the document linking quality and estimate the document conflation rate for the S2ORC dataset. Using semi-automatically curated ground truth corpora, we estimated that the overall document linking quality is high, with 92.6% of documents correctly linking to six major databases, but the linking quality varies depending on subject domains. The document conflation rate is around 2.6%, meaning that about 97.4% of documents are unique. We further quantitatively compared three near-duplicate detection methods using the ground truth created from S2ORC. The experiments indicated that locality-sensitive hashing was the best method in terms of effectiveness and scalability, achieving high performance (F1=0.960) and a much reduced runtime. Our code and data are available at https://github.com/lamps-lab/docconflation.

CCS Concepts

• **Information systems** → *Digital libraries and archives*; *Data cleaning*; *Deduplication*; *Entity resolution*; *Data cleaning*.

Keywords

scholarly big data, data quality, document linking, document conflation, deduplication

ACM Reference Format:
Jian Wu, Ryan Hiltabrand, Dominik Soós and C. Lee Giles. 2018. Scholarly Big Data Quality Assessment: A Case Study of Document Linking and Conflation with S2ORC. In *Proceedings of The 22th ACM Symposium on Document Engineering (DocEng '22).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Since the inception of scholarly big data, several large-scale scholarly big datasets have been released, such as CORE [11], Microsoft Academic Graph (MAG; [19]), and CiteSeerX [8], and much research has been conducted based on mining these datasets [24]. One widely used scholarly big dataset is the Semantic Scholar Open Research Corpus (S2ORC) [13], containing more than 130 million English academic paper records spanning multiple disciplines. This corpus has been used in many research projects, such as generating the CORD-19 dataset [20], scientific claim verification [18], and citation intent classification [16].

When using these datasets, most studies assumed perfect metadata quality. However, existing studies have found that automatically extracted metadata could contain errors that may not be ignored when used for downstream analysis. For example, it was found that named entities (people, organization) appearing in the acknowledgement sections are not all actually acknowledged [23]. The accuracy of citation data problem was also revealed for Web of Science and Scopus datasets [17].

Scholarly big data quality covers many aspects [9], such as coverage, metadata accuracy, and linkage accuracy. Automatic assessments of all aspects of a large scale scholarly dataset is challenging. In this work, we showcase quantitative assessments of document linking and conflation quality of S2ORC. The quality metrics of the two aspects can be taken into account to estimate potential uncertainties for future studies involving document linking, citation analysis, and bibliometric studies. The ground truth datasets we compiled can be used as benchmarks for developing automatic document linking and conflation models. In addition, we compared three methods for near-duplication detection and found that the Locality-Sensitive Hashing (LSH) method is the most effective and scalable method, which could be used for *automatic* document linking and conflation rate assessment.

2 Related Work

Data quality for digital libraries is a long-standing problem. Early works pointed out that completeness and accuracy errors were designated as important aspects of quality dimensions in digital libraries, e.g., [6]. Bui assessed the metadata quality of the National Science Digital Library [2]. A systematic survey of metadata quality in digital repositories was conducted by Park [15]. Recently, researchers started paying attention to the quality of large datasets used for machine learning, deep learning, and downstream analytical works [3]. Several works included assessments of scholarly big datasets, e.g., CiteSeerX [22] and MAG [19].

Document conflation (aka near-duplicate detection) has also been studied. Williams et al. compared *simhash* and *shingle* methods using a set of near-duplication pairs compiled from CiteSeerX [21]. A recent survey reviewed several methods for record duplication detection [1]. Ge et al. proposed a greedy graph-based algorithm [7] for conflating text documents. Our work provides a case study of assessing document quality of a real-world scholarly big dataset.

DocEng '22, September 20–23, 2022, Virtually hosted from San Jose, CA, USA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

3 Previous S2ORC Quality Assessment

In the original paper [13], the authors evaluated paper clustering (i.e., document conflation) by randomly sampling 500 paper clusters, restricting to those with PDFs, and comparing whether the title and authors of papers in each cluster match metadata in PDFs. Document conflation or near-duplicate detection is a task to identify papers that were published (officially or unofficially) in different versions. The evaluation in [13] indicated that the title matching accuracy was 93% and the author matching accuracy was 89%. The authors also evaluated the bibliography linking, in which 500 papers were parsed by GROBID [14] and 100 papers were processed by a LaTeX parser. The evaluation was performed by checking whether the titles and authors in the bibliography entries matched with the linked papers. The title matching accuracy was 1.00 for both GROBID the LaTeX parser. The author matching accuracy was 0.96 for GROBID and 0.92 for the LaTeX parser.

