Design and Implementation of Citation Verification and Retrieval Function Based on Institutional Repository: Postprint

Authors: Liu Yanmin, Zhu Zhongming, Zhang Wangqiang

Date: 2023-08-26T00:00:00+00:00

Abstract

[Purpose/Significance] To enhance the efficiency of citation retrieval services, facilitate self-service querying of inclusion and citation reports for researchers, and expand the scientific research evaluation functionalities offered by institutional repositories, a citation verification and citation retrieval function based on the CSpace institutional repository system was developed.

[Method/Process] By investigating the key workflows and detailed issues of manual citation verification and citation retrieval services, the necessity of extending institutional repositories with citation verification and citation retrieval capabilities was established. A functional implementation flowchart was designed, and features were implemented for displaying research outputs, statistics on inclusion types and citation counts, non-self citation counts, citing literature, and export of detailed information from SCI, EI, CSCD, and other databases.

[Results/Conclusion] After testing and evaluation, the citation verification and citation retrieval function based on the institutional repository achieved over 95% accuracy in inclusion and citation retrieval. Compared with manual retrieval, service efficiency was significantly improved, enabling libraries to save approximately 75% of human resources. This effectively realizes the institutional repository's support for citation verification and citation retrieval services.

Full Text

Preamble

Design and Implementation of Cited Reference Retrieval Function Based on Institutional Repository

Liu Yanmin[1], Zhu Zhongming[2], Zhang Wangqiang[2]
[1]Lanzhou University Library, Lanzhou 730000
[2]Lanzhou Library of Chinese Academy of Sciences, Lanzhou 730000

Abstract

[**Purpose/Significance**] To improve the efficiency of citation retrieval services, facilitate researchers' self-service querying of citation reports, and expand the research evaluation service functions of institutional repositories, this study developed a cited reference retrieval function based on the CSpace institutional repository system. [**Method/Process**] Through investigation of key processes and detailed issues in manual citation retrieval services, we propose the necessity of extending institutional repositories with cited reference retrieval functionality, design an implementation flowchart, and realize functions including display of research outputs, indexing type and citation counts, other-citation statistics, citing literature export, and detailed information export for SCI, EI, CSCD, and other databases. [**Result/Conclusion**] Testing and evaluation demonstrate that the institutional repository-based cited reference retrieval function can achieve over 95% accuracy in indexing and citation retrieval. Compared with manual retrieval, service efficiency improves significantly, helping libraries save approximately 75% of human resources. This truly enables institutional repositories to provide effective support for citation retrieval services.

Keywords: cited reference retrieval; institutional repository; function design; citation search
Classification Number: G250.7
DOI: 10.13266/j.issn.0252-3116.2018.12.012

Introduction

Citation retrieval is an information consulting service where retrieval institutions search for papers indexed and cited in domestic and international authoritative databases according to user needs, providing proof of research capability and level. Specifically, through author name, affiliation, journal title and volume/issue, conference information, article title, and other approaches, the service verifies whether papers are indexed by authoritative databases such as SCI (Science Citation Index), SSCI (Social Science Citation Index), A&HCI (Arts & Humanities Citation Index), EI (The Engineering Index), CPCI (Conference Proceedings Citation Index), Chinese Social Sciences Citation Index (CSSCI), and Chinese Science Citation Database (CSCD), and whether they have been cited, ultimately issuing a retrieval certificate report. The indexing and citation report serves as an important reference for project applications and professional title evaluations.

According to our investigation, most libraries currently provide citation retrieval services through a manual process: clients submit retrieval request forms, staff manually search specified databases for the provided papers, manually organize

and statistically process the downloaded data, and finally generate citation reports. Taking Lanzhou University Library as an example, from 2013 to 2016, the library produced an average of 500 citation reports annually. Each report required several hours to complete, with some taking an entire day or even two days, particularly during peak retrieval periods. This work is highly repetitive, inefficient, and labor-intensive, unable to meet all user demands while also affecting service quality. Therefore, using computer programs to complete online the processes of paper list submission, indexing and citation statistics, other-citation counting, and automatic citation report generation has become an inevitable development trend for citation retrieval work.

Current Status of Citation Retrieval Services and the Necessity of Extending Institutional Repository Functions

2.1 Current Status of Citation Retrieval Services

Citation retrieval is a distinctive service project of Chinese libraries. Manual citation retrieval services primarily consist of 10 steps: request form acceptance and confirmation, paper list verification, retrieval task assignment, paper indexing retrieval, paper citation retrieval, retrieval result formatting, retrieval statistics, report writing, user verification of reports, and payment and report collection. This service process is cumbersome and inefficient, with major problems including time-consuming multi-database retrieval, tedious result organization, non-reusable retrieval results, and inability to automatically generate reports.

