An Empirical Study of Research Topic Innovation Based on Self-citation Networks and Main Path Analysis: Postprint

Authors: Wei Ruibin

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

Abstract

[Purpose/Significance] This study explores how to quickly identify relevant papers from a large corpus and investigates methods for paper topic innovation, providing reference for researchers in reading and utilizing journal articles. [Methods/Process] Based on the conceptual definitions of paper topic innovation and self-citation networks, this paper proposes ideas, methods, and tools for researching paper topic innovation. An empirical study is conducted using papers from Indiana University, USA in the field of Library and Information Science as a case study. [Results/Conclusion] Using main path analysis can quickly identify papers connected through citations from self-citation networks, upon which topic innovation analysis can be conducted.

Full Text

Preamble

Vol. 62 No. 3, February 2018, ChinaXiv Cooperative Journal
Empirical Research on Paper Topic Innovation Based on Self-citation Network and Main Path Analysis
Wei Ruibin
School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu 233030

Abstract

[**Purpose/Significance**] This study explores methods for rapidly identifying relevant papers from large corpora and investigates approaches to analyzing paper topic innovation, providing references for researchers reading and utilizing journal articles. [**Method/Process**] Building upon definitions of paper topic

innovation and self-citation networks, this paper proposes research ideas, methods, and tools for studying paper topic innovation. An empirical study was conducted using papers from Indiana University in the field of library and information science. [**Result/Conclusion**] Main path analysis can quickly identify papers connected through citation relationships within self-citation networks, enabling subsequent analysis of topic innovation.

Keywords: topic innovation; main path analysis; self-citation network; research topics
Classification Number: G250
DOI: 10.13266/j.issn.0252-3116.2018.03.008

Innovation is the fundamental basis for the scientific community to gain social recognition. Treating innovation as a behavioral norm of the scientific community requires that "scientific research results should always be novel. A study that does not add new content to what is already fully understood and comprehended contributes nothing to science." Researchers can obtain recognition and rewards from the scientific community and society through knowledge production, and innovation has become the fundamental basis for scientists to gain social recognition [1]. Journal papers represent an important manifestation of researchers' innovative achievements. The innovative aspects can be understood through the content of journal papers. Hu Yingkui et al. [2], Sheng Jie [3], and Zhu Daming [4] explored specific methods for evaluating paper innovativeness from the perspective of academic journal editors, as well as four basic elements for assessing the innovativeness of scientific journal papers. Cao Yan et al. [5] used the Delphi method to construct an evaluation index system for nursing paper innovativeness. These studies provide relatively specific and scientific references for evaluating the innovativeness of individual papers. However, for already published papers, these methods have significant limitations for subsequent researchers seeking to quickly discover innovations in their research content. This study attempts to conduct exploratory research on paper topic innovation by combining main path analysis with content analysis. Due to the relative nature of innovation, this study confines the relevant research scope within the research output of a single institution.

2 Paper Topic Innovation, Self-citation Networks, and Main Path

2.1 Paper Topic Innovation

Paper innovation can be examined from different perspectives, such as research topic, research method, or research conclusion. This study focuses on topic innovation in journal papers. When determining whether a paper's topic is innovative, the conclusion is typically reached through comparison with existing research results. Judging paper topic innovation can be divided into three levels: first, whether there is innovation; second, in which aspects innovation exists; and third, the degree of innovation. This study addresses the first and second levels,

namely, whether the paper's topic possesses innovativeness.

