A Blocking-Based Approach to Enhance Large-Scale Reference Linking

Tarek Saier, Meng Luan and Michael Färber

Karlsruhe Institute of Technology (KIT), Institute AIFB, Kaiserstr. 89, 76133 Karlsruhe, Germany

Abstract

Analyses and applications based on bibliographic references are of ever increasing importance. However, reference linking methods described in the literature are only able to link around half of the references in papers. To improve the quality of reference linking in large scholarly data sets, we propose a blocking-based reference linking approach that utilizes a rich set of reference fields (title, author, journal, year, etc.) and is independent of a target collection of paper records to be linked to. We evaluate our approach on a corpus of 300,000 references. Relative to the original data, we achieve a 90% increase in papers linked through references, a five-fold increase in bibliographic coupling, and a nine-fold increase in in-text citations covered. The newly established links are of high quality (85% F1). We conclude that our proposed approach demonstrates a way towards better quality scholarly data.

Keywords

entity resolution, references, blocking, bibliometrics, scholarly data, digital libraries

1. Introduction

Scholarly data is becoming increasingly important and with it its quality and coverage. Connections between publications in the form of literature references are of particular importance, as they are used as a basis for various analyses, decision making, and applications. Some examples are research output quantification [1], trend detection [2], summarization [3], and recommendation [4, 5].

However, reference linking methods[1] described in the literature are only able to link around half of the references contained in the original papers to the cited publications [6, 7]. This lack in coverage is especially affecting references to non-English publications [8], which are in general underrepresented in scholarly data [9, 10, 11, 12] along with publications in the humanities [13, 14].

We see the reason for this lack in linked references in two key shortcomings of current methods. First, references are linked using simple string similarity measures that are often relying *only* on publications' title and author information (which is not always contained in references; see Figure 1). Second, references are exclusively linked to a target collection of

ULITE2022: Understanding Literature References in Academic Full Text, JCDL 2022, Cologne, Germany, June 24, 2022

tarek.saier@kit.edu (T. Saier); lm19940625@163.com (M. Luan); michael.faerber@kit.edu (M. Färber)

0000-0001-5028-0109 (T. Saier); 0000-0001-5458-8645 (M. Färber)

[1] We use "link[ing/ed] references" w.r.t. to connections to cited papers rather than in-text citation markers.

[10] I. Bonalde et al., Phys. Rev. Lett. **85**, 4775 (2000).

[25] Bonalde I, Yanoff B D, Salamon M B, Van Harlingen D J, Chia E M E, Mao Z Q and Maeno Y 2000 Phys. Rev. Lett. **85** 4775

[4] Jaume, S.C. and Sykes, L.R., Pure and Applied Geophysics **155**, 279-305.

Jaume, S.C. and L.R. Sykes, Evolving Towards a Critical Point: A Review of Accelerating Seismic Moment/Energy Release Prior to Large and Great Earthquakes, Pure Appl. Geophys., 155, 279, 1999.

Figure 1: Examples of challenging reference pairs from our evaluation that where successfully matched. **Top:** references from `arXiv:cond-mat/0503317` (no title, first author only) and `arXiv:cond-mat/0104493` (no title, all authors). **Bottom:** references from `arXiv:cond-mat/0104341` (no title, full venue, page rage, no year) and `arXiv:physics/0504218` (with title, venue abbreviation, start page only, with year).

