Postprint: Investigating Historical Root Literature in the Digital Library Field Based on RPYS i/o

Authors: Wu Chuang, Xie Fuxiu, Wang Chunlei, Wanguo Liu, Sun Bo

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

Abstract

[Purpose/Significance] Exploring the historical roots and evolution of a discipline or research field holds significant importance for its construction and development.

[Method/Process] Utilizing the visual online tool RPYS i/o, we conducted both standard RPYS (Reference Publication Year Spectroscopy) and multi-dimensional RPYS analyses to identify literature that has profoundly influenced the origin and evolution of the digital library domain.

[Results/Conclusion] The standard RPYS analysis provided by this tool can accurately identify classic literature related to the field' s origins, while the multi-dimensional RPYS analysis can further reveal literature that has made enduring contributions throughout the origin process of this research field.

Full Text

Preamble

Volume 62, Issue 5, March 2018

Exploring Historical Root Literature in the Digital Library Field Based on RPYsi/o

Wu Chuang, Xie Fuxiu, Wang Chunlei, Liu Wanguo, Sun Bo

Northeast Normal University Library, Changchun 130024

Abstract

[**Purpose/Significance**] Exploring the historical roots and evolution of a discipline or research field is crucial for its development. [**Method/Process**] This study employs the visualization online tool RPYsi/o to conduct both standard Reference Publication Year Spectroscopy (RPY S) and multi-dimensional RPY S analyses, identifying literature that has significantly influenced the origin and evolution of the digital library field. [**Result/Conclusion**] The standard RPY S analysis provided by the tool can accurately identify classic literature related to the field' s origins, while multi-dimensional RPY S analysis can also reveal literature that has made enduring contributions during the field' s formative period.

Keywords: RPYsi/o; digital library; standard RPY S; multi-dimensional RPY S; historical roots

Classification Number: G253.1

DOI: 10.13266/j.issn.0252-3116.2018.05.010

1. Introduction

The term "digital library" emerged in the 1990s. In 1993, the National Science Foundation (NSF), Defense Advanced Research Projects Agency (DARPA), and National Aeronautics and Space Administration (NASA) jointly launched the Digital Library Initiative (DLI). In 1994, the Library of Congress announced a major investment to build a national digital library, an action that propelled digital library construction worldwide. Digital libraries were also previously called electronic libraries or virtual libraries, though a unified definition of the concept has remained elusive. Over more than two decades of development, digital libraries have brought not only technological transformations but also revolutionary changes in service concepts and methods.

Existing studies on the origins and development of digital libraries have employed systematic review methods based on manual reading, which are inefficient and prone to subjectivity, making them ill-suited for handling massive scholarly resources. Both domestic and international researchers have also used bibliometric methods such as co-word and co-citation analysis. For instance, Yang Guoli, Yan Weidong, Yang Jiulong, Hong Lingzi, and L. Godeaux have all employed these methods to analyze research status and hotspots in the digital library field. However, these studies have not specifically examined literature predating the field' s emergence—that is, questions of origin.

Based on this gap, this paper utilizes the emerging bibliometric analysis tool RPYsi/o to explore the historical roots and evolutionary processes of the digital library field, identifying from massive literature collections the classic works potentially related to the field' s origins. This is significant for clarifying the

conceptual connotations of digital libraries and for the further development of the field, and may provide guidance for other researchers seeking historical root literature in different research areas.

---

2. Principles and Features of RPYsi/o

In 1964, with funding from the U.S. Air Force Office of Scientific Research (AFOSR), E. Garfield, I. H. Sher, and R. J. Torpie discussed models and methods for using citation data to explore the historical roots of scientific fields. In 2003, E. Garfield, A. I. Pudovkin, and V. S. Istomin developed HistCite, a citation mapping analysis software that can quickly depict the developmental history of a disciplinary field and identify its important literature. At the 14th International Conference on Scientometrics and Informetrics (ISSI 2013) in 2013, W. Marx and L. Bornmann first proposed "Reference Publication Year Spectroscopy" (RPY S) as a new method for investigating the historical roots of disciplinary fields. Chinese researcher Li Xin summarized RPY S as "a two-dimensional distribution map that takes the publication years of all references cited by all documents in a field as the x-axis and the total citation frequency of all references for each reference publication year (RPY) as the y-axis." RPY S can serve as a supplement to the HistCite method in terms of algorithm and visualization.