Paper topic innovation is the result of comparing and analyzing paper topics within a specific dataset. Many factors must be considered when judging paper topic innovation: Time factor—when determining paper topic innovation, a paper is typically compared with previous papers, with the analysis based on professional terms reflecting its topic. Scope factor—for example, using a scholar's journal papers as a dataset allows exploration of individual research innovation; using an institution's research output in a specific field as a dataset allows study of innovation within that dataset scope. This study's research objects are primarily confined within a single institution. Institutional paper topic innovation can be judged from two perspectives: one is comparison among researchers within the institution, and the other is comparison with research results outside the institution in a broader context. This study focuses on the first type of comparison. Judging paper topic innovation is typically based on professional terms reflecting the research topic, such as Paper A's research topic being co-word analysis and Paper B's research topic being information retrieval. These professional terms can be derived from the paper's title, abstract, keywords, or full text, as in T. Amjad et al. [6] who categorized authors and journals into three research topics: multimedia retrieval, medical information retrieval, and database and query processing.

2.2 Self-citation Network

Through paper citation relationships, connections between different research topics can be discovered, providing another angle for judging paper topic innovation. I. Hellsten et al. [7] proposed author self-citation networks and pointed out that author self-citation networks can better identify authors' new research topics. J. Y. Lee [8] argued that author self-citation networks can successfully identify an author's core papers and leading achievements. The author believes that self-citation networks can be divided into different levels. Author self-citation networks [9] refer to citation networks formed by citation relationships among papers written by an author and his/her collaborators. Institutional self-citation networks are citation networks formed when researchers from the same institution cite their own literature or the work of other researchers within the same institution. In this study's institutional self-citation network, nodes represent papers published by institutional researchers, and connections between nodes represent citation relationships. Self-citation networks can also include more macro-level regional and national self-citation networks.

Typically, due to spatial and other factors, a researcher has relatively frequent academic exchanges with other researchers in the same institution and discipline, has a better understanding of each other's research topics, and has greater potential for collaboration. Institutional self-citation networks, on the one hand, reflect the dissemination process of knowledge among researchers within the same institution during scientific research, and on the other hand, objectively reflect changes in research topics of researchers within the same institution.

2.3 Main Path Analysis

Main path analysis was first proposed by N. Hummon and P. Dereian [10] from the perspective of network connectivity. Its primary goal is to identify series of literature with maximum connectivity in citation networks to outline the development trends of research fields and major literature, figures, and events in the evolution of fields. W. Goffman [11] and M. Jahn et al. [12] demonstrated through research on the mast cell research field that a discipline is defined by a few extremely important events or people in its historical development. This conclusion provided theoretical support for the emergence of main path methods in citation networks.

The theoretical premise of main path analysis is to view citation networks as channel systems for delivering knowledge and information. If knowledge circulates through citation relationships, then a citation relationship that participates in paths among many papers is more important than another citation relationship that rarely participates in such paths. The most important citation relationships form one or more main paths, which may constitute the skeleton of a research tradition [13]. W. Goffman [11] and M. Jahn et al. [12] showed that if a paper can integrate knowledge from previous papers and make substantial contributions to new knowledge growth, it is likely to be heavily cited and may make subsequent citation of earlier papers somewhat redundant. Therefore, such papers become important hubs in the channel system, through which large amounts of knowledge and information flow [13]. Han Yi and Jin Bihui [14] deeply analyzed the background, basic connotations, and algorithmic implementation of citation network main path analysis, and summarized the main problems in theoretical and applied research. Wei Ling and Fang Shu [15] found that scholars' existing corrections and extensions to main path methods have focused on three aspects: selection principles for main paths, determination of search starting points, and weight settings for arcs.

From the research of Song Ge [16], Han Yi [17], and Zhu Qingsong [18], main path analysis methods can help researchers quickly identify important research results from citation networks and intuitively reflect citation and cited relationships among research results. This study uses main path analysis to connect papers through citation relationships while incorporating temporal factors to analyze topic innovation in these papers.

3 Research Ideas, Tools, and Methods

3.1 Research Ideas

3.1.1 Data Collection Use an institution's name as the search object to obtain its paper data from specific citation databases. Due to limitations in data format and processing tools, this study uses journal paper data collected from the Web of Knowledge platform.