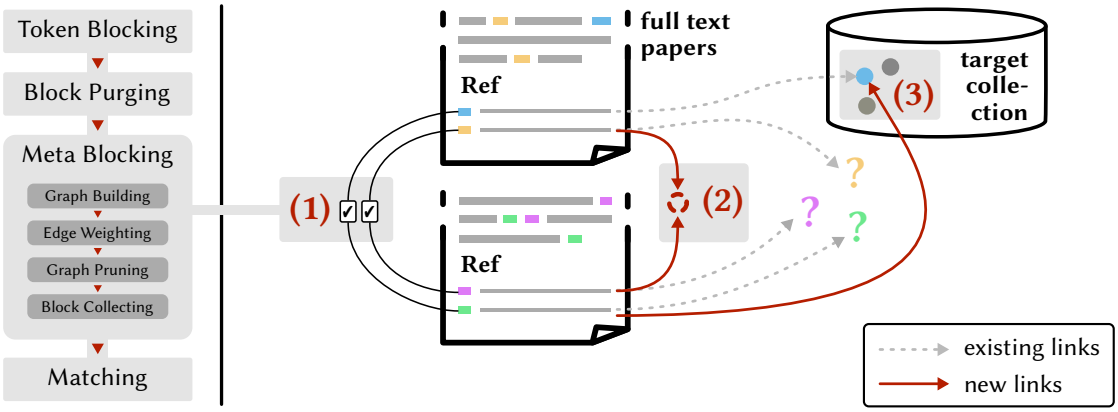

Figure 2: Schematic depiction of the use case. A corpus of full text papers, where some references are already linked to a target collection (blue), and some are not (orange, pink, green). At **(1)** we apply our blocking and matching approach to identify all references that point to the same publication. In doing so, we establish new links in the form of **(2)** bibliographic coupling and **(3)** links to the target collection.

paper records—usually a large metadata set like DBLP[2] or OpenAlex[3], or a set of IDs like DOIs or PMIDs. This means references to literature which is not contained in the target collection, as well as to non-source items [15], cannot be linked (see "?" markers in Figure 2).

Linking references can be seen as a task of entity resolution (ER) [16], which is concerned with identifying entities referring to the same object within or between large data sets. Because the task requires a one-to-one comparison between each of the involved entities, it is inherently of quadratic complexity. To make approaches scalable, entities are assigned into groups of likely matching candidates prior to comparison, a technique called blocking [17]. While blocking-

[2]See https://dblp.org/.
[3]See https://openalex.org/.

based approaches are used in the domain of scholarly data to, for example, identify duplicate paper records [18, 19, 6] (where information such as abstracts are used) and authors [20], they are not utilized for bibliographic references.

We therefore address both of the aforementioned problems with current reference linking approaches, (1) the use of simple matching methods based on title and authors, as well as (2) the reliance on a target collection of paper records, by proposing (1) the use of a blocking and matching process utilizing seven reference fields (title, author, journal, year, etc.) that (2) operates *within* the set of bibliographic references of a corpus, and is thereby *independent* of a target collection of papers (see marker "(1)" in Figure 2).

We showcase the feasibility and benefits of our approach, implementing a pre-processing, blocking, and matching pipeline and evaluating it on a corpus containing 300,000 references. We show that relative to the original data, our approach gives us a **90% increase** in papers linked to the target collection, a **five-fold increase** in bibliographically coupled [21] papers (see marker "(2)" in Figure 2), and a **nine-fold increase** in in-text citation markers covered.[4] The new links are furthermore of high quality (85% F1). This paves the way towards higher quality scholarly data, especially regarding the coverage of so far underrepresented literature and non-source items.

In summary, we make the following contributions.

- We propose a blocking-based approach for matching bibliographic references that is independent of a target collection of paper records.
- We perform a large-scale evaluation showing that our approach results in a manifold increase in high quality reference links.
- We make our data and code publicly available.[5]

2. Related Work

Blocking-based approaches have been used in the domain of scholarly data, though to the best of our knowledge not for bibliographic references. We therefore report on (1) exemplary uses of blocking in the scholarly domain for entities other than references, and (2) approaches to linking bibliographic references using methods other than blocking.

Simonini et al. [18] develop BLAST (Blocking with Loosely-Aware Schema Techniques) which adapts Locality-Sensitive Hashing. Among data sets from other domains, they also evaluate their approach for the task of linking 2,600 DBLP paper records to the ACM[6] and Google Scholar.[7] Sefid [19] proposes several models to match paper records utilizing the papers' title, header, and citation information. The models are evaluated in three scenarios matching 1,000 paper records from CiteSeer^x [22] to IEEE, DBLP, and Web of Science. Lastly, Färber et al. [20] detect duplicates among 243 million author records in the Microsoft Academic Knowledge Graph [23] and evaluate their approach using ORCiD IDs.

