Digital Humanities-Driven Core Author Discovery and Topic Mining: Postprint

Authors: Wu Shuai, Yang Xiuzhang, REN Tianshu, Liu Jianyi

Date: 2025-08-14T14:35:40+00:00

Abstract

[Purpose / Significance] Driven by digital humanities, this study conducts statistical analysis from a macro perspective on the four-decade achievements of the *Journal of Redology* from a researcher's viewpoint, and integrates data mining techniques from a micro perspective to analyze potential research areas in Redology, thereby better promoting the development of Redology studies. [Method / Process] The analysis primarily employs bibliometric statistics and topic mining methods. First, core research authors in Redology over the past forty years are identified through bibliometric statistics, followed by topic mining of the achievements in the Redology field. The research hotspots of Redology over the forty-year period are investigated from two aspects: core author discovery and topic evolution analysis. [Results / Conclusion] Redology research as a whole can be divided into four domains, namely character relationship studies, social institution studies, contemporary academic discourse, and version conjectures on *Dream of the Red Chamber*.

Full Text

Introduction

Driven by digital humanities, this study statistically analyzes the achievements of *Studies on "A Dream of Red Mansions"* over its forty-year publication history from a researcher's perspective at the macro level, while integrating data mining techniques at the micro level to analyze potential research fields in Redology, thereby promoting the development of Redology research. As the foremost of China's Four Great Classical Novels, *A Dream of Red Mansions* represents the pinnacle of ancient Chinese chapter-based novels and has been adapted into television dramas nine times, serving as an important basis for studying ancient Chinese literature [**?**]. With the development of art and culture, Redology research and literary criticism have emerged in recent years in a "hundred flowers blooming, hundred schools contending" manner [**?**].

Studies on "A Dream of Red Mansions", included in the Chinese Social Sciences Citation Index (CSSCI), plays a guiding role in its discipline and represents academic achievements with strong scholarly value, novel research, and high creativity. As the only Redology journal selected for the CSSCI database, *Studies on "A Dream of Red Mansions"* has published numerous papers over four decades, serving as the primary exchange platform for Redology research [**?**]. Since its inception, the journal has maintained a rigorous academic attitude, balancing scholarly value with accessibility, achieving high academic standards while remaining engaging and readable, thus earning favor among Redology researchers and literature enthusiasts. As an important carrier for communication and dissemination of Redology research, the journal has effectively promoted the field's development. With internet technology applied to journal publishing, research achievements have proliferated as scholars express diverse opinions, studying *A Dream of Red Mansions* from different angles using various methods, gradually making Redology a comprehensive and integrated discipline.

However, Redology researchers often hold different interpretations of *A Dream of Red Mansions*, making it difficult to accurately reflect contemporary research themes. To address this limitation, this paper examines 5,582 journal papers from *Studies on "A Dream of Red Mansions"* collected by CNKI between April 2, 1979 and April 2, 2019, employing bibliometric statistics and topic mining methods. First, we identify core authors in four decades of Redology research through bibliometric analysis, then conduct topic mining on Redology achievements. We explore research hotspots in forty years of Redology from two perspectives: core author discovery and topic evolution analysis.

1 Related Research

Studies on "A Dream of Red Mansions" serves as the main academic journal for *A Dream of Red Mansions* research, publishing content related to ideological studies, artistic value, historical materials, Redology research, author history, and cultural artifact verification, enjoying high academic reputation in Redology and classical literature research. The journal represents the stage-specific professional level and academic quality of *A Dream of Red Mansions* research, providing important reference and research value.

1.1 Traditional Literature Research Status

In the big data era, data mining technology has developed rapidly, yielding numerous academic research achievements. Currently, relatively few studies domestically and internationally have used data mining [**?**] and machine learning [**?**] algorithms to deeply mine journal literature. Pei Jie [**?**] employed traditional bibliometric methods to explore typical characteristics and development patterns in Japanese translation studies of *A Dream of Red Mansions*. Zhang Qing-shan et al. [**?**] used traditional bibliometric methods to clarify the disciplinary nature of Redology, define its scope, and construct its framework.