Current RPY S researchers have developed two software packages: (1) In 2014, L. Leydesdorff of the University of Amsterdam developed RPY S.exe, freely available to researchers at http://www.leydesdorff.net/software/rpys/. (2) A. Thor of Leipzig University of Applied Sciences developed CRExplorer.exe, available at http://andreas-thor.github.io/cr-e/. Compared to RPY S.exe, CRExplorer.exe includes a data "disambiguation" function that can identify variants of cited references, integrating data that are in fact the same reference but with non-standard formatting.

RPYsi/o is an online tool developed in 2016 by J. A. Comins of Virginia Tech Applied Research Corporation (VTARC) and L. Leydesdorff of the University of Amsterdam. It can perform two types of RPY S analyses: (1) **Standard RPY S Analysis**: This approach operates from the perspective of references, positing that among all references published before a research field' s emergence, a few documents will have citation frequencies far exceeding those of other documents published in the same or adjacent years. These documents are likely classic works that played important roles in the field' s origin and evolution, and they must appear at peak points on the map. Therefore, by analyzing peaks in citation frequencies before a field' s emergence in the reference publication year spectroscopy, we can explore its historical root literature. (2) **Multi-dimensional RPY S Analysis**: This method performs a standard RPY S analysis for each year' s references, calculating the deviation of each RPY' s total citation frequency from the median of total citation frequencies for the

previous year, two years prior, the current year, the following year, and two years after. Using rank transformation, deviation values are sorted—larger deviations receive higher rank values—which are then converted into visual heat values. Greater heat values are represented by darker colors. Thus, darker colors in the heat map indicate larger deviations, showing that an RPY' s citation frequency is significantly higher than in surrounding years. Multi-dimensional RPY S analysis maps can display the annual citation heat and dynamic changes of references over time, helping to identify references with long-term contributions to a discipline or research field.

The differences between RPYsi/o and analysis software like RPY S.exe and CRExplorer.exe are: (1) It is web-based with simple operation and good interactivity; (2) It can perform both standard and multi-dimensional RPY S analyses; (3) Based on DOI and Google search engines, it provides links to access identified classic literature. The main limitation of the current RPYsi/o version is that it can only analyze datasets up to 15MB and publication years from 1900-1999. However, compared to existing software, its ability to perform multi-dimensional RPY S analysis is its greatest advantage, and providing DOI links to access classic literature is another distinctive feature.

Regarding the application effectiveness of RPY S, more than ten foreign studies have used RPY S.exe or CRExplorer.exe to explore historical roots in various fields, including the Higgs boson, graphene and solar cells, the "Darwin's finches" legend in biology, GPS, and climate change. In China, Li Xin, Lu Wei, and Li Xuhui first used RPY S in 2016 to study the historical origins of health information literacy, and in 2017 further explored RPY S analysis using citation analysis and sentiment analysis as examples. These studies demonstrate that RPY S can reveal influential literature in a field and even identify "sleeping beauties"—works that were once neglected. Only a few foreign studies have used the RPYsi/o online tool to explore historical roots, including the developers' investigations into basal cell carcinoma in biomedicine and the journal *Philosophy of Science*, as well as Dalian University professor Hou Jianhua' s study on the origins of citation analysis. No Chinese-language studies have been reported.

---

3. Exploring Historical Root Literature in the Digital Library Field

3.1 Data Source

We selected the Web of Science (WoS) Core Collection as our data source, specifically including: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, and CPCI-SSH. The search strategy was: Topic = "electronic librar" *OR* " *digital librar*" OR "virtual librar* "; time span from 1985 to present; retrieval date: August 31, 2017; document type: article. After refining, we obtained 3,621 papers related to digital libraries. We exported records in" Full Record and Cited References" format

(500 records maximum per export), merged multiple txt files into one data file, and renamed it data.txt (13MB) for final analysis.