[4] With the "coverage" of in-text citation markers we refer to markers associated with linked references, relative to markers belonging to unlinked references.

[5] See https://github.com/IllDepence/ulite2022.

[6] See https://dl.acm.org/.

[7] See https://scholar.google.com/.

Lo et al. [6] introduced the data set S2ORC, which contains 9.6 million open access papers and has recently seen extensive use in area of scholarly document processing. The authors link references to papers within their data set using a heuristic similarity measure based on n-grams and the Jaccard similarity, which only uses the paper title. Using this method, 26 million out of 50 million references (52%) are successfully linked. The authors report that the low number is *"due to large numbers of papers (mostly in the field of physics) for which the bibliography entries are formatted without paper titles."* Saier et al. [7] introduce unarXive, a data set created from papers' LaTeX sources containing over 1 million publications. Bibliographic references in the data set are linked to the Microsoft Academic Graph [24, 25]. The linking procedure is based on string similarity of papers' titles and author information. With this procedure 17 million out of 40 million references (42%) are successfully linked. Lastly, CiteSeer^x [26, 22] in another large data set containing paper records. Similar to S2ORC, references are linked to paper records within the data set itself. In the case of CiteSeer^x the linking is performed through a heuristic assignment based on title and author information. We are not aware of information on the percentage of references that are successfully linked in CiteSeer^x.

3. Approach

Our approach consists of the following three steps: (1) *pre-processing* to convert references into a normalized, structured format, (2) *blocking* to allow us to process large amounts of references, and (3) *matching*. These steps are explained in more detail below.

Pre-processing References as they appear in papers are hard to match for several reasons, such as the variety of citation styles, variants of author names, venue abbreviations, sparsity of information, and typing errors [27] (see Figure 1). To mitigate these issues, we pre-process references in three steps: first, we apply GROBID's [28] reference string parsing module,[8][9] then we expand journal and conference abbreviations, and lastly all strings are lowercased and Unicode normalized. For the abbreviation expansion we use a mapping for 47.6k journal titles provided by JabRef[10] and 2.6k conference titles crawled from various web sources. Following [30] we select seven reference fields for the blocking step: title, author, year, volume, journal, booktitle, and pages.

Blocking Following [31], we build our blocking pipeline from components for (1) block building, (2) block cleaning, and (3) comparison cleaning. As shown in Figure 2, we use token blocking, block purging, and meta-blocking respectively for each of the steps.

Token blocking is chosen for the block building step because it is schema-agnostic and therefore robust against the varying level of information contained in or missing from bibliographic references. In this step, references are assigned to blocks based on all tokens (i.e., words) contained in the identified and normalized reference fields. As a result, references at this point are associated with multiple blocks, which leads to a high level of redundancy.

[8] See https://grobid.readthedocs.io/en/latest/Grobid-service/#apiprocesscitation.
[9] GROBID was chosen according to the results of [29].
[10] See https://github.com/JabRef/abbrv.jabref.org.

Block purging [32] removes oversized blocks based on a comparison cardinality metric, which we determine heuristically and set it to 0.01. Intuitively, the removed blocks originate from common tokens, meaning that matched reference strings within them are highly likely to also share smaller blocks. Purging therefore reduces the number of overall comparisons with minimal effect on the final result quality.

