Thematic Analysis of Big Data Major Curriculum and Its Implications for Library, Information and Archival Science: Postprint

Authors: Yang Jie, Zhao Xing

Date: 2023-04-01T15:51:22+00:00

Abstract

[Purpose/Significance] Against the backdrop of the big data wave and the "New Liberal Arts" initiative, the talent cultivation paradigm of China's library, information and archives management discipline is in urgent need of reform. Meanwhile, the development of big data-related majors is burgeoning, offering valuable references for building a new talent cultivation paradigm for the LIS discipline.

[Method/Process] This study adopts a temporal topic network model and computational method; by collecting, processing, and statistically analyzing texts of big data major cultivation programs from 259 institutions of higher education, it conducts topic mining along the temporal dimension, summarizes the hierarchy of data science courses, analyzes the connections between core knowledge of the LIS discipline and big data majors, and provides recommendations for data science courses suitable for the LIS discipline.

[Results/Conclusions] The results indicate that the adopted temporal topic network model method can effectively align with the analysis of talent cultivation paradigm development in big data majors and can serve as a methodological approach for researching disciplinary topics. Furthermore, recommendations for data science courses oriented toward the LIS discipline are provided, which can serve as a reference for talent cultivation in the LIS discipline.

Full Text

Thematic Analysis of Big Data Program Curricula and Its Implications for Library and Information Science

Yang Jie[1], **Zhao Xing**[1,2] [1]Department of Information Management, School of Economics and Management, East China Normal University, Shanghai 200062

[2]Institute for Academic Evaluation and Development, East China Normal University, Shanghai 200062

Abstract: [Purpose/Significance] Against the backdrop of the big data wave and the "New Liberal Arts" initiative, China's library, information, and archival science (LIS) discipline urgently needs to reform its talent cultivation paradigm. Meanwhile, the construction of big data-related programs is flourishing, offering valuable references for developing a new talent cultivation model for LIS. [Method/Process] This study employs a sequential topic network model and computational method. By collecting, processing, analyzing, and statistically examining the curriculum documents of big data programs from 259 higher education institutions, we conduct temporal topic mining to summarize the hierarchical structure of data science courses. We analyze the connections between core LIS knowledge and big data programs, and provide recommendations for data science courses appropriate for LIS disciplines. [Result/Conclusion] The results demonstrate that the sequential topic network model effectively aligns with the analysis of talent cultivation paradigm development in big data programs and can serve as a methodological approach for studying disciplinary topics. Additionally, the study offers data science course recommendations for LIS disciplines, providing references for talent cultivation in LIS.

Keywords: talent cultivation; data science; new liberal arts; sequential topic network model **Classification Number:** G254.9 **DOI:** 10.13266/j.issn.0252-3116.2022.02.012

The construction of "New Liberal Arts" represents a crucial component in enhancing China's national soft power and promoting cultural prosperity, as well as a key focus of high-quality education system reforms. Cultivating talent constitutes the core of "New Liberal Arts" construction, requiring adherence to educational principles, collaborative cultivation, and diversified models to develop applied and interdisciplinary talents suited to the new era. Library, information, and archival management (hereinafter referred to as "LIS") possesses inherent interdisciplinary characteristics bridging arts and sciences, necessitating significant innovation in its talent cultivation model under the "New Liberal Arts" framework. When data science first emerged, Ye Ying and Ma Feicheng noted that data science and information science share common theoretical logic and technical methods. "Data and intelligence empowerment" has become a new development trend for LIS. Sun Jianjun et al. argued that data management and analytical technologies offer new possibilities for LIS development, while Zhao Xing et al. proposed expansion directions for LIS based on data intelligence and knowledge discovery. Drawing upon Ye Ying's summary of core LIS knowledge, this study explores data science curriculum construction schemes for LIS to provide a foundation for innovative development in LIS talent cultivation models.