One caveat of the document linking and conflation assessment in [13] was that the samples were drawn based on *post-processing* data. Therefore, the reported results reflect the *precision* of document conflation. The fraction of near-duplicate documents that should have been conflated into a cluster, which reflected the *recall*, was not incorporated. We conduct a more careful assessment of document linking and document conflation quality.

4 Data

The S2ORC version we evaluated was the one released on July 5, 2020. The corpus covered 136M+ paper nodes with 12.7M+ full-text papers connected by 467M+ citation edges by unifying data from different sources.

Document linking Data Due to the large base number of the S2ORC corpus, manual verification of even 1% (about 1 million) papers is infeasible. To mitigate this potential bias, we selected samples using different criteria, aiming at estimating the quality from different aspects. For document linking assessment, we compiled three ground truth corpora based on different criteria.

- **DL-Ran** comprises 500 randomly selected papers from S2ORC. This corpus was used to estimate the overall linking accuracy.
- **DL-Sub** contains five corpora each containing 100 papers randomly selected from five subjects, including physics (Phy), chemistry (Chem), computer science (CS), medicine (Med), and psychology (Psych). These corpora were used for comparing linking accuracy in major subject domains.
- **DL-Nolink** consisted of 500 papers that were not linked to any external databases. This corpus was used for estimating the overall fraction of missing links.

Document Conflation Data Because S2ORC does not contain a field indicating two or more papers are near-duplicates, it is not possible to directly assess the conflation rate. Therefore, we compiled a ground truth corpus – DC3 curated from a parent sample of 150,000 randomly selected S2ORC papers. To compare near-duplication detection algorithms, we used DC3 and the CiteSeerX dataset from an existing study, consisting of 360 pairs of manually identified near-duplicate documents from CiteSeerX [21].

5 Linking Quality Assessment

Ground truth. In S2ORC, each paper has a unique ID. If the paper is found in another database, the two documents are linked by adding the external database ID as a metadata field of the S2ORC paper. S2ORC links papers to arXiv, ACL Anthology, PMC, PubMed, and MAG. We also check if the paper has a valid DOI. To build the ground truth, we queried the title of *each paper* on Google under the digital library domain using the `site:` option[1], such as:

```
the impact of migration on trade site:arxiv.org.
```

We attempted to identify the counterpart paper in the external digital libraries by matching titles, authors, years, and venues. If we could not find the paper on the first two pages of Google's search result, we directly queried the title on the digital library's native search interface. If we could not find the paper on both search interfaces, we determined this paper was not cross-listed on the external database.

Assessment. To assess the document linking quality, we directly compare the external digital library IDs in the ground truth against the IDs recorded in S2ORC. The results are shown in Table 1. Because a paper may be linked to different external databases, we also reported the micro- and macro- precision, recall, and F1 scores. Table 1 also reports the accuracy, which incorporates true positives and true negatives. To make a comparison with the assessment in the original S2ORC paper, we also calculated the document-level precision P_D, which is the fraction of S2ORC *papers* with all links correctly identified.

The document linking precision obtained in the DL-Ran (0.926) was lower than what was reported in the original S2ORC paper [13] by 7.4% (title matching) and was roughly consistent with the author matching result[2]. The assessment results across the five subjects indicated that the "Medicine" subject have the lowest linking quality, with 15% of the papers containing at least one wrong link to external databases. Assuming the number of papers containing at least one wrong link within 100 randomly selected documents follows a Gaussian distribution and a conservative range of deviation is $[0, 20]$, we apply the 1/5th rule [4] to estimate the standard deviation $\sigma \approx 20/5 = 4$, so the difference between DL-Med and other disciplines is significant. The differences between disciplines other than DL-Med are insignificant. Psychology papers have the highest fraction of documents with all links correct (0.990), highest linking accuracy (0.998), and highest Micro-F1 (0.997). The relatively low P_D found in the DL-Nolink corpus (0.908) indicates that ($\approx$ 10%) of papers that do not link to any external databases in S2ORC should have been linked to at least one external database.

6 Conflation Rate Estimation

The conflation rate indicates the proportion of near-duplicate documents in a digital corpus. The time complexity of a pairwise comparison is $O(n^2)$, so a pairwise comparison of all papers in S2ORC is infeasible. To build the ground truth, we first randomly selected 150,000 papers. We then used a brute-force method to find near-duplicate candidates and then manually verified the candidates. This method contains the following steps.