Meta-blocking [33], our comparison cleaning step, reduces unnecessary comparisons within blocks by generating a weighted graph of entities (references in our case) based on their shared blocks, removing edges based on a pruning scheme, and lastly creating a new block collection based on the reduced graph. For both the weighting and the pruning of edges several schemes exist. In Section 4 we describe how we determined the most suitable combination of schemes for our use case. Here, we briefly mention the schemes involved. Available graph weighting schemes include the Common Blocks Scheme (CBS), the Enhanced Common Blocks Scheme (ECBS), the Aggregate Reciprocal Comparisons Scheme (ARCS), and the Jaccard Scheme (JS). For graph pruning, we consider Cardinality Node Pruning (CNP), which relies on cardinality to select the top edges for each node, as well as Weight Edge Pruning (WEP), which removes edges based on their assigned weight.

Matching To determine which references within a block refer to the same publications, we utilize a weighted average of Jaccard similarities across our seven reference fields. Based on [34] as well as preliminary experiments, we set the weights for title, author, journal, booktitle, year, volume, and pages to 8, 6, 5, 5, 3, 3, and 2 respectively, and set the threshold for a match to 0.405.

4. Evaluation

We use a large corpus of scholarly publications to perform two types of evaluations. (1) A large-scale evaluation utilizing the corpus' existing reference links as ground truth, and (2) a manual evaluation to also assess the correctness of newly created reference links. In the following, we describe the data used, evaluations performed, and results obtained.

Data For our evaluation we use the data set unarXive [7]. We chose this data set over similar data sets such as S2ORC [6], because it not only contains paper's full text with annotated in-text citation markers, but also a dedicated database of all raw references in plain text. From unarXive we sample the 300,000 most recent references to conduct our evaluation. The 300,000 references originate from 9,917 papers from the disciplines of physics (7,347), mathematics (1,686), computer science (789), and other STEM fields (95). The publications cited through the references cover publication years from 1743 up to 2020. Four examples of references used in the evaluation are shown in Figure 1.

Large-Scale Evaluation Our large-scale evaluation is performed in two steps. First, we determine the most suitable configuration of graph weighting and pruning scheme for our meta-blocking step, then we apply our pipeline to the evaluation corpus and determine the number of additionally linked entities.

Table 1
Performance of five graph weighting and graph pruning scheme combinations for meta-blocking.

Weighting scheme	Pruning scheme	#Comparisons	#Matches	RR[1] (%)	PC[2] (%)	PQ[3] (%)
CBS[4]	CNP[8]	39,050	3,053	99.96	54.47	7.82
ECBS[5]	CNP	39,050	3,201	99.96	**57.11**	**8.20**
ARCS[6]	CNP	39,050	2,890	99.96	51.56	7.40
ARCS	WEP[9]	24,175	1,285	**99.98**	22.93	5.32
JS[7]	WEP	42,919	2,272	99.96	40.54	5.29

Metrics: [1]Reduction Ratio, [2]Pair Completeness, [3]Pairs Quality
Weighting schemes: [4]Common Blocks Scheme, [5]Enhanced Common Blocks Scheme, [6]Aggregate Reciprocal Comparisons Scheme, [7]Jaccard Scheme
Pruning schemes: [8]Cardinality Node Pruning, [9]Weight Edge Pruning

Table 2
Number of linked papers, references, and in-text citations given in the original corpus and newly created through the application of our approach.

	Linked to target collection		
	#Papers	#Referencecs	#In-text Citations
Given	1,590	13,975	23,707
New	1,443	2,442	7,824
	Linked through bibliographic coupling		
	#Papers	#Referencecs	#In-text Citations
Given	-	-	-
New	8,895	53,940	219,630
	Combined (linked in either way)[1]		
	#Papers	#Referencecs	#In-text Citations
Given	1,590	13,975	23,707
New	8,931	55,197	227,454

[1] Note that the combined entity counts are not simply the sum of the numbers above, because a single entity can be linked in both ways.