Dilemmas and Reflections

Practical Issues

Examining the current state of undergraduate programs in most LIS departments—specifically the Information Management and Information Systems major—13 and 16 institutions discontinued this major in 2020 and 2021 respectively, ranking 3rd and 2nd among all discontinued undergraduate programs in regular higher education institutions. In contrast, big data programs have proliferated rapidly, with over 50 institutions applying to establish Data Science and Big Data Technology or Big Data Management and Application programs in each of the past two years. The most pessimistic prediction suggests that LIS-related programs may face "extinction, merger, replacement, renaming, or marginalization" within a decade. This stark reality demonstrates the urgent need for LIS talent cultivation model reform to align with the "New Liberal Arts" era.

Academic Reflections

Since the "New Liberal Arts" initiative was proposed, numerous scholars have offered new perspectives on LIS future development. Chu Jingli suggested that LIS should evolve into a "hard discipline" by strengthening the integration of technologies and methods while maintaining its core identity. Ma Feicheng et al. advocated seizing the opportunities presented by New Liberal Arts construction to emphasize interdisciplinary integration while preserving humanistic traditions. Zhang Jiuzhen proposed that LIS interdisciplinary integration under the New Liberal Arts framework should be "subject-centered and purpose-driven." Regarding the preservation of disciplinary core identity, Ye Ying summarized the core knowledge and research methods of library and information science, providing a clear foundation for LIS development. Zhou Wenjie noted that the "old" core that LIS New Liberal Arts construction needs to preserve includes: developing data science based on scientific data, supporting digital humanities as infrastructure, "compiling" digital memory through knowledge organization, upholding humanistic care in public cultural services, and expanding the evidence-based decision-making function of reference services.

Innovative Development

The "new" in "New Liberal Arts" signifies innovation. The construction of New Liberal Arts talent cultivation models can draw upon new methods and technologies from "New Engineering" programs. LIS can explore new development directions by learning from big data programs. This approach is not original to our study—Chen Mo et al. examined data science and big data technology programs from an information science perspective, categorizing them into foundation, methodology, and application courses. Tao Jun et al. analyzed data science curricula in multiple international iSchool programs to provide recommendations for LIS data science course construction. Zhao Xing et al. employed

content analysis to study big data management and application program curricula. Li Haibo et al. investigated data science course group construction for information management programs, offering new ideas for data science competency cultivation. Yan Hui et al. used design ethnography and future interviews to predict that data science and LIS would most likely develop a complementary relationship. This study explores big data program cultivation models to provide new insights for LIS talent cultivation.

Thematic Analysis of Big Data Programs

Current State of Big Data Programs

Big data programs in China first launched in 2016, with Peking University, University of International Business and Economics, and Central South University establishing the Data Science and Big Data Technology major. Subsequently, 32, 248, 203, 143, and 62 institutions launched this major from 2017 to 2021 [Figure 1: see original paper]. This major awards either engineering or science degrees, with some universities like Xiamen University, Renmin University of China, and Shanghai University of Finance and Economics offering it in both science and engineering faculties. By early 2021, 674 institutions had established this major, including 29 former "985 Project" universities and 73 former "211 Project" universities. The Big Data Management and Application major emerged in 2017, with 5, 25, 52, and 68 institutions launching it from 2018 to 2021. By 2021, 140 universities offered this major, including 5 former "985 Project" and 25 former "211 Project" institutions, all awarding management degrees.

Through email inquiries and official university websites, we collected cultivation plans from these institutions. While data from some universities were unavailable, we ultimately gathered big data program curricula and core courses from 86 Big Data Management and Application programs and 173 Data Science and Big Data Technology programs, totaling 259 institutions. The sample encompasses various tiers of universities, including former "985 Project" universities (e.g., Peking University), former "211 Project" universities (e.g., Central China Normal University), other provincial/ministerial co-constructed universities (e.g., Guangdong Ocean University), provincial universities (e.g., Liaocheng University), and independent colleges (e.g., Chengyi College of Jimei University).

