Construction and Application of Citation Linkages Between Papers and Citing Literature: Solving the Bottleneck Problem in Citation Analysis

Authors: Wang Lixue, Wang Lixue

Date: 2022-04-15T10:52:52Z

Abstract

[Purpose/Significance] Researchers require citation linkages that identify the cited papers referenced by citing documents; however, the commonly downloaded data only contains the references listed in citing documents. Only by establishing a mapping between reference entries and cited papers can citation analysis research overcome its bottlenecks and be revitalized.

[Method/Process] This paper proposes a combined approach using DOI matching and multi-field concatenation to locally establish associations between cited papers and reference entries in citing documents, based on the splitting or parsing of downloaded paper data.

[Results/Conclusion] Creating paper citation linkages is a fundamental data processing step that can support the development of various analyses and applications, thereby opening up broad prospects for scientometrics.

Full Text

Mapping between Papers and Citing Papers to Enable Citation Analysis: A Solution to the Bottleneck Problem in Citation Analysis

National Science Library, Chinese Academy of Sciences, Beijing 100190, China

Abstract

[**Purpose/Significance**] Researchers require citation links that identify which cited papers are referenced by citing papers. However, commonly available download data only contains the references listed in citing papers. Only by establishing mappings between reference items and cited papers can citation analysis research break through its bottleneck and gain new vitality.

[**Method/Process**] This paper proposes a method that combines DOI matching and multi-field matching to locally create associations between cited papers and reference entries of citing papers, based on the splitting or parsing of downloaded paper data.

[**Result/Conclusion**] Creating paper citation links is a fundamental data processing step that can support various analyses and applications, opening vast development space for scientometrics.

Keywords: Citation Analysis; Citing Paper; Citation Link; Citation Data; Bibliometrics

Since Garfield established the citation index, citation analysis has achieved comprehensive development in theory, methods, and tools, with numerous applications in revealing literature relationships, paper measurement and evaluation, disciplinary structure analysis, and scientific law exploration. The data foundation of citation analysis is the citation relationship between papers formed by citing papers referencing references, i.e., "cited paper = reference → citing paper." In researchers' contexts, citation relationships generally refer to "cited paper → citing paper" associations, which implicitly assume the "cited paper = reference" mapping relationship. However, commercial databases currently only provide downloadable data in the "cited paper = reference → citing paper" format without including the former mapping relationship. Meanwhile, due to the lack of readily available tools, ordinary researchers cannot easily complete the missing mapping relationships on their own. Consequently, paper citation relationship data manifests as "difficult to integrate online and incomplete offline," creating multiple "breakpoints" in normal citation networks (see the dashed links in Figure 1) that severely hinder citation analysis from developing in deeper and more scientific directions in terms of methodological research, indicator exploration, and law analysis. This represents the "bottleneck" problem in citation analysis research and application.

This paper is a research outcome of the National Social Science Fund project "Research on Joint Ranking Methods for Chinese and Foreign Discipline Journals Across Selection Systems" (19BTQ190).

Author Introduction: Wang Lixue (ORCID: 0000-0001-9108-7671), Associate Research Librarian, EMAIL: hiwanglixue@163.com.

To break through the above bottleneck and create "cited paper → citing paper" associations, the fundamental task is to establish the "cited paper = reference" mapping relationship, i.e., to connect the dashed link marked in Figure 1. Swedish scientometrician Olle Persson once asserted that comparing citing and cited source items would have tremendous potential [1]. We can also foresee that if paper data containing complete citation relationships could be easily obtained, innovative researchers would certainly be able to give citation analysis greater room for development in methodological research, indicator exploration, and application expansion, stimulating strong research vitality and opening up more valuable application prospects. To this end, this study takes paper datasets downloaded from the Web of Science (WoS) platform and their corresponding citing paper datasets (including references) as examples, proposes methods for locally creating associations between cited papers and citing papers, introduces practical use cases in methodological research and evaluation indicator exploration, and further discusses the shortcomings of the proposed scheme and optimization measures.