To chose a graph weighting and pruning scheme, we use the 13,976 references in our corpus which are are already linked to the target collection as ground truth. Following [33], we select five combinations of schemes to evaluate. The combinations are evaluated using the metrics pair completeness (PC), which expresses the ratio of detected matches with respect to all true matches, pair quality (PQ), which estimates the portion of true matches within all executed comparisons in the block collection, and reduction ratio (RR), which measures the number of unnecessary comparisons that are saved through blocking. Table 1 shows the results of our evaluation. We achieve the best results using ECBS weighting and CNP pruning. Accordingly, we apply our pipeline with this configuration on the full evaluation corpus of 300k references, where our approach performs 496,051 comparisons after blocking and identifies 71,826 matches.

As shown earlier in Figure 2, we can use the matches identified by our pipeline to create two types of new links. First, new links to the target collection, and second, links between references created through bibliographic coupling. New links to the target collection are established whenever a reference with no existing link is matched to a reference with an existing link (see marker "(3)" in Figure 2). In cases where neither of the references in a match have an existing link, we create a bibliographic coupling (see marker "(2)" in Figure 2). In Table 2 we show on the level of papers, references, and in-text citations how many links were already given in our corpus and how many new links we are able to establish. Regarding links to the target collection, we are able to link *1,443 new papers* (90.75% increase) through *2,442 references* (17.47% increase), which are connected to *7,824 in-text citation markers* (33.00% increase). As for bibliographic coupling, we connect *8,895 papers* through *53,940 references* connected to *219,630 in-text citation markers*. Comparing the number of given links to the combined number of new links, we see a 90% increase in papers linked to the target collection, a five-fold increase in bibliographically coupled papers, and a nine-fold increase in in-text citation markers covered.

Manual Evaluation To assess the quality of our newly linked references, we take a random sample of 500 reference comparisons from the matching procedure and manually verify if our approach correctly labeled each pair as a match or non-match. This is done by inspecting both original reference strings (prior to pre-processing) and determining whether they refer to the same publication or not. Because in some disciplines such as physics it is common to see references without a title given, this process involves looking up and verifying publications' details online.[11] Examples of two reference pairs are shown in Figure 1. Comparing our predicted matches with the manually established ground truth, we measure a precision of 93.20% and a recall of 79.34%. Accordingly the F1-score is 85.71%. This shows us that our newly established links are of good quality, suggesting our approach facilitates the creation of more accurate scholarly data and, accordingly, higher quality analyses and downstream applications based scholarly data sets.

5. Discussion and Future Work

To improve the quality of reference linking in large scholarly data sets, we proposed a blocking-based reference linking approach that is independent of a target collection of paper records. In a large-scale evaluation, we first determined the most suitable meta-blocking scheme for our particular application case. Subsequently applying our approach to a corpus of 300,000 references, we saw a manifold increase in linked papers, references, and in-text citation markers. The newly established links are of high precision and have a high recall, which we confirmed through a manual evaluation on a sample of our results. This demonstrates the benefits and quality of our approach.

Key limitations of the work presented are (1) the size and discipline coverage of the evaluation corpus, (2) the usage of a comparatively basic blocking technique, and (3) the lack of a thorough evaluation of time performance.

[11] For further details see https://github.com/IllDepence/ulite2022/tree/master/5_manual_evaluation.

In the future we want to address these points by expanding our work through using more advanced blocking methods such as progressive blocking [35, 36], using larger evaluation corpora such as the whole unarXive data set, including data from more diverse disciplines such as the humanities, and evaluating the time performance of our approach. Because references in our evaluation corpus are linked to in-text citation markers, we furthermore plan to explore application scenarios utilizing the paper full texts.

Author Contributions

Tarek Saier: Conceptualization, Data curation (support), Formal analysis, Investigation (support), Methodology (support), Software (final evaluation), Visualization, Supervision, Writing – original draft (lead), Writing – review & editing. Meng Luan: Data curation, Formal analysis, Investigation, Methodology, Software, Writing – original draft (support). Michael Färber: Supervision, Writing – review & editing.

References

[1] J. E. Hirsch, An index to quantify an individual's scientific research output, Proceedings of the National academy of Sciences 102 (2005) 16569–16572.