3.2 Data Import

The RPYsi/o platform URL is http://comins.leydesdorff.net/. The interface is shown in [Figure 1: see original paper]. Google Chrome or Safari browsers are recommended; some browsers (e.g., Firefox) are less suitable. Uploading the data.txt file enables online analysis. Note that the tool requires datasets under 15MB.

3.3 Interpreting Standard RPY S Results

The upper portion of the results displays the standard reference publication year spectroscopy for the digital library field from 1900-1999 [Figure 2: see original paper]. The x-axis shows reference publication years, while the y-axis presents two data series: a bar chart showing total citation frequency for each year' s references, and a spline curve showing the deviation of each publication year' s total citation frequency from the median of total citation frequencies for the previous year, two years prior, the current year, the following year, and two years after. Hovering over each publication year displays the total citation frequency and deviation values. For example, total citation frequencies for 1981-1985 are 295, 301, 470, 380, and 433 respectively, with a median of 380. The 1983 frequency of 470 deviates from this median by 90, creating a peak on the spline curve. Clicking and dragging on the map displays the spectroscopy for selected year ranges, showing clearer variations. For instance, dragging from 1900 to 1944 clearly presents the 1900-1945 range [Figure 3: see original paper], while dragging from 1946 to 1960 clearly shows peaks in that period [Figure 4: see original paper].

The lower portion of the results page presents a searchable literature list [Figure 5: see original paper], showing the top 40 most-cited references for selected years. The list includes five columns: author, publication year, source publication, total citation frequency, and link to access the document. This helps users locate important historical root literature. For example, 1983 shows a clear peak, indicating frequently cited references with significant influence. Searching "rpy1983" retrieves the top 40 most-cited 1983 publications, sortable by citation frequency.

3.4 Standard RPY S Results Analysis

Based on the spectroscopy results, we analyze literature important to the digital library field's origins and evolution. Figure 2 shows that from 1900-1960, overall annual citation frequencies were low (below 100). From 1961-1985, frequencies steadily increased (from ~100 to over 400). From 1986 onward, frequencies grew exponentially (from over 600 to more than 4,000), indicating rapid development. Based on these observations, we divide the pre-emergence period into three

segments: 1900-1960, 1961-1985, and 1986-1993, analyzing peak years according to W. Marx et al.' s 2014 methodology, which focuses on single most-cited documents at peak points.

3.4.1 1900-1960 Digital Library Field Standard RPY S Analysis Figure 2 shows one major peak in 1945 during 1900-1960, with additional notable peaks in 1913, 1926, 1938, 1956, and 1960 [FIGURE:3 and FIGURE:4]. Table 1 lists the most-cited references at these six peaks.

The 1945 peak features V. Bush' s article "As We May Think" in *The Atlantic Monthly*, translated by Professor Xu Yuequan as "我们可以这样设想" ("We Can Imagine Thus"). The article detailed computer technology' s prospects for information collection, storage, discovery, and retrieval, mentioning libraries six times and envisioning library mechanization. Bush introduced the concept "Memex" (memory extender), a device for storing all materials accessible via a screen, keyboard, buttons, and levers.

The 1913 peak features German psychologist H. Ebbinghaus' s paper on the "forgetting curve," studying practice' s effect on memory. The 1926 peak features A. J. Lotka's "Lotka's Law,"a bibliometric law revealing the relationship between author frequency and document quantity. The 1938 peak features H. G. Wells' s *World Brain*, proposing a universal knowledge system accessible to all. The 1956 peak features G. A. Miller' s famous paper "The Magical Number Seven," showing human memory limitations. The 1960 peak features J. Cohen' s "kappa coefficient," a statistical measure widely used for assessing agreement.

3.4.2 1961-1985 Digital Library Field Standard RPY S Analysis Figure 2 shows six clear peaks from 1961-1985: 1965, 1967, 1973, 1975, 1979, and 1983. Table 2 lists the most-cited references.