1.1 Paper Citation Relationship Data

For many years, scholars have conducted numerous studies using paper citation relationships, mainly including using citation frequency (highly cited) for evaluation, conducting citation network analysis with citation associations (co-citation, bibliographic coupling, etc.), and attempting correlation analysis by combining citations with other document features. In related research, the majority directly use "reference → citing paper" data, relatively few use citing paper data, and very few construct and use "cited paper → citing paper" associations. Regarding original citation relationship data, most studies obtain it from databases such as WoS, Scopus, CSCD, CSSCI, and CNKI, which is relatively common. Some researchers have access to underlying citation data, such as SCI CD-ROM data and WoS BP data, but this is not universally applicable.

Research using reference data can be roughly divided into two categories: The first type simply uses citation association data, i.e., the reference fields of citing papers, to statistically analyze authors, journals, years, etc., contained in reference information items, or to conduct co-citation analysis of references and bibliographic coupling analysis of citing papers. However, due to limited information items, these are mostly used as process results. The second type conducts other association analyses on papers with citation relationships, combining references with other internal and external features to analyze references and keywords, cited authors and keywords, cited authors and paper authors, journal mutual citation, reference years and co-citation relationships, and more recently, citation content analysis. These require richer data items but can still be satisfied primarily by single datasets.

Research using citing paper data also has two main types: One type involves using citing paper datasets alone or combining cited paper and citing paper datasets for measurement and analysis at the dataset level, but does not require one-to-one associations between cited papers and citing papers. The other approach obtains citing papers for each cited paper individually, but because it skips the process of mapping cited papers to specific reference entries, it can mostly only be used for small-scale datasets, and the resulting datasets are difficult to generalize for subsequent use.

Among cases that associate cited papers with specific reference entries, Olle

Persson [2] once introduced Bibexcel software's functionality in this regard, using strings combined according to WoS reference styles or DOIs for exact matching and association. Unfortunately, what the software feeds back to users is only the correspondence between cited papers and citing papers, not the specific reference entries. Zhu Qingsong and Leng Fuhai [3] used HistCite software' s "Local Citation Score" (LCS) function to associate dozens of cited papers with reference entries of citing papers. This software is suitable for analyzing single data files and requires considerable subsequent manual operations to export results and enable use. Qin Xiaohui and Le Xiaoqiu [4] mentioned using self-written Java programs to construct forward and backward citation networks centered on single papers, but did not elaborate on the details of association construction.

1.2 Citation Relationship Data Tools

No specialized tools for processing citation relationship data have been seen in the library and information science field, but most bibliometric analysis tools or platforms have certain citation relationship statistics functions. More than ten tools have been frequently mentioned in papers and reports, including Citespace, Bibexcel, VOSviewer, CitNetExplorer, Sci2, SciMAT, HistCite, TDA, SATI, ItgInsight, Bicomb, RefViz, bibliometric.com, and COOC. Investigation reveals that existing tools or platforms cannot effectively support users in building local paper citation databases. In fact, they are mainly oriented toward data statistics applications for ordinary users, with functional focuses not on basic data operations for professional users. HistCite and Bibexcel can support associations between cited papers and citing papers to a certain extent, but the operations are cumbersome and functions are limited. Citespace software' s co-citation analysis actually targets all reference entries within the dataset. It should be noted that various toolkits in R, Python, or other programming environments can also conduct bibliometric analysis, but due to time and energy constraints, this paper has not expanded exploration in this direction.

Bibexcel software [2] uses the "Citations among docs" function to associate cited papers with citing papers through two methods: First, multi-field matching (Make citation links among WoS-records) combines required information from each paper record according to WoS reference field styles for matching in the dataset. Second, DOI matching (Make citation links based on DOI). After running the function, two columns of numbers are obtained: the sequential numbers of citing papers and cited papers in the data file. Using the "Add field to units" function, the ID of citing papers (the UT field in WoS data) and the ID of cited papers can be added to the list respectively, thereby obtaining the citation relationship between citing papers and cited papers. Testing shows that the numbers of citation relationships obtained by Bibexcel' s multi-field matching and DOI matching are inconsistent, and in practical use, the union of the two can be taken. The software is not designed to support users in building local analysis databases and is not quite capable of meeting the needs of creating paper citation relationship databases.