Thematic Relationships in Big Data Talent Cultivation Models

To analyze core themes and overall architecture in big data talent cultivation, this study employs co-occurrence technology to examine relationships among various themes. After preprocessing the curriculum texts through segmentation, stop-word removal, and topic indexing, we conducted thematic relationship mining to generate a chord diagram of big data talent cultivation themes [Figure 2: see original paper].

The chord diagram presents themes for both Data Science and Big Data Technology programs and Big Data Management and Application programs. Different arcs represent important themes in each program's cultivation plan, with arc length indicating theme importance and network density representing thematic interconnections. Analysis of [Figure 2: see original paper] reveals that Data Science and Big Data Technology programs focus on data analysis as their core, with computer technology and statistics as foundational techniques. Big Data Management and Application programs emphasize data management, with statistics, management science, and computer science as core technologies. The former cultivates abstract thinking, mathematical formalization, data science theory, and fundamental competencies, while the latter develops data collection, processing, analysis, and visualization skills. Under these cultivation models, students can apply data science methods to address complex business, management, or data engineering problems.

Temporal Evolution of Big Data Talent Cultivation Models

S. C. Deerwester et al. proposed a non-probabilistic topic model called Latent Semantic Indexing (LSI) in 1990, which T. Hofmann later extended into a probabilistic topic model, leading to the rise of probabilistic topic models. Traditional topic models lack temporal dimensions. This study draws upon temporal topic models from Liao Junhua et al., A. Bruns, M. J. Westgate et al., Z. F. Zhang et al., and Y. Zheng et al., introducing time dimensions through slicing methods.

To deeply analyze the paradigm development of big data programs temporally, we used Python 3.8.5, Gephi 0.9.2, and VOSviewer 1.6.16, applying the ForceAtlas2 algorithm for layout calculation. The specific methodology is as follows:

(1) Slice the text dataset into j segments based on different time periods.

(2) For each slice, perform stop-word removal, word frequency statistics, word co-occurrence matrix generation, and network distribution calculation based on probability.

(3) Combine and overlay all network distributions to obtain the sequential topic network.

The formulas are as follows:

$$m_t = \sum_{i=1}^{j} m_i \quad (1)$$

where m_t represents the total network distribution and m_i represents the network distribution of each slice.

$$p(w_1, \dots, w_m) = \prod_{i=1}^{m} P(w_i | w_1, \dots, w_{i-1}) \quad (2)$$

where p denotes the occurrence probability of word combinations in word vector encoding. Word co-occurrence matrix rows serve as word vectors, which are dimensionally reduced into node and edge data for subsequent visualization.

$$TRAN_x(n) = |X(n) - X(n-1)| + n \quad (3)$$

where $TRAN_x(n)$ represents the positional offset of a slice's co-occurrence network in the temporal network along a certain dimension. This offset is determined by the difference between adjacent slice networks $|X(n) - X(n-1)|$ and the slice's ordinal position n.

We analyzed the big data program curriculum texts as follows: Sliced texts by program establishment year; Removed stop words using the Harbin Institute of Technology stop-word list and a custom list; Calculated word frequencies using Python's pandas package and computed co-occurrence matrices using word vectors; Used VOSviewer to calculate network distributions for each slice; Combined all network distributions using Gephi to generate the sequential topic network for big data programs [Figure 3: see original paper].

Analysis of [Figure 3: see original paper] shows that when big data programs first launched, talent cultivation models emphasized technical directions such as data systems and engineering projects. As more institutions established these programs, cultivation models became diversified, broad-based, and open. Specifically, nodes for social sciences, statistics, and other themes became denser in cultivation texts, indicating that big data programs increasingly integrated data science with management, medicine, economics, and other disciplines, creating new cultivation models in e-commerce, information management, fintech, and smart healthcare. The cultivation paradigm gradually expanded from foundational theories and basic applications to diverse and broad application domains.