[2] C. Chen, CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature, Journal of the American Society for Information Science and Technology 57 (2006) 359–377. doi:10.1002/asi.20317.

[3] A. Elkiss, S. Shen, A. Fader, G. Erkan, D. States, D. Radev, Blind men and elephants: What do citation summaries tell us about a research article?, Journal of the American Society for Information Science and Technology 59 (2008) 51–62.

[4] S. Ma, C. Zhang, X. Liu, A review of citation recommendation: from textual content to enriched context, Scientometrics 122 (2020) 1445–1472.

[5] M. Färber, A. Jatowt, Citation recommendation: approaches and datasets, International Journal on Digital Libraries 21 (2020) 375–405. doi:10.1007/s00799-020-00288-2.

[6] K. Lo, L. L. Wang, M. Neumann, R. Kinney, D. Weld, S2ORC: The Semantic Scholar Open Research Corpus, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 4969–4983.

[7] T. Saier, M. Färber, unarXive: a large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata, Scientometrics (2020). doi:10.1007/s11192-020-03382-z.

[8] T. Saier, M. Färber, T. Tsereteli, Cross-Lingual Citations in English Papers: A Large-Scale Analysis of Prevalence, Formation, and Ramifications, International Journal on Digital Libraries (2021). doi:10.1007/s00799-021-00312-z.

[9] M.-A. Vera-Baceta, M. Thelwall, K. Kousha, Web of Science and Scopus language coverage, Scientometrics 121 (2019) 1803–1813.

[10] X. Liu, X. Chen, CJK Languages or English: Languages Used by Academic Journals in China, Japan, and Korea, Journal of Scholarly Publishing 50 (2019) 201–214.

[11] H. F. Moed, V. Markusova, M. Akoev, Trends in Russian research output indexed in Scopus and Web of Science, Scientometrics 116 (2018) 1153–1180.
[12] O. Moskaleva, M. Akoev, Non-English language publications in Citation Indexes - quantity and quality, in: Proceedings 17th International Conference on Scientometrics & Informetrics, volume 1, Edizioni Efesto, Italy, 2019, pp. 35–46.
[13] G. Colavizza, M. Romanello, Citation Mining of Humanities Journals: The Progress to Date and the Challenges Ahead, Journal of European Periodical Studies 4 (2019) 36–53.
[14] C. Kellsey, J. E. Knievel, Global English in the humanities? A longitudinal citation study of foreign-language use by humanities scholars, College & Research Libraries 65 (2004) 194–204.
[15] P.-S. Chi, Which role do non-source items play in the social sciences? A case study in political science in Germany, Scientometrics 101 (2014) 1195–1213. doi:10.1007/s11192-014-1433-1.
[16] V. Christophides, V. Efthymiou, K. Stefanidis, Entity Resolution in the Web of Data, Synthesis Lectures on the Semantic Web: Theory and Technology 5 (2015) 1–122. doi:10.2200/S00655ED1V01Y201507WBE013.
[17] G. Papadakis, D. Skoutas, E. Thanos, T. Palpanas, Blocking and Filtering Techniques for Entity Resolution: A Survey, ACM Computing Surveys 53 (2020) 31:1–31:42. doi:10.1145/3377455.
[18] G. Simonini, S. Bergamaschi, H. V. Jagadish, Blast: A loosely schema-aware meta-blocking approach for entity resolution, Proc. VLDB Endow. 9 (2016) 1173–1184. doi:10.14778/2994509.2994533.
[19] A. Sefid, Record Linkage Between CiteSeerX and Scholarly Big Datasets, Master's thesis, The Pennsylvania State University, 2019.
[20] M. Färber, L. Ao, The Microsoft Academic Knowledge Graph Enhanced: Author Name Disambiguation, Publication Classification, and Embeddings, Quantitative Science Studies 3 (2022) 51–98. doi:10.1162/qss_a_00183.
[21] K. W. Boyack, R. Klavans, Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?, Journal of the American Society for Information Science and Technology 61 (2010) 2389–2404. doi:10.1002/asi.21419.
[22] J. Wu, K. Kim, C. L. Giles, CiteSeerX: 20 Years of Service to Scholarly Big Data, in: Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse, AIDR '19, 2019. doi:10.1145/3359115.3359119.
[23] M. Färber, The Microsoft Academic Knowledge Graph: A Linked Data Source with 8 Billion Triples of Scholarly Data, in: Proceedings of the 18th International Semantic Web Conference, ISWC'19, 2019, pp. 113–129. doi:10.1007/978-3-030-30796-7_8.
[24] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. P. Hsu, K. Wang, An Overview of Microsoft Academic Service (MAS) and Applications, in: Proceedings of the 24th International Conference on World Wide Web, WWW '15 Companion, ACM, 2015, pp. 243–246. doi:10.1145/2740908.2742839.
[25] K. Wang, Z. Shen, C. Huang, C.-H. Wu, D. Eide, Y. Dong, J. Qian, A. Kanakia, A. Chen, R. Rogahn, A Review of Microsoft Academic Services for Science of Science Studies, Frontiers in Big Data 2 (2019) 45. doi:10.3389/fdata.2019.00045.