In terms of machine-assisted citation retrieval services, novelty search institutions have conducted extensive practical research. Tsinghua University was the first to develop a proxy search service system. In 2005, Li Xiaodong and Lu Zhenbo from Peking University Library proposed using tool software to achieve author paper data collection, automatic retrieval, and automatic downloading functions, but did not further design and implement a citation retrieval system. Fan Yafang et al. proposed using Excel's filtering functions and literature management software such as EndNote Web and NoteExpress to assist manual retrieval, enabling batch removal of self-citations, formatted output of other-citation literature lists, and statistics on total other-citation counts, citing journal types, and author numbers. This method can assist manual retrieval and improve service efficiency.

In terms of automated management of citation retrieval business processes, Shi Xiaoqing from Shandong University designed and constructed an efficient library citation retrieval system based on B/S architecture, detailing the planning and design of permissions and main functions for six types of user roles. Users can submit retrieval applications online, retrieval staff can manage reports, and reviewers can audit and archive reports. Xu Shiyan proposed introducing jBPM workflow technology to design a new comprehensive citation retrieval service platform. The advantages of these two systems lie in process automation, but

they fail to automate the key retrieval functions and report generation. Yan Chaobin and Chen Jiayong from Beijing University of Posts and Telecommunications proposed integrating citation retrieval services into the institutional repository ecosystem, independently developing an institutional repository inspired by citation retrieval needs, achieving a precise association mechanism between literature and authors that can accurately link authors' indexed and cited literature lists. However, there are data accuracy issues requiring actual verification in databases, and functions such as self-service citation report export have not been implemented.

The CALIS Technology Center and Peking University Library jointly developed a CALIS paper indexing and citation retrieval system. This system essentially completes 7 of the 10 manual citation retrieval steps, with only request form acceptance and confirmation, preliminary paper list checking, and user report verification still requiring traditional methods. After six rounds of testing and nearly half a year of trial use and improvement, the system basically meets the needs of university libraries, greatly alleviating human resource constraints. Currently, over 150 university libraries have opened trial accounts, including 7 "985" universities, with 19 formal purchasers. However, paper list submission and acceptance require manual completion, reports lack journal impact factors and other evaluation indicators, and Chinese literature retrieval effects are suboptimal.

At the end of 2011, the Institute of Software of the Chinese Academy of Sciences developed a "Citation Report Automatic Generation Prototype System." Based on this prototype, Wang Xueqin et al. optimized data preprocessing functions and algorithms, added retrieval data sources, human-computer interaction modules, and self-citation exclusion functions. After testing and evaluation, the system's work efficiency, accuracy, and stability all achieved expected targets, and it has been deployed at the Documentation and Information Center of the Chinese Academy of Sciences with overall good results. However, it still needs improvement in user concurrency control, user permission management, Chinese and conference paper citation retrieval, and multi-format report generation.

2.2 Necessity of Extending Citation Retrieval Functions in IR

As an important means for many universities and research institutions to preserve and manage research knowledge achievements, institutional repositories (IR) have accumulated considerable academic resources through years of construction and maintenance. How to enhance IR systems' research evaluation services for universities and research institutions has become an important future development issue. The following lists the necessity and favorable conditions for extending citation retrieval functions in IR.

First, IR deposits metadata of a large number of academic achievements by institutional authors, providing a good data foundation for citation retrieval. In the initial construction stage, metadata was primarily imported in batches

through professional platforms, including journal papers and conference papers indexed by English databases such as Web of Science and EI, and Chinese databases including CNKI, CSCD, and CSSCI. IR also provides academic resource automatic collection services through ScienceRouter, supporting the collection of institutional publicly published achievements from large mainstream academic resource databases and sharing data through interfaces. Academic resource databases include IEEE, Springer, PubMed, Elsevier, Google Scholar, and CiteSeerx. IR automatically obtains the latest institutional output data from SciRouter and synchronously updates batch-imported databases. Due to the broad coverage of papers deposited in IR, basically encompassing all institutional academic achievements, the research outputs retrieved for citation retrieval are actually a subset of IR. For citation retrieval users, there is no need to provide literature to be retrieved. IR's standardized processing of imported metadata solves problems such as inconsistent list formats and incomplete data in user-submitted paper lists.