Three Levels of Big Data Courses

The above analysis demonstrates that big data programs emphasize cultivating fundamental mathematical abilities, data science thinking, and practical data science skills. In recent years, they have begun integrating knowledge from social sciences and other disciplines to provide students with more diversified development paths. Related courses have evolved from foundational theories and basic applications to broader knowledge domains.

Drawing on Chen Mo et al.'s framework, we divide big data core courses into three levels based on our thematic analysis: Foundation Theory and Methods, Big Data Theory and Methods, and Integrated Methods and Applications .

The Foundation Theory and Methods level cultivates fundamental mathematical thinking and basic data science literacy, including statistics courses and algorithm/programming courses such as Data Structures and Algorithms, and Python Programming and Applications.

The Big Data Theory and Methods level comprises core courses that develop students' core competencies in data science and big data, covering four areas: data storage and management, data collection and processing, data analysis and mining, and big data technologies. Data storage and management includes courses on databases, data warehouses, and Hadoop storage. Data collection and processing includes web scraping, data gathering, and information retrieval. Data analysis and mining includes data modeling, statistical analysis, machine learning, and visualization. Big data technologies includes fundamental analytics techniques such as distributed algorithms.

The Integrated Methods and Applications level provides multiple direction choices (tailored by institutions based on their circumstances), enabling students to apply data science fundamentals in specialized domains like business big data and medical big data. This effectively stimulates student interest and enhances practical abilities. For example, Fudan University's big data program offers diverse pathways including "Big Data Analytics in Science, Medicine, and Engineering," "Big Data Analytics in Social Sciences," and "Advanced Brain-Inspired Computing."

Innovative Development of LIS Talent Cultivation

Connections Between LIS Core Knowledge and Big Data Programs

In learning from big data program construction experiences to build new LIS talent cultivation models, we must grasp the core foundations of LIS education. S. R. Ranganathan established foundational principles of library science focused on library services. As times evolved, LIS core themes changed. Ye Ying expressed LIS core knowledge through hierarchical levels: conceptual, theoretical, and systematic. Core academics are divided into information organization, information retrieval, and information analysis, each refined across three levels.

Information organization is refined at the conceptual level into classification and indexing, at the theoretical level into classification systems, subject methods, cataloging, and indexing techniques, and at the systematic level into literature and knowledge systems. Information retrieval is refined at the conceptual level into precision and recall, at the theoretical level into search algorithms like Boolean retrieval, and at the systematic level into search engines. Information analysis is refined at the conceptual level into breadth, speed, precision, and accuracy, at the theoretical level into citation analysis and content analysis, and at the systematic level into quantitative and qualitative analysis. Based on this, we identified core themes in LIS talent cultivation systems: information organization, information retrieval, and information analysis.

We searched and counted these core terms across the 259 big data program curricula, using the total frequency $\sum_{i=1}^{j} x_j$ of core terms x_j as an indicator of the connection strength between LIS and big data programs. We summed frequencies for all core terms at each level to examine connections between LIS

core knowledge and big data programs based on the three-level course structure .

Evidently, data science has permeated all levels of information organization, information retrieval, and information analysis. Terms like "classification," "search algorithms," and "quantitative analysis" appear frequently, indicating that courses from the Foundation Theory and Methods level and portions of the Big Data Theory and Methods level align well with LIS core themes, offering valuable references for new LIS talent cultivation paradigms.

Data Science Curriculum Construction for LIS

Based on the above discussion, we identified alignment points between big data course levels and LIS core knowledge [Figure 4: see original paper]. The upper portion shows LIS core knowledge, while the lower portion displays the data science course chain. By mapping highly relevant course categories to corresponding LIS core knowledge, we derived alignment points to summarize a data science course group for LIS.

The data science course group for LIS talent cultivation should broadly adopt the big data program framework while emphasizing practical application. Institutions can offer Python programming at the Foundation Theory and Methods level, then sequentially offer data storage and management, data collection and processing, and data analysis and mining courses centered on Python, such as Python-based web scraping, data analysis, and visualization. At the Integrated Methods and Applications level, collaboration with business schools and other faculties can deliver big data courses like business intelligence, forming a cohesive "Python Programming Fundamentals → Python Web Scraping → Python Data Analysis and Visualization → Business Intelligence" course chain tailored for LIS.