3.1.2 Data Processing Data collected from Web of Knowledge can be used

with HistCite's GraphMaker function to quickly generate corresponding self-citation networks. Self-citation network data can be directly saved as .net documents. Meanwhile, HistCite software can automatically statistics such as local citations and global citations for each paper, providing data support for subsequent analysis. .net documents can be directly imported into Pajek software for further processing. In Pajek, network metrics such as degree centrality for each paper can be obtained, and visual citation network graphs can be generated.

3.1.3 Data Analysis By combining citation network graphs generated by Pajek with relevant data for each paper, the position of each paper in the entire citation network can be further analyzed. Analysis of original paper information can determine its research topic. Combined with the temporal attributes of papers, institutional topic innovation can be analyzed.

3.2 Research Methods and Tools

This study applies main path analysis (MPA) to research self-citation networks. Main path analysis is a special technique for analyzing temporal flows. Pajek version 4.06 differs functionally from previous versions, with the new version providing different main path analysis methods for users to select, such as: Network→Acyclic Network→Create Weighted Network + Vector→Traversal Weights→Search Path Count (SPC)/Search Path Link Count (SPLC)/Search Path Node Count (SPNP); Network→Acyclic Network→Create (Sub)Network→Critical Path Method→CPM. Wei Ling and Fang Shu [11] provided specific explanations for these different methods in their research. This study uses the second method to find main paths in citation networks.

4 Empirical Research

4.1 Data Acquisition

This study uses the Web of Science Core Collection as the data source, selecting papers published by Indiana University scholars in the field of information science and library science from 1986 to 2017. The final retrieval yielded 731 records. The specific search formula is as follows: Address: (Indiana University) Refined by: Web of Science Categories: (INFORMATION SCIENCE LIBRARY SCIENCE) Timespan: 1986-2017. Indexes: SCI-EXPANDED, CPCI-S, CPCI-SSH, CCR-EXPANDED, IC.

As shown in [Figure 1: see original paper], from 1990 to 1996, the institution's publication quantity showed a rapid growth trend. Beginning in 1997, there was a decline, but overall it remained between 20-40 papers, reaching 50 papers in a few years. This reflects that Indiana University's paper output in information science and library science is relatively stable.

This study first uses HistCite as the research tool to automatically generate citation networks for the 731 papers and saves the citation network raw data as

.net documents, which are then imported into Pajek for subsequent processing.

4.2 Data Analysis

4.2.1 Citation Network Overall Analysis Degree refers to the number of connections a node has, which is a discrete attribute. In citation networks, it represents the number of times a paper cites other papers or is cited by other papers. Pajek's partition function can calculate the degree of each node (including the number of papers it cites in the network and the number of times it is cited by other papers), or calculate them separately. The entire citation network has 731 nodes, 497 connections, and a density of 0.0009, with an average degree of 1.3598. These data reflect that citation relationships among papers in this network are not particularly close.

As shown in , there are 382 nodes with degree 0, accounting for 52.26% of the total, meaning more than half of the papers have no citation or being-cited relationships with other papers. There are 133 papers that have only one citation or being-cited relationship with other papers. This reflects that Indiana University scholars' citation behavior in information science and library science exhibits concentration and dispersion.

Based on nodes' indegree centrality and outdegree centrality, this study classifies the 238 nodes into three types:

(1) **Knowledge Output Papers**. These nodes have outdegree centrality greater than indegree centrality (difference ≥ 3), accounting for about 14% of the total. These papers represent early research results on a particular topic within the institution and play a leading role in subsequent related research. For example, Paper 41 [19] is cited by 9 other papers in the network but does not cite any other papers. This paper proposes a new model for explaining decision support system functional performance.

(2) **Knowledge Absorption Papers**. These nodes have indegree centrality greater than outdegree centrality (difference ≥ 3), accounting for about 13% of the total. Such papers are typically newer stage research results formed on the basis of reviewing many previous studies. For example, Paper 148 [20] cites 10 papers in the network but is only cited by one other paper. This paper proposes an integrated theoretical model for collaborative research based on social presence theory and other foundations.