Gao Huai-sheng [?] applied traditional bibliometric methods to review Redology's development history, finding that each stage of Redology's development relies on literature. Sun Wei-ke et al. [?] systematically reviewed 2017 Redology research achievements, finding that Redology emphasizes the integration of multiple methods, effectively promoting field integration and interdisciplinary cross-fertilization.

Traditional literature research tends to focus on original reading, expert lectures, core literature reading, and forum participation, with core literature reading being the most common approach. Core literature reading typically uses keyword searches and download volume filtering, representing a relatively single method that insufficiently reveals deep-level Redology research themes and lacks data mining methods to study hot topics and temporal development in *Studies on "A Dream of Red Mansions"*.

1.2 Topic Mining Research Status

As big data technology advances, increasing numbers of scholars recognize the importance of data's potential value, dedicating themselves to combining data mining or machine learning methods to derive valuable conclusions from massive literature data. Shen Lin [?] summarized text content to systematically establish furniture types and combination patterns in *A Dream of Red Mansions*. Wu Di et al. [?] deeply excavated literature materials related to *A Dream of Red Mansions* collected in *Xiangyan Congshu* to further understand the dissemination of Redology research. Chen Xiao [?] studied the image world of *A Dream of Red Mansions* in the Qing Dynasty, exploring connections between text and images to broaden Redology data mining categories. Cai Yong-ming et al. [?] proposed a CA-LDA model for Chinese short text topic analysis, increasing the probability of grouping words with identical collocation relationships into the same topic and providing new research methods for short text literature data. Wu Shuai et al. [?] used data measurement and social network analysis methods to explore library and information science development. Yang Xiu-zhang et al. [?, ?] employed bibliometric and social network analysis methods to review literature development in the Qingshui River Basin and used composite indices and knowledge graphs to discover core authors. Applying data mining methods to deeply explore literature data can reveal data's potential value. This paper uses bibliometric statistics and topic mining methods to investigate forty years of *Studies on "A Dream of Red Mansions"*, reflecting to some extent the stage-specific professional level and academic quality of *A Dream of Red Mansions* research.

2 Research Framework

This study analyzes 5,582 papers published in *Studies on "A Dream of Red Mansions"* over four decades to mine high-citation papers, core authors, main research institutions, and core topics. The specific analysis process consists of four steps, as shown in Figure 1 [Figure 1: see original paper].

First, we use Python's Selenium module to custom-crawl *Studies on "A Dream of Red Mansions"* papers collected by CNKI and save them as CSV files. Second, we preprocess the data, including data cleaning, relationship extraction, and outlier handling, saving the processed data as CSV files. Third, we identify core authors by combining citation counts and publication volume, primarily using Price's Law to preselect core author candidates and composite indices to select core authors. Fourth, we conduct topic mining on forty years of journal papers, including temporal topic evolution analysis, co-word network analysis, and social network analysis.

2.1 Data Collection

This study aims to deeply mine core authors and core topics of *Studies on "A Dream of Red Mansions"* papers collected by CNKI. Using Python's Selenium module, we crawled journal papers from *Studies on "A Dream of Red Mansions"* published between April 2, 1979 and April 2, 2019. The crawled fields mainly include: paper title, author, publication date, citation count, download count, keywords, and abstract.

2.2 Data Cleaning

Some data information in *Studies on "A Dream of Red Mansions"* papers collected by CNKI is incomplete, requiring preprocessing to standardize data formats. The preprocessing mainly includes: data cleaning, outlier detection and handling, and related numerical processing.

3 Core Author Group Discovery

While big data provides massive amounts of diversified information, it also brings the problem of information overload, particularly prominent in academic research. As online submission replaces traditional methods, academic achievements grow rapidly, making accurate identification of core authors increasingly difficult. Core authors constitute the solid foundation of disciplinary research [**?**], determining research direction and academic achievement quality. Traditional identification methods rely solely on publication volume while ignoring paper quality. To address this, this paper proposes a method for identifying core authors in *Studies on "A Dream of Red Mansions"* based on Price's Law and composite indices. First, we identify core author candidates based on Price's Law, then select core authors through a composite index combining publication volume and citation count.