As Ke Ping noted, LIS talent cultivation in the New Liberal Arts context cannot indiscriminately add inappropriate data science courses. Data science curriculum construction for LIS must remain grounded in core knowledge transmission of information organization, information retrieval, and information analysis. By integrating big data methods and technologies into traditional LIS domains, we can develop distinctive data science course groups. We selected relevant courses from big data programs that can cultivate LIS students' information organization, retrieval, and analysis capabilities .

When constructing data science courses for information organization capabilities, incorporate knowledge graphs, classification and indexing, and semantic networks. For information retrieval capabilities, beyond foundational data collection and retrieval, strengthen teaching on intelligent information retrieval and user recommendation algorithms integrated with deep learning and reinforcement learning. For information analysis capabilities, offer algorithm and programming courses such as Python programming based on Jupyter Notebook platforms, and incorporate cutting-edge data mining algorithms and citation

network analysis fundamentals.

It is crucial not to simply replicate big data courses but to adapt them based on LIS core knowledge and institutional contexts. LIS-oriented data science courses should prioritize practicality and integration over high-level data science theory and thinking.

Notably, data science curriculum construction is just one developmental path. Some institutions have already established data science as a self-designed secondary discipline within LIS. However, LIS also features other distinctive directions, such as Renmin University of China's focus on "Digital Humanities," Fudan University's initiatives in "Ancient Book Preservation and Intangible Cultural Heritage," and East China Normal University's "Business Analytics" program launched in 2015. Regardless of the development model, maintaining both tradition and innovation remains the prerequisite for reform.

LIS talent cultivation reform in the big data era has just begun and faces numerous challenges: How to deeply integrate data science courses with disciplinary characteristics? How to align LIS talent cultivation with social needs? How to internationalize LIS talent cultivation while retaining Chinese characteristics? These questions have been extensively discussed without clear consensus.

Methodologically, this study's sequential topic network model has limitations. Future research could: Develop more scientific slicing range metrics; Explore more suitable layout algorithms for clear, complete evolutionary network visualization; Apply PCA dimensionality reduction to mitigate co-occurrence sparsity.

References

[1] Ye Ying, Ma Feicheng. The rise of data science and its relationship with information science[J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(6): 575-580. [2] Sun Jianjun, Li Yang, Pei Lei. Reflections on the transformation of library and information science in the era of data and intelligence empowerment[J]. Document, Information & Knowledge, 2020(3): 22-27. [3] Zhao Xing, Qiao Lili, Ye Ying. An interdisciplinary expansion of library and information science toward data intelligence and knowledge discovery: Integration of data, scholarship, and creation[J]. Journal of Library Science in China, 2020, 46(6): 16-25. [4] Ye Ying. On the core knowledge and effective methods of library and information science: The utility of double-proof and model-based methods[J]. Journal of Library Science in China, 2021, 47(3): 58-66. [5] Yan Hui, Han Leiqian, Wu Meng, et al. A study on the development prospects of library science, information science, and archival science by 2029[J]. Library & Information, 2019(6): 2-17, 153. [6] Chu Jingli. "New Liberal Arts" calls for library and information science to become a "hard" discipline[J]. Library & Information, 2020(6): 1-3. [7] Ma Feicheng, Li Zhiyuan. Development prospects of China's library and information science discipline under the New Liberal Arts background[J]. Journal of Library Science in China,