[1] arXiv: arxiv.org; ACL: aclanthology.org; PMC: ncbi.nlm.nih.gov/pmc; PubMed: pubmed.ncbi.nlm.nih.gov; MAG: academic.microsoft.com; DOI: doi.org

[2] The original paper did not specify the external databases that were linked to.

Sample	N_D	N_{GT}	N_{S2ORC}	Macro-P	Macro-R	Macro-F1	Micro-P	Micro-R	Micro-F1	Accuracy	P_D
DL-Ran	500	937	916	0.987	0.973	0.976	0.985	0.963	0.974	0.984	0.926
DL-Phy	100	187	192	0.967	0.990	0.985	0.969	0.995	0.982	0.988	0.940
DL-Chem	100	176	177	0.975	0.980	0.997	0.983	0.989	0.986	0.992	0.970
DL-CS	100	161	164	0.985	0.995	0.987	0.982	0.994	0.988	0.993	0.960
DL-Med	100	219	206	0.997	0.935	0.955	0.995	0.936	0.965	0.975	0.850
DL-Psych	100	160	159	1.000	0.995	0.997	1.000	0.994	**0.997**	**0.998**	**0.990**
DL-Nolink	500	276	0	–	–	–	–	–	–	–	0.908

Table 1: Document linking assessment. N_D: the number of papers. N_{GT}: the number of links found in the ground truth corpus. N_{S2ORC}: the number of links in S2ORC. We do not calculate $P/R/F1$ for DL-Nolink because there are no true positive samples.

(1) We took the $n_w \in \{4, 5, 6\}$ longest words from each paper's title, and sort them by lengths in descending order. If two words of the same length appeared, we put the leftmost first.
(2) We generated two keys for each paper. The first key contained the concatenation of the first $(n_w - 1)$ words. The second key contained the concatenation of the second to the n_w-th word.
(3) We appended the first two authors' last names to Keys 1 and 2, resulting in 4 keys. For example, if a paper's title is "one two three four five six seven eight" and the authors are "John Doe" and "Tom Smith", the keys generated when $n_w = 5$ are listed below. If the title length is less than n_w, we use all words.
```
three_seven_eight_four_Doe
three_seven_eight_four_Smith
seven_eight_four_five_Doe
seven_eight_four_five_Smith
```
(4) We matched the keys of different papers to find near-duplicate candidates that share at least one common key.
(5) We allow near-duplicate papers in DC3 published within ±1 years. To justify this choice, we constructed another dataset called DC1, in which near-duplicate papers were published in the same year.
(6) We manually verified candidates by examining their titles, authors, venues, and publication year in the metadata. We downloaded and compared PDF files if metadata alone was not sufficient to determine duplication. Papers identified near-duplicates are conflated into a cluster.

Corpus	N_C	N_D	$N_C(S = 2)$	$N_C(S \geq 3)$
DC1	513	1047	500 (97.5%)	13 (2.5%)
DC3	3286	7191	2927 (89.1%)	359 (10.9%)

Table 2: Properties of DC1 and DC3. N_C: the number of clusters; N_D: the number of papers; S: cluster size, i.e., the number of near-duplicates in a cluster.

Following the steps above, we generated the DC3 dataset (Table 2). The significant increase of N_C of DC3 compared with DC1 indicates that the majority of near-duplicate papers were not published in the same year. In particular, the number of clusters with more than 3 near-duplicates significantly increases from 2.5% to 10.9% in DC3. Therefore, compared with DC1, DC3 better represents the majority of near-duplicates in its parent sample. Using DC3, the conflation rate of S2ORC is estimated as $(N_D - N_C)/N_P \approx 2.6\%$, which is lower than the CiteSeerX conflation rate (11%) [22].

7 Near-duplicate Detection Methods

The documents in S2ORC were conflated using a *fuzzy matching* method [13] by calculating the Jaccard index between unigrams extracted from titles. If the Jaccard index between a pair of papers is greater or equal to a threshold J_0, the two papers are determined to be near-duplicates. It requires comparing all unigrams in a title of a paper against the other. This method has a time complexity of $O((mn)^2)$ where m is the average number of unigrams in paper titles and n is the number of papers to be conflated.