HistCite software has a “Local Citation Score” (LCS) function that can obtain citing paper sets [3]. The software also uses multi-field matching patterns to “exactly match” reference entries of citing papers but does not support matching using DOI fields. Using the export function in the software, records can be sorted by “LCS” and the sequential number of each paper recorded, then clicking the LCS value after each record opens the local citing paper list, which can be exported in CSV format. However, it is still necessary to manually supplement the ID of each cited paper in the data file according to the paper’ s sequential number to obtain the correspondence between cited papers and citing papers. The software can only process one TXT document at a time and requires considerable manual operation, thus it is only suitable for small-data-volume applications.

Operationally, Bibexcel and HistCite are extremely sensitive to author name spelling forms based on multi-field combined exact matching strings, and cannot match if there are slight differences. Bibexcel’ s matching results still require manual association between cited papers and reference entries of citing papers. HistCite’s matching results for citing papers also require manual correspondence according to the sequential numbers of cited papers. Overall, they cannot meet practical needs.

2 Method for Constructing Paper Citation Links

Citation analysis research in theoretical exploration, methodological innovation, and application expansion all relies on the support of basic citation relationship data. To obtain the required paper citation associations, we propose a combined method of “DOI matching” and “multi-field matching” (see Table 1) for paper datasets downloaded from WoS and their corresponding citing paper datasets. This method matches cited papers to reference entries of citing papers (cited paper = reference), thereby enabling the creation of basic citation relationships for local datasets and supporting various database-level analysis and statistical operations.

Table 1. Matching Basis for Journal Paper Citation Link Construction

Matching Basis	DOI Matching	Multi-field Configuration
First author surname	√	M1 M2 M3 M4 M5 M6
Publication year	√	√ √ √ √ √ √ √
Journal name	√	√ √ √ √
Volume	√	√ √ √ √
Issue	√	
Starting page	√	
DOI	√	

Note: The table shows matching fields where √ indicates the field is used. M1-M6 represent different multi-field matching configurations.

DOI matching is not complicated but requires two preprocessing steps: First, character normalization can partially correct character recognition errors introduced during the OCR stage of citation data. Second, to match DOIs contained in references, it is necessary to parse the information in each reference entry to judge the completeness and number of DOIs. In some databases, reference entries are marked with multiple DOIs, and DOI information in such cases cannot be used for citation link creation.

Multi-field matching combines relevant information of cited papers according to the characteristic style of the data source to match with parsed and reconfigured references of citing papers. If there is a need to consider merging multi-source datasets, all reference entries must be precisely parsed and a unified multi-field matching rule defined. This paper adopts M1, M2, and M3 from Table 1. Through exploration, four aspects require further discussion in multi-field matching: paper page numbers, usage scenarios and rationale for first author names and surnames, source journal names, and processing in multi-source data fusion scenarios.

(1) **Paper page numbers**: Page numbers are one of the key pieces of information for effectively distinguishing papers, especially when other matching fields are incomplete. If page number information is missing, paper citation links created through multi-field matching may incorrectly match unrelated papers, and additional verification must be applied in such cases. If page numbers exist, only the starting page is needed.

(2) **Processing of first author names**: Clarifying different spelling forms of the same author' s name has always been a challenge. In actual WoS data, some Chinese scholars' pinyin names have reversed surnames and given names, or two-character names marked only with one initial (e.g., "Qian, Xuesen" marked as "Qian, X."). Foreign scholars mainly have spelling differences due to whether middle names are fully written. Therefore, balancing fault tolerance and accuracy, we use only the author' s surname when other matching fields are complete to create as many paper citation links as possible (as in M1 in Table 1). When volume or issue is missing, we use the author' s name initials (as in M2 and M3 in Table 1) to reduce error probability.

(3) **Source journal names**: Different databases mark journal names slightly differently, some using abbreviations (e.g., WoS data) and others using full names (e.g., Scopus), while domestic Chinese journal names are basically all full names. To handle matching with multi-source data, a journal information table should be created and maintained to facilitate mapping between abbreviations and full names.

(4) **Multi-source data fusion scenarios**: Constructing paper citation links with multi-source data requires identifying and marking common

paper records from each source, then unifying the reference styles from each source and using paper DOIs in combination. It should be noted that WoS database reference entries do not include “issue” information, and its rules cannot distinguish multiple papers that differ only in issue number (e.g., those with DOIs 10.7500/AEPS20170601011 and 10.7500/AEPS20170120004), thus it cannot be used as a unified style in multi-source data scenarios.