2020, 46(6): 4-15. [8] Zhang Jiuzhen. LIS discipline construction needs to align with new era development[J]. Library & Information, 2020(6): 17-18. [9] Zhou Wenjie. From multiple heterogeneity to integrated unity: An evaluation of trends in LIS New Liberal Arts construction[J]. Information and Documentation Services, 2021, 42(2): 14-21. [10] Chen Mo, Li Guangjian, Chen Congcong. Talent cultivation in data science and big data technology programs from an information science perspective[J]. Library and Information Service, 2019, 63(12): 5-11. [11] Tao Jun, He Xiaodong. A comparative study of data science curriculum structures for library and information science[J]. Research on Library Science, 2019(6): 10-16. [12] Zhao Xing, Yu Xiaoting, Wan Lingyu. Content analysis of cultivation characteristics for big data management and application programs under the New Liberal Arts background[J]. Library & Information, 2020(6): 26-34, 92. [13] Li Haibo, Xie Jianmin. Research on data science course group construction for information management programs under the New Liberal Arts background[J]. Information Science, 2020, 38(8): 128-133. [14] Yan Hui, Han Yanfang, Zhang Yuhao, et al. A predictive study on the relationship between library and information science and New Liberal Arts interdisciplinary fields[J]. Information and Documentation Services, 2021, 42(1): 21-27. [15] DEERWESTER SC, DUMAIS ST, LANDAUER TK, et al. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 1990, 41(6): 391-407. [16] HOFMANN T. Probabilistic latent semantic indexing[C]//Proceedings of the 22nd annual international SIGIR conference. New York: ACM Press, 1999: 50-57. [17] Liao Junhua, Sun Keying, Zhong Lixia. A network hot topic evolution analysis system based on temporal topic models[J]. Library and Information Service, 2013, 57(9): 96-102, 118. [18] BRUNS A. How long is a tweet? Mapping dynamic conversation networks on Twitter using Gawk and Gephi[J]. Information, Communication & Society, 2012, 15(9): 1323-1351. [19] WESTGATE MJ, BARTON PS, PIERSON JC, et al. Text analysis tools for identification of emerging topics and research gaps in conservation science[J]. Conservation Biology, 2015, 29(6): 1606-1614. [20] ZHANG ZF, LI QD. QuestionHolic: Hot topic discovery and trend analysis in community question answering systems[J]. Expert Systems with Applications, 2011, 38(6): 6848-6855. [21] ZHENG Y, MENG ZP, XU C. A short-text oriented clustering method for hot topic extraction[J]. International Journal of Software Engineering and Knowledge Engineering, 2015, 25(3): 453-471. [22] JACOMY M, VENTURINI T, HEYMANN S, et al. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software[EB/OL]. [2021-12-09]. https://www.ncbi.nlm.nih.gov/sites/pmc/articles/PMC4051631/. [23] RANGANATHAN SR. The five laws of library science[M]. 2nd ed. Madras: The Madras Library Association, 1957. [24] Ke Ping. New LIS: The development of library, information, and archival management as a first-level discipline in New Liberal Arts construction[J]. Information and Documentation Services, 2021, 42(1): 15-20.

Author Contributions: Yang Jie: Responsible for data collection, data analysis, data visualization, and paper writing. Zhao Xing: Responsible for topic

selection and paper revision.

Abstract: [Purpose/Significance] Under the background of the big data tide and the new liberal arts, the talent training paradigm of Chinese library and information science urgently needs to be reformed. At the same time, the construction of big data related majors is in the ascendant, which has reference significance for the construction of a new paradigm of talent training for library and information science. [Method/Process] This paper innovatively proposes a sequential topic network model and calculation method. By collecting, processing, counting and analyzing the text of big data professional training programs of 259 colleges and universities, topic mining is carried out in the time dimension, and the levels of data science courses are summarized. The connection between the core knowledge of library and information science and big data majors is analyzed, and suggestions for data science courses suitable for library and information science are given. [Result/Conclusion] The results show that the sequential topic network model can better fit the analysis of the development of talent training paradigm for big data majors, and can become a method for researching subject topics. In addition, suggestions for data science courses for library and information science are also given, which can be used as a reference for talent training in library and information science.

Keywords: talent cultivation; data science; new liberal arts; sequential topic network model

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