The first peak (1965) features D. J. D. Price' s "Networks of Scientific Papers" in *Science*. Price, a founder of scientometrics, used SCI citation data to demonstrate citation networks and their application to scientometric research. The second peak (1967) features B. G. Glaser and A. L. Strauss' s grounded theory, a qualitative research method widely applied in library and information science. The third peak (1973) features H. Small' s co-citation analysis concept, highly influential in citation analysis. The fourth peak (1975) features G. Salton' s vector space model for information retrieval, foundational for modern search technology and applied in the SMART system. The fifth peak (1979) features C. J. Van Rijsbergen' s *Information Retrieval*, a classic textbook by a pioneer in modern information retrieval. The sixth peak (1983) features Salton' s *Introduction to Modern Information Retrieval*, providing theoretical foundations widely cited in the field.

3.4.3 1986-1993 Digital Library Field Standard RPY S Analysis Figure 2 shows 1986 as a peak year, 1987 as a relative trough, and 1988-1993 as

a flat line (median = 0), indicating that RPY S is ineffective when total citation frequencies grow continuously. However, the rapid growth trend shows the field entered a period of rapid development, paving the way for digital libraries' emergence in 1994. Table 3 details the most-cited documents.

These focus on information retrieval: M. J. Bates (1986) on query models; G. W. Furnas et al. (1987) on vocabulary matching; G. Salton et al. (1988) on term-weighting approaches; F. D. Davis (1989) on user acceptance models; S. Deerwester et al. (1990) on latent semantic indexing; C. C. Kuhlthau (1991) on information-seeking from a cognitive perspective; D. Goldberg (1992) on collaborative filtering; and E. Fox et al. (1993) on digital library usability evaluation.

3.5 Interpreting Multi-dimensional RPY S Results

Uploading data.txt for Multi-RPY S analysis produces Figure 6 [Figure 6: see original paper], showing the multi-dimensional reference publication year spectroscopy for 1900-1999. The x-axis shows reference publication years, and colors represent annual citation heat values—darker colors indicate higher values. Continuous dark bands suggest sustained citation, helping identify historically important references.

The lower results page presents a searchable list [Figure 7: see original paper] showing yearly citation frequencies for references by publication year (searchable also by author and journal). The six-column list includes: author, publication year, source, citation frequency, citing year, and access link.

3.6 Multi-dimensional RPY S Results Analysis

Figure 6 shows three publication years with continuous dark bands: (1) 1945 references cited continuously during 1994-2006; (2) 1975 references cited during 1999-2004, 2008-2012, and 2013-2017; (3) 1983 references cited during 1999-2004 and 2006-2016.

Searching RPY 1945 shows V. Bush' s "As We May Think" cited annually from 1994-2006 [Figure 7: see original paper]. The 1975 list shows Salton' s vector space model cited frequently after digital libraries'emergence in the 1990s [Figure 8: see original paper]. The 1983 list shows Salton' s *Introduction to Modern Information Retrieval* cited repeatedly from 1997 onward [Figure 9: see original paper], demonstrating its enduring importance.

3.7 Research Conclusions

Based on standard and multi-dimensional RPY S analyses, we identified 20 documents important to digital library origins, with three showing sustained influence. Eight appeared in *Journal of the Association for Information Science & Technology*, *Journal of the American Society for Information Science*, and *Communications of the ACM*, confirming these journals' importance. Three were by G. Salton, confirming his pivotal role.

We propose a three-stage evolution: (1) **Fantasy Period (1913-1960)**: Ebbinghaus and Miller identified short-term memory limitations; Wells and Bush imagined systems for long-term storage and easy access. (2) **Foundation Period (1961-1985)**: Basic research in information retrieval and scientometrics laid groundwork, including Price' s and Small' s seminal scientometric papers, Salton' s vector model, and classic textbooks by Van Rijsbergen and Salton. (3) **Development Period (1986-1994)**: Research became more specialized, addressing query models, vocabulary matching, term-weighting, semantic indexing, user interaction, and collaborative filtering, driving rapid field growth.

Using RPYsi/o, we explored and identified literature significantly influencing digital library origins and evolution. The tool accurately identifies origin-related classics and visualizes enduring contributions. However, final determination of root literature requires expert analysis, and the tool has limitations in dataset size and temporal coverage.

References

[1] Deng Xianglian. Analysis of the origin and conceptual connotations of digital library research[J]. Library Work and Research, 2003(1): 18-20.