Second, IR's work claim mechanism achieves seamless matching between authors and research achievements. The IR platform establishes an author alias database and author unique identifiers, using machine-automated methods to match possible associations between works and authors, pushing association information to relevant authors for claiming, and saving claim results, thereby achieving accurate association between author information and related work information. In manual citation retrieval processes, staff must verify indexing in designated databases based on paper lists, primarily using literature titles combined with authors, journal names, and conference information. For user-submitted literature to be checked, citation retrieval is conducted using "first author & cited work," with automatic confirmation of erroneous citations in retrieval results. This process is cumbersome and inefficient. In IR, users can self-query all their academic achievements, and citation retrieval staff can avoid repeatedly searching multiple databases.

Third, IR and citation retrieval work complement and promote each other. For IR builders, citation retrieval is no longer a one-time repetitive task. Through citation retrieval work, IR can supplement works not yet collected and urge authors and IR builders to update data. Since the user groups for university library citation retrieval services are basically fixed, there may be clients who commission queries continuously within several years, or the same user updating indexing and citation status of paper lists. Due to the drawbacks of manual library operation processes, report updates are equivalent to re-retrieval, not only increasing repeated retrieval costs but also failing to meet users' needs for rapid report acquisition. The citation retrieval function in IR better solves these problems faced by libraries and users.

Finally, IR can promote comprehensive diversification of citation retrieval services. As one of the important means of scientific and technological evaluation, commonly used indicators for citation retrieval are indexing counts, citation times, and other-citation times. To develop citation retrieval from simple evalua-

tion to comprehensive evaluation requires comprehensive and multi-dimensional evaluation indicators. IR's subject analysis support services provide relatively comprehensive evaluation indicators. Research achievements include whether they are Class A papers, ESI highly cited papers, hot papers, etc. Indexing types include SCIE, SSCI, A&HCI, PubMed, MEDLINE, BIOSIS, CPCI-S, ESI, CPCI-SSH, CSSCI, CSCD, EI, etc. In IR's journal distribution analysis of published papers, journal configuration items include JCR partitions, CAS partitions, JCR impact factors, and journal types. This metadata lays a good foundation for promoting comprehensive diversification of citation retrieval services.

These factors constitute the important basis for extending citation retrieval functions in IR.

IR Citation Retrieval Function Framework and Implementation

3.1 Design Concept and Function Framework

Based on the existing resources and storage structure of the Chinese Academy of Sciences IR system, combined with current automation needs for citation retrieval services, we designed and developed a self-service citation report export function in IR, aiming to provide effective support for research achievement evaluation. IR's harvesting and paper deposit mechanisms and work claim mechanism can display user paper list lists. By extracting main fields needed for research achievements in citation retrieval, fields can be extended according to user needs. IR's indexing category fields for papers are default-collected when submitting journal paper and conference paper metadata, with high attention paid to this metadata collection during resource construction, resulting in good data quality. Moreover, IR performs deduplication processing for literature repeatedly indexed by multiple databases. When importing entries into the database, title + author matching is first conducted for duplication checking. If entries are duplicated, corresponding indexing types are added to the indexing category field. For example, one paper in Lanzhou University IR is simultaneously indexed by SCIE, EI, PubMed, Medline, BIOSIS, and CSCD databases (see Figure 1 [Figure 1: see original paper]), with indexing categories exceeding retrieval requirements. If manually retrieved, five database platforms would need to be searched, and because staff are accustomed to using only one or two retrieval formulas, workload multiplies and omissions are likely to occur. Through deduplication matching, IR provides comprehensive indexing types, improving both retrieval precision and recall rates, and ensuring the feasibility of automated indexing function implementation.

Figure 2 [Figure 2: see original paper] shows the IR citation retrieval function framework, which consists of two modules: the author literature list display module and the citation report export module. These two modules basically implement data preprocessing, indexing and citation retrieval statistics, and

report generation functions, both relying on the complete data structure of the IR system. The underlying data sources of IR include database platforms such as WOS, EI, CSSCI, CSCD, and CNKI, as well as citation frequency interface programs provided by Web of Science and CSCD.

3.2 Key Function Implementation

The IR citation retrieval function module implements the author literature list (i.e., data preprocessing function), paper indexing and citation retrieval function, citing literature collection and other-citation/self-citation distinction function, and citation report export function. The author paper list is shown in Figure 3 [Figure 3: see original paper]. The citation statistics table includes major fields such as JCR impact factor, SCI citation times, CSCD citation times, paper title, author, source journal, and publication time.

The main steps for manually counting total indexing counts and total citation times are: exporting papers to be queried from databases, extracting corresponding fields, filling them into retrieval result forms, and using Excel to calculate total citation times. IR statistics for total citation times only require submitting literature, with background programs automatically calculating total citations, achieving high efficiency and accuracy.