3.1 Price's Law Analysis

This study combines first authors' publication volume and citation count from *Studies on "A Dream of Red Mansions"* to screen core author candidates, using Price's Law to preselect candidates. The specific procedures are as follows.

First, we determine the minimum citation count. The most highly cited paper in *Studies on "A Dream of Red Mansions"* is Wang Jinbo's 2010 publication "The First Neglected Complete English Translation of *A Dream of Red Mansions* (120 Chapters)—An Introduction to Father Bonsall's English Translation," cited 71 times, denoted as Nc_{max}. Using Price's minimum citation formula (1), we calculate the minimum citation count, denoted as M_c. Authors with cumulative citations reaching or exceeding 7 times qualify as core author candidates.

Second, we determine the minimum publication volume. The most prolific author in *Studies on "A Dream of Red Mansions"* is Feng Qiyong, with 124 publications, denoted as N_{pmax}. Using Price's minimum publication formula (2), we calculate the minimum publication volume, denoted as M_p. Authors with 9 or more publications qualify as core author candidates.

Third, we confirm core author candidates. We screen authors from *Studies on "A Dream of Red Mansions"* meeting formula (1) or formula (2), deduplicate the results, and ultimately select 363 qualified core author candidates who published 2,957 papers, accounting for 52.97% of the total papers collected in *Studies on "A Dream of Red Mansions"*, with total citations of 10,324, representing 66.41% of the total citation count.

3.2 Comprehensive Index Selection

From the 363 core author candidates identified by Price's Law, we set a comprehensive index threshold of 2 to select 35 core authors from *Studies on "A Dream of Red Mansions"*. The specific steps are as follows.

First, we determine average publication volume. The total publication volume of the 363 core author candidates from *Studies on "A Dream of Red Mansions"* is denoted as X_{total}; the total number of candidates is denoted as n. Using the average publication formula (3), we calculate the average publication volume, denoted as $\bar{x}$.

Second, we determine average citation count. The total citation count of papers by the 363 core author candidates is denoted as Y_{total}; the number of candidates is denoted as n. Using the average citation formula (4), we calculate the average citation count, denoted as $\bar{y}$.

Third, we apply comprehensive index selection. With $\bar{x}$ as the average publication volume and $\bar{y}$ as the average citation count of core author candidates, we use the comprehensive index formula (5) to calculate each candidate's score, denoted as $score_i$. In the calculation, x_i represents the total publications of candidate i, and y_i represents their total citations. Setting the comprehensive index threshold at 2, we select 35 core authors from *Studies on "A Dream of Red Mansions"*.

$$score_i = \frac{x_i}{\bar{x}} + \frac{y_i}{\bar{y}}$$

Table 1 presents the 35 core authors of *Studies on "A Dream of Red Mansions"* selected through the comprehensive index. The top core author is Feng Qiyong with 124 publications, average citations per paper of 1.89, and a comprehensive index of 11.72; his most highly cited work is "Interpreting *A Dream of Red Mansions*" with 14 citations. The second core author is Hong Tao with 25 publications, average citations per paper of 12.92, and a comprehensive index of 7.21; his most highly cited work is "*A Dream of Red Mansions* English Translation and East-West Culture and Language" with 58 citations. The third core author is Hu Wenbin with 50 publications, average citations per paper of 4.18, and a comprehensive index of 6.74; his most highly cited work is "*A Dream of Red Mansions* and Chinese Name Culture" with 60 citations.

Analyzing average citations per paper, eight core authors from *Studies on "A Dream of Red Mansions"* achieve 7 or more citations per paper. Hong Tao ranks first with 12.92 citations per paper, followed by Wang Jinbo (10.42), Liu Yongliang (9.33), Rao Daoqing (7.88), Duan Jiangli (7.63), Chen Weizhao (7.59), Yu Xiaohong (7.36), and Mei Xinlin (7.00).