After the above processing, paper records can be matched with reference entries, and paper citation links can be created in local databases. To verify actual effectiveness, we randomly selected a dataset from WoS containing 592 citation relationships. The results showed: HistCite obtained 428 relationships, Bibexcel obtained 655 relationships in total, and our method obtained 592 relationships. After manual verification, we found that although HistCite had fewer matching relationships, it had no errors; Bibexcel had 75 incorrectly matched relationships, but the causes of errors were difficult to determine; HistCite had 3 relationships that Bibexcel failed to match; our method obtained 164 more relationships than HistCite and 12 more than Bibexcel.

3 Applications of Paper Citation Links

After creating paper citation links, researchers’ available data expands from isolated paper datasets to associated “cited paper → citing paper” datasets, enabling forward and backward direct and indirect citation analysis along paper citation chains, as well as multi-angle correlation analysis combined with document features. Some potential applications after linking cited papers and citing papers are shown in Table 2. Due to table dimension limitations, only potential applications of two-field associations can be simply displayed, but using more fields is expected to open up more possibilities.

Table 2. Application Potential of Paper-Citing Paper Associations

Analysis Dimension	Cited Paper Features	Citing Paper Features	Potential Applications
Author	(First/Corresponding) author institution	Author self-citation/mutual citation/direct citation	Author citation patterns, author potential knowledge flow
Institution	Institution self-citation/mutual citation/direct citation	Institution citation models, institution potential knowledge flow	Institutional collaboration analysis

Analysis Dimension	Cited Paper Features	Citing Paper Features	Potential Applications
Theme	Author’ s theme, institution’ s theme	Theme association or evolution, topic heat	Disciplinary frontier identification
Discipline	Author’ s discipline, institution’ s discipline	Discipline self-citation/mutual citation/direct citation	Discipline knowledge flow, discipline citation models
Journal	Author’ s journal, institution’ s journal	Journal potential knowledge flow, journal citation patterns	Journal evaluation and mapping

Based on clarified paper citation relationships, we have already carried out several practical applications, such as self-citation statistics under multiple rules, academic influence contribution analysis, co-citation analysis within datasets, and paper evaluation using weighted citations.

3.1 Custom Rule Self-Citation Statistics Paper citation indexing has always been a traditional business of libraries, often providing objective data support for project completion, award applications, and talent selection. When counting paper citation data, it is often necessary to list or exclude self-citation counts separately. However, the definition of self-citation can follow rules such as author self-citation, co-author self-citation, or institutional self-citation, which may not align with database platform rules. Therefore, manual judgment of self-citation counts is often required in practice.

After constructing paper citation links, each citation source can be clearly identified. Combined with data parsing work such as author name normalization, authorship situation clarification, and affiliation splitting, citations from domestic and foreign sources, inside and outside institutions, and within specific author groups can be distinguished, enabling convenient statistics of self-citation counts under various rules. Since the principle of paper self-citation statistics is relatively clear, specific details will not be elaborated here.

3.2 Academic Influence Contribution Analysis Universities and research institutes frequently need to count research papers, which may involve quantity statistics, affiliation signatures, citation counts, etc. More in-depth analysis covers responsible persons and their secondary institutions, citation situations, and academic influence distribution. Previous citation statistics could only use cumulative total citation counts and other statistical values provided by database platforms, such as WoS citation counts and ESI global top 1% discipline indicators frequently used before the “Breaking the Four Only” policy. However, such statistics can only limit the publication year of statistical objects but cannot set the citation year range as needed. For example, to understand the “recent two-year situation” of a unit’ s papers, one can only look at papers published in the recent two years, but it is difficult to analyze the citation situation of the unit’ s historical papers in the recent two years.

Based on localized paper citation relationships, citation data from database platforms can be further limited and operated as needed. We have mainly explored two types of applications: One is counting paper citation situations by citing year to reveal more timely academic influence distribution, such as “the recent two-year citation situation of a unit’ s papers.” The other is limiting the journal scope and document types of citing papers according to ESI statistical rules to count the performance and changing trends of specific units that have entered and not entered the global top 1% disciplines, as well as the contribution and performance characteristics of internal secondary institutions in terms of citations, to assist in research unit disciplinary development planning.