(1) Citing Literature Information Collection and Processing Function Implementation. Citing literature data in citation retrieval reports involves large data volumes. If a paper's citation times are less than 100, retrieving and exporting citing literature is relatively easy. However, for individual papers with citation times reaching several thousand—for example, one paper by Geng Baizong from Lanzhou University's School of Physical Science and Technology has been cited 1,144 times—manual downloading requires 23 page turns due to WOS platform limitations (maximum 50 items per marking). The IR citation retrieval report export function achieves automatic acquisition of citing literature by analyzing the metadata field wos_{{{citing}}_{{url}}} (citing literature URL) in IR database entries, using web information extraction technology to locate web elements describing literature from HTML pages and extract corresponding data content. Using jQuery technology, it obtains the total number of citing literature and metadata fields including title, author, WOS record number, and source journal. To solve the slow speed of exporting citing literature, we designed scheduled tasks to store citing literature information crawled from URLs into IR's self-built database, greatly improving citing literature export efficiency.

(2) Other-Citation and Self-Citation Distinction Function Implementation. In literature citation retrieval services, other-citations are used as the main indicator for evaluating the impact of scientific and technological achievements, obtained by excluding self-citations. Self-citation exclusion methods generally include excluding the author being retrieved, excluding group authors, or excluding all paper authors. The other-citation calculation method used in this

study is strict other-citation, which excludes all paper authors. IR collects citing literature for papers, storing fields including author and author affiliation information. Moreover, citing literature and original literature both come from the Web of Science database platform, where authors are matched using full names. For individual authors with non-standard full names, standardization is performed in the database. The other-citation calculation implementation process is shown in Figure 4 [Figure 4: see original paper].

Citation Retrieval Function Testing and Effect Evaluation

After implementing the IR system's citation retrieval function, we tested and evaluated the function from three aspects: timeliness, stability, and indexing/citation accuracy. The test data selected were from the well-established Lanzhou University IR, covering resources in key disciplines such as physics, chemistry, biology, and economics. The total number of test papers was 12,524, including 7,216 papers jointly indexed by SCIE and EI, 3,805 papers indexed by CSCD, 1,259 papers indexed by CSSCI, 234 papers indexed by CPCI-S, and 10 papers indexed by SSCI. The test involved 356 Lanzhou University scholars. Among them, the highest citation times for SCIE-indexed papers was 1,168, and the highest citation frequency for CSCD was 252.

For Chinese literature (papers indexed by Chinese databases, i.e., CSCD and CSSCI), the system time and indexing/citation accuracy are shown in Table 2 .

Table 2 Chinese Literature Report Export Time and Accuracy

Number of Articles	System Time (Indexing, Citation, Report Export) (s)	Citation Accuracy (%)
20-50	70-90	-
50-100	120-300	-
100-150	300-450	-

For Chinese literature, only indexing types and citation times need to be counted, and report export only requires exporting detailed information for CSCD and CSSCI indexing. Therefore, speed is faster in IR, with indexing details coming from the institutional repository database and citation times obtained through interfaces. For 100 papers, exporting a retrieval report takes only about 30 seconds, saving approximately 5-7 hours compared with manual retrieval, and about 1-2 hours compared with the CALIS citation retrieval system. Indexing and citation accuracy reaches 98%, attributed to strict data control and standardization by work submitters during IR construction. The CALIS citation retrieval system's accuracy for Chinese literature is high when original texts are standardized, but when original texts are non-standard, the system's automatic matching accuracy is only about 90%, with citation retrieval accuracy lower than indexing accuracy, requiring manual confirmation for about 10% of literature.

For English literature report export, more time is required. For SCIE-indexed papers, other-citation times need to be calculated and citing literature exported. The time spent and accuracy for English literature report export are shown in Table 3 .

Table 3 English Literature Report Export Time and Accuracy

Total Citation Times	System Time (Indexing, Citation, Other-Citation, Citing Literature Export) (s)	System Indexing/Citation Accuracy (%)
10-50	30-180	10-45
50-200	200-750	60-180
200-500	750-1500	180-300

As shown in Table 3, English literature requires relatively more time for calculating other-citation times and exporting citing literature. As total citation times increase, time spent also increases. When citation times are around 500, report export takes about 5 minutes, while manual organization requires about 30 hours, and the Chinese Academy of Sciences citation report automatic generation system takes about 7-8 hours. For papers with citation times up to 1,168, system testing shows report export takes only about 10 minutes. However, indexing and citation accuracy decreases as citation times increase because English databases may contain erroneous citations requiring manual verification. The more citing literature, the greater the possibility of errors. During testing of 12,524 papers' indexing and citation reports, the system ran stably.