4 Literature Topic Mining

As the core content of journal papers, keywords can roughly reflect research themes, methods, and hot topics. Conducting topic mining on keywords from *Studies on "A Dream of Red Mansions"* can clarify the journal's main research directions, methods, and hot themes. This paper's literature topic mining includes temporal topic evolution analysis, co-word network analysis, and social network analysis.

4.1 Temporal Topic Evolution Analysis

CiteSpace temporal sequence topic evolution analysis primarily uses temporal development as the thematic axis. Based on 5,582 papers collected over forty years, we generated the temporal topic evolution analysis shown in Figure 2 [Figure 2: see original paper]. Each node represents a topic, and connections between nodes indicate co-occurrence relationships. The timeline spans from 1979 to 2019.

By examining the temporal distribution of word frequency and combining keywords with high change rates, we identify emerging terms and track frequency trends to confirm frontier topics and development trends in *Studies on "A Dream of Red Mansions"*. Core themes include "Liu Xinwu," "Daiyu," "Yihong Gongzi," "version," "author," "tombstone," and "social scientist."

Overall, Chinese Redology research on *A Dream of Red Mansions* over four decades has evolved from points to lines to planes, involving not only deep literature mining but also archaeological research and film/television production. This enables more objective and accurate restoration of the original work's themes, understanding of its historical context, and grasp of the author's cre-

ative tendencies, providing theoretical foundations for further Redology development.

4.2 Co-word Network Analysis

Using Python, we constructed a keyword co-occurrence matrix for 5,582 papers published in *Studies on "A Dream of Red Mansions"* since its inception. When two keywords appear together in the same paper, they are considered co-occurrent and build a relationship edge with weight +1; otherwise, no relationship edge exists with weight 0. The specific co-occurrence matrix construction formula is shown in calculation formula (6).

$$score_i = \begin{cases} 1 & \text{if keywords co-occur in same paper} \\ 0 & \text{if keywords do not co-occur in same paper} \end{cases}$$

In co-occurrence matrix analysis, the frequency of two keywords appearing together is recorded as word frequency. Higher frequency indicates closer keyword relationships and stronger association with research content; zero co-occurrence indicates no relationship between keywords.

To better reflect core Redology research content through the co-occurrence matrix, we removed the following keywords: "Dream of Red Mansions" (our research theme), "Cao Xueqin" (the author), "Studies on A Dream of Red Mansions" (our target journal), and "chapter novel" (the work's form). Based on keyword co-occurrence analysis of *Studies on "A Dream of Red Mansions"* papers, we obtained the high-frequency co-occurrence word list shown in Table 2 .

The top 20 co-occurrences are: "Baoyu" and "Daiyu" (582 times), "Jia Baoyu" and "Yihong" (369 times), "Daiyu" and "Baochai" (184 times), "Baoyu" and "Baochai" (178 times), "Baoyu" and "Fengjie" (121 times), "Baoyu" and "Jiamu" (102 times), "Daiyu" and "Fengjie" (88 times), "Baoyu" and "Jiafu" (83 times), "Baoyu" and "Qingwen" (83 times), "Daiyu" and "Jiamu" (77 times), "Jiaxu version" and "Gengchen version" (76 times), "Grand View Garden" and "Baoyu" (70 times), "Jiamu" and "Fengjie" (66 times), "manuscript" and "version" (65 times), "Jiafu" and "Daiyu" (65 times), "Baoyu" and "Jia Zheng" (62 times), "Baoyu" and "Gengchen version" (55 times), "Mr." and "Redology" (53 times), "Baoyu" and "Lady Wang" (52 times), and "Jiamu" and "Lady Wang" (51 times).