(3) **Knowledge Balanced Papers**. These nodes have roughly equal indegree and outdegree centrality (difference ≤ 2), accounting for about 73% of the total. These papers represent intermediate-stage research results within the institution. For example, Paper 126 [21] cites 3 papers in the network and is cited by 5 other papers. This paper further improves and expands on the media synchronicity theory (MST) proposed by the author in 1999.

From the perspective of the entire citation network, when citation frequencies

are similar, knowledge output papers have higher innovativeness, followed by knowledge absorption papers, and then knowledge balanced papers. When comparing paper innovativeness based on this standard, it is necessary to select similar papers within the same time window. For example, if knowledge output papers were published earlier, their high citations may result from cumulative time advantages. Comparing knowledge output papers with knowledge absorption papers based solely on citation frequency would be unreasonable.

4.2.2 Main Path Analysis When the number of nodes is small, a paper's position and role in the citation network can be judged by observing its position in the network structure. However, when the number of nodes is large, quickly identifying important papers becomes more difficult. This study first uses Pajek to process the 238 papers, with results shown in [Figure 3: see original paper] (Chinese content in the figure was added manually). Based on paper content, this study divides the journal papers in [Figure 3: see original paper] into the following six aspects:

(1) **Electronic Journals and Academic Communication**. Four papers in this area were published between 1996-2000, with S. P. Harter as the representative figure. Papers 50 and 159 studied the impact of electronic journals on academic communication from the perspective of journal article references. Papers 197 and 227 studied policies and practices in electronic journal publishing in academic communication. Viewed from the entire citation network of this dataset, Papers 50, 159, 197, and 227 were cited 2, 3, 6, and 2 times respectively. This reflects that the institution has a series of related achievements in this field, but this method only presents a small portion of the papers. Additionally, Papers 159 and 197 have global citations of 53 and 72 respectively, reflecting that these two papers have also gained recognition from academic peers outside the institution.

(2) **Web Citations**. From 2001-2008, web citations became a research topic of interest for some scholars at Indiana University, with B. Cronin and L. Vaughan as representative figures. As shown in [Figure 3: see original paper], Paper 244 occupies a relatively important position in the network, playing a significant leading role for other related research in the network. B. Cronin [22] argued that web-based citation analysis brings new opportunities to the bibliometrics field. This paper serves as both a summarizing continuation of related research by peers within the institution and a foundation for subsequent related research, playing a connecting role in the network structure. B. Cronin [23] used the concept of symbolic capital in the paper title but actually conducted correlation analysis on 25 scholars using three indicators: citation counts, web click rates, and media mention rates. From the perspective of paper citations, this paper also had significant influence on subsequent related research.

(3) **Research Collaboration**. From 2003-2005, using bibliometric informa-

tion to study research collaboration became a research topic, with B. Cronin as the representative figure. For example, B. Cronin and D. Shaw used traditional bibliometric methods to study research collaboration in psychology and philosophy fields, collaboration patterns in 20th-century chemistry, and the impact of research collaboration on academic writing. V. Larivière et al. [26] studied the impact of team size on academic influence through three indicators: number of authors, number of addresses, and number of countries, finding that larger team size and broader author distribution correlate with more citations.

(4) **Academic Impact Evaluation**. From 2005-2009, scholars represented by L. Meho used traditional bibliometric indicators and methods to study the academic impact of researchers and institutions. For example, L. Meho studied research output of LIS researchers and institutions, analyzed the academic impact of information scientists using the h-index, and conducted studies on 25 LIS researchers using three data sources: Web of Science, Scopus, and Google Scholar. L. Vaughan and D. Shaw [24] conducted related research on web citation data from four disciplines and citation data from Web of Science (WoS) and other sources. These studies represent current "altmetrics" research in library and information science. According to Yu Houqiang and Qiu Junping [25], the concept of altmetrics was proposed by Priem in 2010. B. Cronin and others began related research in 2001, demonstrating the innovativeness and foresight of their research.