[2] Lesk M. A personal history of digital libraries[J]. Library Hi Tech, 2012, 30(4): 592-603.

[3] Iris X, Krystyna K. Chapter 1: Introduction to digital libraries, in Discover Digital Libraries[M]. Oxford: Elsevier, 2016.

[4] Yang Guoli. Visualizing international digital library research progress: A study based on keyword co-occurrence and document co-citation[J]. Library Journal, 2012(6): 20-25.

[5] Yan Weidong. Visual analysis of digital library development[J]. Public Library, 2012(1): 30-34.

[6] Yang Jiulong, Du Wenlong. Evolutionary analysis of domestic digital library research based on knowledge visualization mapping[J]. Library Science Research, 2012(5): 5-9, 39.

[7] Hong Lingzi, Huang Guobin, Yu Yang. Comparative analysis of domestic and international digital library research papers based on CiteSpace[J]. Library Tribune, 2014(6): 91-100.

[8] Godeaux L. A co-word analysis of digital library field in China[J]. Scientometrics, 2012, 91(1): 203-217.

[9] Garfield E, Sher I H, Torpie R J. The use of citation data in writing the history of science[M]. Philadelphia: Institute for Scientific Information, 1964.

[10] Garfield E, Pudovkin A I, Istomin V S. Why do we need algorithmic historiography?[J]. Journal of the American Society for Information Science and Technology, 2003, 54(5): 400-412.

[11] Marx W, Bornmann L, Barth A. Detecting the historical roots of research fields by reference publication year spectroscopy (RPY S)[C]//Proceedings of 14th International Society of Scientometrics and Informetrics conference. Stolberg: Facultas v, 2013: 493-506.

[12] Li Xin, Li Qian. A supplementary perspective to traditional bibliometrics and scientific evaluation: Full-time domain RPY S[J]. Library and Information Service, 2017(4): 89-99.

[13] Comins J A, Leydesdorff L. RPYsi/o: Software demonstration of a web-based tool for the historiography and visualization of citation classics, sleeping beauties and research fronts[J]. Scientometrics, 2016, 107(3): 1509-1517.

[14] Marx W, Bornmann L. On the origins and the historical roots of the Higgs boson research from a bibliometric perspective[J]. European Physical Journal Plus, 2014, 129(6): 1-13.

[15] Marx W, Bornmann L, Barth A. Detecting the historical roots of research fields by reference publication year spectroscopy (RPY S)[J]. Journal of the Association for Information Science and Technology, 2014, 65(4): 751-764.

[16] Marx W, Bornmann L. Tracing the origin of a scientific legend by reference publication year spectroscopy (RPY S): The legend of the Darwin finches[J]. Scientometrics, 2014, 99(3): 839-845.

[17] Comins J A, Hussey T W. Detecting seminal research contributions to the development and use of the global positioning system by reference publication year spectroscopy[J]. Scientometrics, 2015, 104(2): 575-580.

[18] Marx W, Haunschild R, Thor A. Which early works are cited most frequently in climate change research literature? A bibliometric approach based on Reference Publication Year Spectroscopy[J]. Scientometrics, 2017, 110(1): 335-353.

[19] Li Xin, Lu Wei, Li Xuhui. An emerging method for exploring historical roots of disciplinary fields: RPY S[J]. Library and Information Service, 2016, 60(20): 70-76.

[20] Li Xin, Zhao Wei, Xiao Xianglong, et al. RPY S analysis of citation analysis research: Origin and evolution[J]. Library Tribune, 2017, 37(11): 56-65.

[21] Comins J A, Leydesdorff L. Citational algorithms for identifying research milestones driving biomedical innovation[J]. Scientometrics, 2017, 110(3): 1495-1504.

[22] Hou J. Exploration into the evolution and historical roots of citation analysis by referenced publication year spectroscopy[J]. Scientometrics, 2017, 110(3): 1437-1452.

[23] Bush V. As we may think[J]. The Atlantic Monthly, 1945, 176(1): 101-108.

[24] Xu Yuequan, Yu Ning. Interpretation of the scientific classic "As We May Think" [J]. Library Journal, 2006, 25(11): 11-14.