[26] J. Wu, K. M. Williams, H.-H. Chen, M. Khabsa, C. Caragea, S. Tuarob, A. G. Ororbia, D. Jordan, P. Mitra, C. L. Giles, CiteSeerX: AI in a Digital Library Search Engine, AI Magazine 36 (2015) 35–48. doi:10.1609/aimag.v36i3.2601.

[27] P. Christen, Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, Springer Science & Business Media, 2012. doi:10.1007/978-3-642-31164-2.

[28] P. Lopez, GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications, in: Research and Advanced Technology for Digital Libraries, 2009, pp. 473–474.

[29] D. Tkaczyk, A. Collins, P. Sheridan, J. Beel, Machine Learning vs. Rules and Out-of-the-Box vs. Retrained: An Evaluation of Open-Source Bibliographic Reference and Citation Parsers, in: Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, JCDL '18, ACM, New York, NY, USA, 2018, pp. 99–108. doi:10.1145/3197026.3197048.

[30] H.-K. Koo, T. Kim, H.-W. Chun, D. Seo, H. Jung, S. Lee, Effects of unpopular citation fields in citation matching performance, in: 2011 International Conference on Information Science and Applications, 2011, pp. 1–7. doi:10.1109/ICISA.2011.5772372.

[31] G. Papadakis, J. Svirsky, A. Gal, T. Palpanas, Comparative analysis of approximate blocking techniques for entity resolution, Proc. VLDB Endow. 9 (2016) 684––695. doi:10.14778/2947618.2947624.

[32] G. Papadakis, E. Ioannou, C. Niederee, P. Fankhauser, Efficient entity resolution for large heterogeneous information spaces, 2011, pp. 535–544. doi:10.1145/1935826.1935903.

[33] G. Papadakis, G. Koutrika, T. Palpanas, W. Nejdl, Meta-blocking: Taking entity resolution to the next level, IEEE Transactions on Knowledge and Data Engineering 26 (2014). doi:10.1109/TKDE.2013.54.

[34] Y. Foufoulas, L. Stamatogiannakis, H. Dimitropoulos, Y. Ioannidis, High-pass text filtering for citation matching, in: Research and Advanced Technology for Digital Libraries, Springer International Publishing, Cham, 2017, pp. 355–366.

[35] G. Simonini, G. Papadakis, T. Palpanas, S. Bergamaschi, Schema-Agnostic Progressive Entity Resolution, IEEE Transactions on Knowledge and Data Engineering 31 (2019) 1208–1221. doi:10.1109/TKDE.2018.2852763.

[36] S. Galhotra, D. Firmani, B. Saha, D. Srivastava, Efficient and effective ER with progressive blocking, The VLDB Journal 30 (2021) 537–557. doi:10.1007/s00778-021-00656-7.