(5) **Academic Network Analysis and Application**. Since 2009, research on various academic networks has become a topic for the institution, with representatives including Ding Ying and Yan Lejia. They conducted related research on various academic networks from a network perspective, such as analyzing collaboration networks using network centrality indicators, analyzing author influence using PageRank algorithms, analyzing academic organization interactions based on citation and collaboration networks, measuring scholar reputation and influence using weighted PageRank algorithms, and conducting research on author citation networks using topic-based PageRank algorithms. Compared with traditional bibliometric research, this type of research fully integrates research results from social network analysis, complex networks, and computer science with literature networks, representing a relatively new research direction in bibliometrics.

(6) **Other Aspects**. In the entire network, there are some special research topics. For example, J. A. Pratt et al. [27] used co-citation analysis, multidimensional scaling analysis, and principal component analysis to study the intellectual structure of the information management systems field using data from 25 academic journals. In terms of content, J. A. Pratt et al. referenced Paper 487 in data selection. Although the two research topics differ to some extent, they share certain commonalities in data sources.

C. R. Sugimoto and B. Cronin [28] conducted "bibliometric profiling" of six outstanding bibliometricians through author style, academic output efficiency and patterns, and collaboration patterns. C. R. Sugimoto et al. [29] quantitatively analyzed the relationship between age and research output, collaboration, and influence among over 1,000 scholars in sociology, economics, and political science. These achievements are based on original paper information and citation information to conduct related research on different questions.

Overall, the first five research themes were relatively stable over certain periods, focusing on in-depth innovation of a research topic. As time changes, research themes continuously shift, reflecting innovation in research breadth. As shown in [Figure 3: see original paper], Papers 50, 159, 197, and 227 were published between 1996-2000, reflecting researchers' sustained attention to this theme. Papers 388 and 527 both studied scholar influence but employed different research methods, demonstrating that while research themes remained stable, there was innovation in research content. Paper 244 [30] is a research achievement on web citation analysis that connects the two research themes of electronic journals and academic communication and web citations through citation relationships, reflecting certain associations between the two themes.

Analysis of paper topic innovation relies on interpretation of original paper information. For a small number of papers, manual processing is feasible, but for large-scale processing, more effective automated and intelligent processing methods are needed. Manual methods have strong subjectivity, and using a single word to summarize paper content cannot guarantee comprehensiveness and accuracy. These aspects require further improvement and refinement in future research.

References

[1] Han Yu, Zhao Xuewen, Li Zhengfeng. Analysis of basic research innovation concepts and reflections on related issues [J]. China Basic Science, 2001(3): 35-40.
[2] Hu Yingkui, Luo Min, Wang Xiuling. Methods for academic journal editors to evaluate paper innovativeness in initial review [J]. Acta Editologica, 2012(4): 353-355.
[3] Sheng Jie. Journal editors' grasp of scientific paper innovativeness [J]. Acta Editologica, 2011(3): 215-217.
[4] Zhu Daming. Four basic elements for evaluating scientific journal paper innovativeness [J]. Science and Technology Management Research, 2011(9): 199-201.
[5] Cao Yan, Zhu Ruifang, Han Shifan. Constructing an evaluation index system for nursing paper innovativeness using the Delphi method [J]. Nursing Research, 2017(17): 2101-2103.
[6] AMJAD T, DING Y, DAUDA A, et al. Topic-based heterogeneous rank [J]. Scientometrics, 2015, 104(1): 1-22.
[7] HELLSTEN I, LAMBIOTTE R, SCHARNEHORST A, et al. Self-citations,