We propose using LSH, a more efficient method with sacrificing a marginal accuracy, LSH is an algorithm that breaks an input string into pieces (shingles) and hashes similar strings into the same "buckets" with high probability [12]. It has been used as an efficient method to resolve near-duplicate news articles [5]. The method is controlled by three parameters: the number of shingles (k_s), the Jaccard similarity threshold (J_0) that was used for calculating the similarity between two sets of shingles, and the permutation C that was the number of ways shingles were ordered. This method first calculates a score for each title. It then places papers with the same scores (near-duplicates) into buckets (or clusters). The best case time complexity of this method is sublinear [10].

As a baseline, we also compare against the strict string matching method, in which we directly compare the full title of a paper against the titles of other papers. Title text was normalized by lowercasing all letters. We indexed title strings using a hash index, so we only needed to compare a single hash value for a paper. The time complexity with this method is is $O(n^2)$.

To find the best performance, we set four different thresholds J_0 for the *fuzzy matching* method. We also used different combinations of parameters of (k_s, J_0, C) for LSH. The performance is evaluated using standard metrics: precision (P), recall (R), and F1. Here, a positive sample is a pair of documents identified as near-duplicates.

Table 3 shows the performance of three methods under different settings. The results indicate that LSH is the best method with both high performance and short runtime. The best combination of parameters is (k_s,J_0,C)=(10, 0.5, 128), which takes only about 1 minute to process all papers in DC3 and achieves an F1= 0.960. The Fuzzy-matching method achieves an almost perfect F1 when J_0 = 0.85 or 0.9, with a dramatic long runtime of 3 hours. The strict title matching was the fastest with a poor F1. We ran experiments on a server with 24 Xeon cores, 384GB RAM, and MySQL 8.0.

We also applied the three methods to the CiteSeerX dataset. LSH again outperformed the other methods, achieving the highest F1

Data	DC3				
Method	Parameters	P	R	F1	T(s)
Fuzzy (J_0)	0.80	1.000	0.996	0.998	12,288
	0.85	1.000	0.999	**1.000**	**10,997**
	0.90	1.000	0.999	**1.000**	**11,435**
	0.95	1.000	0.991	0.996	10,683
Strict	–	1.000	0.416	0.588	14
LSH (k_s,J_0,C)	(5,0.5,128)	0.836	0.934	0.882	63
	(5,0.5,256)	0.965	0.857	0.908	61
	(10,0.5,128)	0.987	0.934	**0.960**	62
	(10,0.5,256)	0.996	0.840	0.911	78
	(5,0.75,128)	1.000	0.738	0.850	43
	(5,0.75,256)	1.000	0.766	0.867	54
	(10,0.75,128)	1.000	0.684	0.812	80
	(10,0.75,256)	1.000	0.706	0.828	107
	(5,0.9,128)	1.000	0.556	0.715	35
	(5,0.9,256)	1.000	0.574	0.729	57
	(10,0.9,128)	1.000	0.525	0.689	79
	(10,0.9,256)	1.000	0.536	0.698	92
Data	CiteSeerX				
Fuzzy	0.8	0.565	0.942	0.706	767
Strict	–	0.738	0.377	0.499	10
LSH (k_s,J_0,C)	(5,0.5,256)	0.811	0.885	**0.846**	**3**

Table 3: Evaluation results of near-duplicate detection methods using DC3 and CiteSeerX. T(s) is the runtime in seconds.

along with the shortest runtime. The F1 of the *fuzzy matching* method is no longer close to perfect, which is likely because of the imperfection of the CiteSeerX metadata. By inspecting the title, we found several short titles were incorrect, such as "REFERENCES".

8 Discussion

The problems of document linking and conflation are connected. In this study, to build a high fidelity ground truth, it was necessary to manually verify each link and near-duplicate candidate pair. This method is not scalable for large-scale datasets. Given that many external databases offer query APIs, it is possible to automatically assess linking quality by querying these APIs. To verify if two bibliographic records in different databases match, an effective and scalable near-duplication detection method is desired.

We notice that the parameter settings of LSH that produced the best performance varied for S2ORC and the CiteSeerX corpus, indicating that parameter tuning may be needed for specific datasets.

Our experiments also indicated that most papers can be conflated with high accuracy by comparing only titles. In particular, the *fuzzy matching* method achieved almost a perfect F1, despite of the long runtime. Our preliminary experiments indicate that using abstract, e.g., [21], may decrease the accuracy because the abstracts of near-duplicate papers exhibit higher variance compared with titles.