3.3 Co-Citation Analysis Within Datasets In previous years, scholars widely used Citespace software for co-citation analysis. However, the software’ s co-citation results depend on the type of basic data used by the user: If the target paper dataset is used, the results show co-citation relationships among references cited by the paper set; if the citing paper dataset is used, the results show co-citation relationships including target papers and other cited references. Neither of the above two results is the co-citation relationship between target papers that users actually expect.

Therefore, by associating cited papers with reference entries of citing papers, co-citation relationships of specific paper sets can be easily obtained, supporting scholars in conducting co-citation analysis for any custom domain. This enables the analysis of a domain’ s knowledge base and research frontiers using bibliometric methods, and extends co-citation associations to other analysis objects such as scholars, journals, institutions, and themes.

3.4 Paper Evaluation Using Weighted Citations In the field of web search engines, Google quickly won user recognition in its early days with the PageRank algorithm, whose inspiration came from paper citations in bibliometrics. Web pages have no initial weight for reference and have closed-loop hyperlink situations, so PageRank uses random paths to calculate each web page’ s weight. In contrast, paper citations can only develop from new to old unidirectionally, and most journal papers have natural initial weights—journal impact factors.

Under the current “Breaking the Four Only” evaluation orientation, society generally no longer one-sidedly emphasizes citation counts or crudely judges papers by their journals. In this context, while respecting the objective fact that many disciplines have widely recognized excellent journals, we use the citing papers’ journal impact factors as citation weights to replace previous citation counts that did not consider the academic influence of citing sources, thereby obtaining papers' weighted citation counts (details will be discussed in a separate paper). This brings two changes: First, weighted citations add connotation of academic influence from the source, better reflecting papers' actual academic impact to some extent. Second, it is expected to reduce some non-standard citation behaviors by raising the threshold for effective citations. The weighted citation approach has been providing quantitative support for the selection and series report meetings of "Widely Concerned Academic Achievements in Beijing" organized by the Beijing Association for Science and Technology for three years.

4 Problems and Outlook

Based on routinely accessible paper data, this paper proposes a paper citation association method that combines "DOI matching" and "multi-field matching" to accurately create paper citation relationships within local datasets. This will provide important foundational conditions for professionals to conduct methodological research, indicator exploration, and evaluation applications in citation analysis. Previous analyses were constrained by the limited information contained in reference entries, but now they are expanded to citing papers and cited papers with rich internal and external features, opening the door for single or mixed literature feature correlation analysis based on citation chains.

To better use the basic data after creating citation links, we must clearly understand potential problems in the current scheme. First is understanding special cases in the data, such as situations with only volume or issue numbers (WoS references have no issue numbers) but no page numbers, where evaluation between mismatches and missed matches is needed during multi-field matching. WoS downloaded data reference entries sometimes have multiple DOIs marked, which cannot be used as an absolutely reliable matching means. Second, basic data from single sources and multiple sources are two completely different situations: single sources should use their original information for multi-field matching, as using corrected information would lead to mismatches; but multiple sources require using corrected information for multi-field matching. Going further is gradually improving the accuracy of the scheme, roughly in three stages: first, simply splitting fields as they are; second, parsing fields according to rules with certain verification; third, maintaining and using multiple thesauri for information normalization.

Through exploration, HistCite has not been updated for many years and has matching omissions, Bibexcel has matching errors that are difficult to control, and the most suitable scheme is the combined use of multi-field matching and DOI matching. We recommend that relevant personnel gradually pursue improved matching accuracy when conditions permit. If scholars in library and information science can conveniently construct basic paper data with accurate citation relationships for research and exploration, it will undoubtedly open broader research and application fields for scientometrics.

[1] PERSSON O. Exploring the analytical potential of comparing citing and cited source items[J]. Scientometrics, 2006,68(3): 561-572.

[2] PERSSON O,DANELL R,SCHNEIDER JW. How to use Bibexcel for various types of bibliometric analysis[M]. // STRÖM F. Celebrating scholarly communication studies: A festschrift for Olle Persson at his 60th birthday. Leuven: International society for scientometrics and informetrics, 2009: 9-24.

[3] 祝清松, 冷伏海. 基于引文内容分析的高被引论文主题识别研究 [J]. 中国图书馆学报, 2014,40(209): 39-49.

[4] 秦晓慧, 乐小虬. 面向单篇文献引文网络的主题来源与走向追踪 [J]. 现代图书情报技术, 2015 (9): 52-59.