[25] Ebbinghaus H. Memory: A contribution to experimental psychology[M]. Boston: University Microfilms, 1913.

[26] Lotka Alfred J. The frequency distribution of scientific productivity[J]. Journal of the Washington Academy of Sciences, 1926, 16(12): 317-323.

[27] Wells H G. World Brain[M]. First UK Edition. London: Methuen & Co., 1938.

[28] Miller G A. The magical number seven[J]. Psychological Review, 1956, 63: 81-97.

[29] Cohen J. A coefficient of agreement for nominal scales[J]. Educational & Psychological Measurement, 1960, 20(1): 37-46.

[30] Price D J D. Networks of scientific papers[J]. Science, 1965, 149(3683): 510-515.

[31] Glaser B G, Strauss A L. Discovery of grounded theory: Strategies for qualitative research[M]. New York: Aldine De Gruyter, 1967.

[32] Wang Ping, Ru Jiaqi. Factors influencing domestic minor library service satisfaction: An exploratory study based on grounded theory[J]. Library and Information Service, 2015, 59(19): 41-46.

[33] Lin Ting. Empirical research on Folksonomy management mechanisms in Chinese university libraries based on classic grounded theory[J]. Library and Information Service, 2015, 59(16): 60-67.

[34] Ke Ping, Zhang Wenliang, Li Xining, et al. Research on librarians'perception of organizational culture in public libraries based on grounded theory[J]. Journal of Library Science in China, 2014, 40(3): 37-46.

[35] Small H. Cocitation in scientific literature: A new measure of relationship between documents[J]. Journal of the American Society for Information Science, 1973, 24(4): 265-269.

[36] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.

[37] Van Rijsbergen C J. Information retrieval[M]. London: Butterworths, 1979.

[38] Salton G. Introduction to modern information retrieval[M]. New York: McGraw-Hill, 1983.

[39] Bates M J. Subject access in online catalogs: A design model[J]. Journal of the Association for Information Science & Technology, 1986, 37(6): 357-376.

[40] Furnas G W, Landauer T K, Gomez L M, et al. The vocabulary problem in human-system communication[J]. Communications of the ACM, 1987, 30(11): 964-971.

[41] Salton G, Buckley G. Term-weighting approaches in automatic text retrieval[J]. Information Processing & Management, 1988, 24(5): 513-523.

[42] Davis F D. Perceived usefulness, perceived ease of use, and user acceptance of information technology[J]. Society for Information Management and the Management Information Systems Research Center, 1989, 13(3): 319-340.

[43] Deerwester S, Dumais S T, Furnas G W, et al. Indexing by latent semantic analysis[J]. Journal of the Association for Information Science & Technology, 1990, 41(6): 391-407.

[44] Kuhlthau C C. Inside the search process: Information seeking from the user's perspective[J]. Journal of the Association for Information Science & Technology, 1991, 42(5): 361-371.

[45] Goldberg D. Using collaborative filtering to weave an information tapestry[J]. Communications of the ACM, 1992, 35(12): 61-70.

[46] Fox E A, Hix D, Nowell L T, et al. Users, user interfaces, and objects: Envision, a digital library[J]. Journal of the American Society for Information Science, 1993, 44(8): 480-491.

Author Contributions

Wu Chuang: Designed research framework and wrote the paper; **Xie Fuxiu:** Participated in framework design and revised sections; **Wang Chunlei:** Participated in framework design and revised sections; **Liu Wanguo:** Provided overall guidance and revision suggestions; **Sun Bo:** Revised sections.

English Abstract

[**Purpose/significance**] This article aims to explore seminal works about the historical roots of a specific research field or subject. The study of historical roots is of great significance for the construction and research. [**Method/process**] We describe a technical advancement for developing research historiographies by introducing RPYsi/o, an online tool for performing standard RPY S and multi-RPY S analyses. Based on RPYsi/o, we take digital library research field as an example. [**Result/conclusion**] The tool enables users to explore seminal works underlying a research field and to plot the influence of these seminal works over time.

Keywords: RPYsi/o; digital library; standard RPY S; multi-RPY S; historical roots

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