The high-frequency co-occurrence words indicate that *Studies on "A Dream of Red Mansions"* research primarily divides into two categories: character relationship studies and *A Dream of Red Mansions* version studies. Character relationships center on "Baoyu" and "Daiyu," with other characters associated with them. Version studies mainly focus on investigating the "Jiaxu version" and "Gengchen version."

4.3 Social Network Analysis

Social network algorithms are near-clustering algorithms that can identify strong and weak relationship networks, visually representing relationships through knowledge graphs. Nodes represent relationship points, and edges represent relationships between nodes. Social network algorithms cluster closely related nodes into similar regions and diffuse sparsely related nodes to the periphery, intuitively discovering core relationship points.

Due to numerous scattered nodes with weight coefficients of 1, 2, and 3 affecting overall network effectiveness, we applied Price's Law for node screening as shown in calculation formula (7).

$$M_f = \sqrt{N_{fmax}}$$

where M_f represents the minimum frequency of high-frequency co-occurrent words, and N_{fmax} represents the maximum frequency of high-frequency co-occurrent words in *Studies on "A Dream of Red Mansions"* papers according to Price's Law. Based on Price's Law, high-frequency co-occurrent words must have frequency greater than or equal to 19.

Therefore, we set the co-occurrence threshold at 19 for social network analysis of relevant high-frequency words, generating the thematic keyword co-occurrence knowledge graph shown in Figure 3 [Figure 3: see original paper]. The graph contains 54 core nodes generating 108 relationship edges. The network's modularity coefficient is 0.549, meeting modularity requirements.

The keyword relationship graph shows that *Studies on "A Dream of Red Mansions"* research divides into four modules: character relationship studies, social system studies, current academic discussions, and *A Dream of Red Mansions* version speculation. Character relationship studies focus on "Baoyu" and "Daiyu," with other characters associated with them; Granny Liu's relationships are relatively simple. Social system studies center on "feudal society," reflecting the creative social context of *A Dream of Red Mansions*. Current academic discussions feature Mr. Feng Qiyong as a Redology research representative, who published 124 papers in the journal with his representative work "Interpreting *A Dream of Red Mansions*." Version speculation primarily focuses on *Zhiyanzhai's Re-annotation of The Story of the Stone* (also known as *The Story of the Stone*), collected by Xu Ye (courtesy name Songge), a top scholar and Grand Secretary in the late Qing Dynasty, though some scholars also examine the "Jiaxu version," "Gengchen version," "Chengjia version," and "Chengyi version."

Conclusion

As *Studies on "A Dream of Red Mansions"* serves as Redology's main academic journal, publishing numerous papers over four decades as the primary exchange carrier for Redology research, researchers often base their in-depth studies on

personal understanding of the content, yielding relatively singular results. This necessitates full consideration of the work's essence and core scholars' cognitive perspectives. This paper employs bibliometric statistics and topic mining methods, first identifying core authors in four decades of Redology research, then conducting topic mining on Redology achievements. From core author discovery and topic evolution analysis, we find that forty years of Redology research focuses on character relationships, social systems, current academic discussions, and version speculation. These findings primarily determine the core research themes and development history of *Studies on "A Dream of Red Mansions"*, providing editorial standards and thematic suggestions for the journal's editorial board and offering better academic direction recommendations for Redology researchers.

References

[1] Wen Qingxin. As a literary phenomenon: The productive critical reception of modern "Dream of Red Mansions-ization" [J]. *Chinese Literature Research*, 2021(2): 155-162.

[2] Zhao Jianzhong. Reflections on the "author" of *A Dream of Red Mansions* and "Cao Studies" on the centennial of "New Redology" [J]. *Studies on Ming-Qing Fiction*, 2021(1): 4-24.

[3] Wang Hui. Laying foundations and upholding integrity—Advancing with determination: A summary of the "40th Anniversary Symposium of the Redology Research Institute and the Launch of *Studies on 'A Dream of Red Mansions'*" [J]. *Studies on "A Dream of Red Mansions"*, 2020(1): 8-16.

