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

\(^1\)We use “link[ing/ed] references” w.r.t. to connections to cited papers rather than in-text citation markers.
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 range, 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 DBLP² or OpenAlex³, 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-

---

²See https://dblp.org/.
³See 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. 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 datasets from other domains, they also evaluate their approach for the task of linking 2,600 DBLP paper records to the ACM6 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 CiteSeerx [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.

With the “coverage” of in-text citation markers we refer to markers associated with linked references, relative to markers belonging to unlinked references.

6See https://dl.acm.org/.
7See 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\textsuperscript{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\textsuperscript{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\textsuperscript{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, 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\textsuperscript{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.

---

\textsuperscript{8}See https://grobid.readthedocs.io/en/latest/Grobid-service/#apiprocesscitation.

\textsuperscript{9}GROBID was chosen according to the results of [29].

\textsuperscript{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.
To choose a graph weighting and pruning scheme, we use the 13,976 references in our corpus which 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.\(^\text{11}\) 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.

\(^\text{11}\)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