The CALIS citation retrieval system's accuracy for English databases mainly depends on whether paper lists are correctly standardized. For paper lists with incorrect information or substantial missing information, the system's automatic matching accuracy for English literature indexing is about 80%, with about 17% requiring manual confirmation and about 2% erroneous retrieval. Citation retrieval accuracy requires manual confirmation for 50% of matches.

Compared with manual retrieval, IR's citation retrieval function greatly shortens retrieval time, improves retrieval efficiency, and saves substantial labor. It can achieve 95% indexing and citation accuracy, meeting peak-period user retrieval needs in universities. Compared with the CALIS citation retrieval system and the Chinese Academy of Sciences citation retrieval system, it streamlines the processes of submitting request forms and retrieval staff searching multiple databases. Users only need to log into the institutional repository, query papers requiring retrieval, and click submit to complete report export. The institutional repository's citation retrieval function has broad future expansion space, allowing users to export needed indicators such as journal partitions and H-index according to requirements, providing possibilities for comprehensive and multi-dimensional research evaluation.

The citation retrieval function developed in this study completely depends on IR. Therefore, it is necessary to further improve and solve issues regarding resource quantity and metadata quality in IR. To address data quality problems in IR, prevention must start from the source: IR builders should use multiple retrieval methods to search databases under the guidance of professional retrieval staff to prevent omissions and avoid inaccurate citing literature retrieval results due to non-standard citations; imported data in IR should be standardized to promptly correct erroneous information in the database; when institutional users claim works, they should promptly verify and update detailed literature information; real-time resource updates should improve resource deposit efficiency and metadata quality through the iSwitch interface. The system needs to further improve automation functions for citation retrieval business processes to better serve research evaluation. We hope that the citation retrieval function developed in this study will be widely applied in university libraries and provide sustainable construction for IR.

References

[1] Ma Fangzhen, Li Feng, Ji Fan, et al. Testing and application effect evaluation of CALIS citation retrieval system[J]. Journal of Academic Libraries, 2016, 34(2): 97-102.

[2] Ma Fangzhen. Discussion on requirements analysis and design key points of citation retrieval system[J]. Journal of Academic Libraries, 2015, 33(4): 80-84, 121.

[3] Zhan Yuhua, Cheng Aiping, Qian Junwen, et al. Development and application of proxy search service system[J]. Library and Information Service, 2005(11): 75-77, 55.

[4] Li Xiaodong, Lu Zhenbo. Design and implementation of paper citation retrieval tool software[J]. Journal of Academic Libraries, 2005(1): 49-50, 62.

[5] Fan Yafang, Chen Kai. Using Excel and EndNote Web to improve paper citation retrieval work efficiency[J]. Library Journal, 2013, 32(1): 32-34, 60.

[6] Zhang Xuejuan, Fan Yafang. Application of NoteExpress in paper citation retrieval work[J]. Information Research, 2017(6): 45-49.

[7] Fan Yafang. Practice and application of using literature management software to improve paper citation retrieval work efficiency[J]. Library Work in Colleges and Universities, 2017, 37(2): 63-66.

[8] Shi Xiaoqing, Wei Jiangxing. Design and implementation of online citation retrieval and novelty search system based on JSP—taking Shandong University (Weihai) as an example[J]. Modern Information, 2014, 34(3): 131-134, 138.

[9] Shi Xiaoqing. Design and implementation of university library citation retrieval system[D]. Jinan: Shandong University, 2014.

[10] Xu Shiyan. Design of citation retrieval comprehensive service platform based on jBPM[J]. New Century Library, 2015(11): 52-56.

[11] Hou Ruifang, Chen Jiayong, Zhou Jie. Construction and consideration of citation retrieval service optimization system[J]. Library Development, 2015(4): 75-79.

[12] Yan Chaobin, Chen Jiayong, Hou Ruifang, et al. University institutional repository construction driven by citation retrieval service support needs[J]. New Technology of Library and Information Service, 2015(5): 94-100.

[13] Wang Xueqin, Hao Dan, Zheng Fei, et al. Applied practice research on "citation report automatic generation system"[J]. Library and Information Service, 2014, 58(16): 131-137.

[14] Hao Dan. Research and implementation of data quality control in citation retrieval[D]. Xi'an: Xidian University, 2012.

[15] Liu Wei, Zhu Zhongming, Zhang Wangqiang, et al. Research and implementation of author identification and work claim mechanism in institutional repository[J]. New Technology of Library and Information Service, 2014(3): 8-13.

Note: Figure translations are in progress. See original paper for figures.