[4] Yang Xiuzhang, Wu Shuai, Xia Huan, et al. Analysis of China's film industry in 2019 from a big data perspective [J]. *Film Literature*, 2020(23).

[5] Wu Shuai. Identification of frontier topics in scientific research [D]. Guiyang: Guizhou University of Finance and Economics, 2021.

[6] Pei Jie. Dream crossing to Japan [D]. Shanghai: Shanghai International Studies University, 2020.

[7] Zhang Qingshan, Qiao Fujin, Miao Huaiming, et al. Discussion on Redology philology [J]. *Journal of China University of Mining and Technology (Social Sciences Edition)*, 2016, 18(5): 89-96.

[8] Gao Huaisheng. Academic summary of the "High-end Forum on Philological Studies of *A Dream of Red Mansions*: Historical Review and Future Prospects" [J]. *Journal of Henan Institute of Education (Philosophy and Social Sciences Edition)*, 2016, 35(3): 3-11.

[9] Sun Weike, He Weiguo, Hu Qing, et al. 2017 annual research report on Chinese Redology development [C]// *2017 Annual Report on Chinese Art Development Research*, 2018: 363-385.

[10] Shen Lin. Research on furniture categories in Cheng Jia's version of *A Dream of Red Mansions* [D]. Changsha: Central South University of Forestry and Technology, 2021.

[11] Wu Di, Wu Jiaru. Textual research on *A Dream of Red Mansions* materials collected in *Xiangyan Congshu* [J]. *Studies on "A Dream of Red Mansions"*, 2017(6): 175-189.

[12] Chen Xiao. The image world of *A Dream of Red Mansions* in the Qing Dynasty [D]. Hangzhou: China Academy of Art, 2012.

[13] Cai Yongming, Chang Qing. Chinese short text topic analysis based on co-word network LDA model [J]. *Journal of the China Society for Scientific and Technical Information*, 2018, 37(3): 305-317.

[14] Wu Shuai, Ren Tianshu, Liu Jianyi, et al. Exploring library and information science development based on data measurement and social network analysis [J]. *Information Research*, 2022(1): 28-40.

[15] Yang Xiuzhang, Wu Shuai, Xia Huan, et al. Exploring Qingshui River Basin culture based on bibliometrics and social networks [J]. *Modern Computer*, 2019(35): 19-26, 37.

[16] Yang Xiuzhang. Research on bibliometric analysis and knowledge graph of Shui nationality literature [J]. *Modern Computer (Professional Edition)*, 2019(1): 25-32.

[17] Yang Xiuzhang, Xia Huan, Yu Xiaomin, et al. Analysis of core author groups in Shui nationality literature based on composite index and knowledge graph [J]. *Computer Era*, 2019(4): 13-17.

Research on Core Author Discovery and Topic Mining Driven by Digital Humanities

Wu Shuai[1,2], Yang Xiuzhang[2], Ren Tianshu[2], Liu Jianyi[2]

([1] College of Information Management, Nanjing Agricultural University, Nanjing 210003, China; [2] School of Information, Guizhou University of Finance and Economics, Guiyang 550025, China)

Abstract: [Purpose/significance] Driven by digital humanities, this study statistically analyzes the achievements of *Studies on "A Dream of Red Mansions"* over its forty-year publication history from a researcher's perspective at the macro level, while integrating data mining techniques at the micro level to analyze potential research fields in Redology, thereby promoting the development of Redology research. [Method/process] The methods of measurement statistics and topic mining were used in turn for analysis. First, the core authors of Redology research over the past four decades were counted. Then, topic mining of the achievements in the field of Redology was carried out. The research hotspots of Redology research over the past four decades were investigated from the two

aspects of core author discovery and topic evolution. [Result/conclusion] The research on Redology can be divided into four areas as a whole: the study of character relationships, the study of social systems, current academic discussion, and speculation on the version of *A Dream of Red Mansions*.

Keywords: *Studies on "A Dream of Red Mansions"*; Topic evolution; Digital humanities; Topic mining; Core author

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