co-authorships and keywords: a new approach to scientists' field mobility? [J]. Scientometrics, 2007, 72(3): 469-489.
[8] LEE JY. Exploring a researcher's personal research history through self-citation network and citation identity [J]. Journal of the Korean Society for Information Management, 2012, 29(1): 157-175.
[9] Wei Ruibin. Mining scholar research topics based on self-citation network and content analysis [J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(6): 635-645.
[10] HUMMON N, DEREIAN P. Connectivity in a citation network: the development of DNA theory [J]. Social networks, 1989, 11(1): 39-63.
[11] GOFFMAN W. Mathematical approach to the spread of scientific ideas—the history of mast cell research [J]. Nature, 1966, 212(5061): 449-452.
[12] JAHN M. Changes with growth of the scientific literature of two biomedical specialties [D]. Philadelphia: Drexel University, 1972.
[13] Nei, Batagelj. Spider: Social Network Analysis Techniques [M]. Translated by Lin Feng. Beijing: World Book Publishing Company, 2012: 328-331.
[14] Han Yi, Jin Bihui. A new perspective on citation network structure analysis based on connectivity: main path analysis [J]. Studies in Science of Science, 2012(11): 1634-1640.
[15] Wei Ling, Fang Shu. Review and prospect of citation network main path research [J]. Information Theory and Practice, 2016(9): 128-133.
[16] Song Ge. Study on the diffusion process of academic innovation [J]. Journal of Library Science in China, 2015(1): 62-75.
[17] Han Yi, Tong Ying, Xia Hui. Comparative study of main path methods and highly-cited paper methods for identifying field evolution structures [J]. Library and Information Service, 2013, 57(3): 11-16.
[18] Zhu Qingsong, Leng Fuhai. Topic evolution analysis based on citation main path literature co-citation [J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(5): 498-506.
[19] DENNIS AR, WIXOM BH, VANDENBERG RJ. Understanding fit and appropriation effects in group support systems via meta-analysis [J]. MIS quarterly, 2008, 32(3): 575-600.
[20] BROWN SA. Predicting collaboration technology use: integrating technology adoption and collaboration research [J]. Journal of management information systems, 2010, 27(2): 9-54.
[21] DENNIS AR, FULLER RM, VALACICH JS. Media, tasks, and communication processes: a theory of media synchronicity [J]. MIS quarterly, 2008, 32(3): 575-600.
[22] CRONIN B. Bibliometrics and beyond: some thoughts on Web-based citation analysis [J]. Journal of information science, 2001, 27(1): 1-7.
[23] CRONIN B, SHAW D. Banking on different forms of symbolic capital [J]. Journal of the American Society for Information Science & Technology, 2010, 53(14): 1267-1270.
[24] VAUGHAN L, SHAW D. Web citation data for impact assessment: a comparison of four science disciplines [J]. Journal of the American Society for Information Science & Technology, 2005, 56(10): 1075-1087.

[25] Qiu Junping, Yu Houqiang. The proposal process and research progress of altmetrics [J]. Library and Information Service, 2013, 57(19): 5-12.
[26] LARIVIERE V, GINGRAS Y, SUGIMOTO CR, et al. Team size matters: collaboration and scientific impacts since 1900 [J]. Journal of the Association for Information Science & Technology, 2015, 66(7): 1323-1332.
[27] PRATT JA, HAUSER K, SUGIMOTO CR. Defining the intellectual structure of information systems and related college of business disciplines: a bibliometric analysis [J]. MIS quarterly, 2001, 25(2): 167-193.
[28] SUGIMOTO CR, CRONIN B. Biobibliometric profiling: an examination of multifaceted approaches to scholarship [J]. Journal of the American Society for Information Science & Technology, 2014, 63(3): 450-468.
[29] SUGIMOTO CR, SUGIMOTO TJ, TSOU A, et al. Age stratification and cohort effects in scholarly communication: a study of social sciences [J]. Scientometrics, 2016, 109(2): 997-1016.
[30] CRONIN B. Bibliometrics and beyond: some thoughts on Web-based citation analysis [J]. Journal of information science, 2001, 27(1): 1-7.

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