9 Conclusion

In this study, we assessed the document linking quality and estimated the document conflation rate of the S2ORC dataset. The document linking quality of S2ORC is generally high. Over 92% of papers are correctly linked to six major databases, but the fraction varies depending on subject domains. The document conflation rate is at least 2.6%, estimated by a ground truth dataset consisting of 3286 near-duplicate clusters of various sizes. Using the same ground truth, we compared three near-duplicate detection methods. The LSH method outperformed the other two with a relatively high performance (F1=0.960) and a much shorter runtime. Our results reveal the data quality issues of applying AI-methods to build scholarly big data, which motivate the effort of improving data quality by looping human effort. The relatively low document linking quality in medicine (Table 1) implies that one should be careful when curating a subset in this discipline using S2ORC metadata. The LSH method can further be used for developing efficient algorithms on plagiarism detection.

Acknowledgments

We gratefully acknowledge partial support from the National Science Foundation (Award #1823288).

References

[1] Saleh Rehiel Alenazi, Kamsuriah Ahmad, and Akeem Olowolayemo. 2017. A review of similarity measurement for record duplication detection. In *ICEEI*.
[2] Yen Bui and Jung-ran Park. 2006. An assessment of metadata quality: A case study of the national science digital library metadata repository. In *Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l'ACSI*.
[3] Li Cai and Yangyong Zhu. 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Sci. J.* 14 (2015), 2.
[4] MO Columb and MS Atkinson. 2015. Statistical analysis: sample size and power estimations. *BJA Education* 16, 5 (2015), 159–161.
[5] Abhinandan S. Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. 2007. Google News Personalization: Scalable Online Collaborative Filtering. In *WWW*.
[6] Christopher J. Fox, Anany Levitin, and Thomas C. Redman. 1994. The Notion of Data and Its Quality Dimensions. *Inf. Process. Manag.* 30, 1 (1994), 9–20.
[7] Youming Ge, Jiefeng Wu, Genan Dai, and Yubao Liu. 2019. Text Deduplication with Minimum Loss Ratio. In *Proceedings of ICMLC*.
[8] C. Lee Giles, Kurt D. Bollacker, and Steve Lawrence. 1998. CiteSeer: An Automatic Citation Indexing System. In *Proceedings of JCDL*.
[9] Thomas N. Herzog, Fritz J. Scheuren, and William E. Winkler. 2007. *Data quality and record linkage techniques*. Springer.
[10] Omid Jafari, Preeti Maurya, Parth Nagarkar, et al. 2021. A Survey on Locality Sensitive Hashing Algorithms and their Applications.
[11] Petr Knoth and Zdenek Zdráhal. 2012. CORE: Three Access Levels to Underpin Open Access. *D Lib Mag.* 18, 11/12 (2012).
[12] Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. 2014. *Mining of Massive Datasets* (2nd ed.). Cambridge University Press, USA.
[13] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of ACL*.
[14] Patrice Lopez. 2009. GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In *Proceedings of ECDL*.
[15] Jung-Ran Park. 2009. Metadata Quality in Digital Repositories: A Survey of the Current State of the Art. *Cataloging & Classification Quarterly* 47, 3-4 (2009).
[16] Muhammad Roman, Abdul Shahid, Shafiullah Khan, Anis Koubaa, and Lisu Yu. 2021. Citation Intent Classification Using Word Embedding. *IEEE Access* 9 (2021).
[17] Nees Jan van Eck and Ludo Waltman. 2017. Accuracy of citation data in Web of Science and Scopus. In *Proceedings of ISSI*.
[18] David Wadden, Shanchuan Lin, Kyle Lo, et al. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of EMNLP*.
[19] Kuansan Wang, Zhihong Shen, et al. 2020. Microsoft Academic Graph: When experts are not enough. *Quantitative Science Studies* 1, 1 (02 2020), 396–413.
[20] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, et al. 2020. CORD-19: The Covid-19 Open Research Dataset. *CoRR* abs/2004.10706 (2020).
[21] Kyle Williams and C. Lee Giles. 2013. Near Duplicate Detection in an Academic Digital Library. In *Proceedings of DocEng*.
[22] Jian Wu, Chen Liang, Huaiyu Yang, and C. Lee Giles. 2016. CiteSeerX data: semanticizing scholarly papers. In *Proceedings of SBD@SIGMOD*.
[23] Jian Wu, Pei Wang, Xin Wei, et al. 2020. Acknowledgement Entity Recognition in CORD-19 Papers. In *Proceedings of SDP@EMNLP*.
[24] Feng Xia, Wei Wang, Teshome Megersa Bekele, and Huan Liu. 2017. Big Scholarly Data: A Survey. *IEEE Transactions on Big Data* 3, 1 (2017), 18–35.