Research on the Construction of Data Literacy Curriculum Clusters in Higher Education in the Age of Big Data: Postprint

Authors: Shen Jiujiu, Xu Ping, Zhang Qin, Gong Huaping

Date: 2023-07-26T00:00:00+00:00

Abstract

[Purpose/Significance] In the big data environment, the management of scientific data is receiving increasing attention from all sectors of society. Consequently, the cultivation of data literacy talents in universities has become increasingly important, and data literacy courses serve as a crucial foundation for fostering data literacy among university students. [Method/Process] This study employs questionnaire surveys and Pearson correlation analysis in SPSS 22 software to analyze the relationship between information literacy courses—which are closely related to data literacy—and data literacy itself. From this analysis, a collection of data literacy courses is extracted, and cluster analysis is subsequently applied to modularly classify these courses. [Results/Conclusion] The research results demonstrate that information literacy-related courses can serve as a foundation for data literacy cultivation. Building upon this foundation, by supplementing and refining corresponding data literacy courses, and based on the data lifecycle theory, a data literacy curriculum cluster is constructed, comprising modules on data awareness and ethics, data acquisition and processing, and data analysis and utilization.

Full Text

Preamble

Volume 63, Issue 19, October 2019
ChinaXiv Cooperative Journal

Research on the Construction of University Data Literacy Curriculum Groups in the Big Data Era

Shen Jiujiu, Xu Ping, Zhang Qin, Gong Huaping
Department of Information Management, School of Management, Nanchang University, Nanchang 330031

Abstract

[**Purpose/Significance**] In the big data environment, the management of scientific data has received increasing attention from all sectors of society. Consequently, the cultivation of data literacy talents in universities has become increasingly important, and data literacy curricula serve as a crucial foundation for developing students' data literacy. [**Method/Process**] This study employs questionnaire surveys and Pearson correlation analysis using SPSS 22 software to analyze the relationship between information literacy courses and data literacy, which are closely related. It extracts a collection of data literacy courses and then uses cluster analysis to modularize and classify these courses. [**Result/Conclusion**] The findings indicate that information literacy-related courses can serve as the foundation for data literacy cultivation. Based on this foundation, by supplementing and improving corresponding data literacy courses and drawing upon data lifecycle theory, this study constructs a data literacy curriculum group comprising three modules: data awareness and ethics, data acquisition and processing, and data analysis and utilization.

Keywords: data literacy; information literacy; curriculum group; correlation analysis; cluster analysis
Classification Number: G252.7
DOI: 10.13266/j.issn.0252-3116.2019.19.005

Introduction

With the advent of the big data era, data-driven information management has become increasingly important in scientific research and social management decision-making. On January 23, 2018, President Xi Jinping chaired the second meeting of the Central Committee for Comprehensively Deepening Reform, which emphasized the need to adapt to the demands of the times, strengthen scientific data management, and actively establish data security mechanisms [1]. This signifies that scientific data management has gained attention at the national leadership level. Consequently, data literacy has become a fundamental competency requiring urgent cultivation. As the primary force in cultivating data literacy, universities should strengthen students' data literacy development, particularly for majors closely related to data and information. These programs should timely adjust their educational objectives and plans for data literacy cultivation to meet the talent demands of the big data era.

A review of current international research on data literacy education reveals relatively mature systems already in place. T. Koltay outlined the main characteristics of data-intensive research paradigms and noted that academic libraries can serve as centralized hubs for research data services [2]. E. Stephenson et al. conducted a pilot course at UCLA, co-taught by librarians and data archivists, using syllabi and assignments to reflect sociological research questions and tools in social sciences as well as information literacy competencies [3]. Domestic research on data literacy education primarily focuses on three aspects: (1) investigations, comparisons, and insights from domestic and international data literacy education; (2) construction of system models, training mechanisms, and educational paradigms for data literacy education in universities and libraries; and (3) review studies related to data literacy education.

Meng Xiangbao et al. summarized foreign data literacy education practices into three models: using LibGuides to establish scientific data management resource navigation, general data literacy education, and disciplinary data literacy education [4]. Huang Ruhua et al. explored the concept of data literacy, studied its teaching content, and discussed how to provide data literacy education for different institutions and classify audiences [5]. Additionally, Wei Lai et al. [6] examined the information literacy cultivation status of major Chinese universities from the perspectives of teaching objects, approaches, and content, offering suggestions for curriculum content development in China.

Through extensive literature review, we found that current research on data literacy education remains at the macro-level paradigm stage and has not yet addressed specific curriculum design, despite the need for concrete implementation frameworks. Therefore, this study consulted expert opinions to extract information literacy-related courses closely related to data literacy from the Information Management and Information Systems major (hereinafter referred to as "Information Management Major"). Using questionnaire surveys and Pearson correlation analysis, we identified courses with significant impact on data literacy. Based on the requirements of the big data era, we integrated relevant new knowledge and content in data management and, according to data lifecycle theory, formed an independent data literacy curriculum system to provide an implementation basis for general data literacy education among university students. The modular division of data literacy courses will help training institutions implement personalized cultivation plans according to specific professional requirements.

1. Theoretical Foundation

1.1 Connotation of Data Literacy

Existing research has not provided a standardized definition of data literacy, which is often considered an extension of information literacy. Two definitions are most frequently cited in academia: First, E. Stephenson et al. [3] defined data literacy as "the ability to find, evaluate, and effectively use data." Second, K. Hogenboom et al. considered data literacy as the ability to read, interpret, analyze, critically think about data, and use data as evidence [7]. Other scholars have also offered definitions. For example, J. R. Carlson et al. identified 12 core competencies for data literacy, including data formats, data management, data collection, data processing applications, data visualization, and data ethics [8]. Hu Hui et al. defined data literacy as the ability to acquire data appropriately, select and evaluate data objectively, and manage, process, and utilize data [9]. Long Qian proposed that university faculty and students' data literacy includes data awareness, data acquisition, data processing and analysis, data communication, data evaluation, and data ethics [10].

Based on previous research, data literacy curriculum construction will focus on five dimensions: data awareness, data ethics, data acquisition, data processing, and data analysis and utilization. Data awareness refers to understanding the importance of data and access channels, representing a subjective perception. Data ethics encompasses rules and legal policies related to data, such as citing data sources and ensuring data authenticity. Data acquisition competency involves mastering methods for obtaining primary and secondary data, such as conducting surveys and using various data retrieval tools. Data processing ability refers to the capacity to use various data processing tools and software (Ucinet, SPSS, AMOS, etc.) to visualize collected data and present it in charts. Data analysis competency involves proficiency in multiple data analysis software, selecting effective analytical methods, and 挖掘内在价值的能力.

1.2 Connotation of Curriculum Group

A curriculum group refers to a small cluster of several courses in a curriculum system that share certain characteristics or perform similar functions [11], representing an organic curriculum system module constructed based on knowledge systems. University curriculum groups typically follow a four-level system: "general education courses for arts and sciences, general education courses for disciplines, foundational courses for majors, and specialized courses" [12]. We argue that a data literacy curriculum group is an organic collection of relevant courses that influence data literacy, aimed at cultivating university students' data literacy. It can serve both as a basis for strengthening data literacy education and as a reference for relevant majors with data literacy requirements to adjust their training plans.

1.3 Relationship between Data Literacy and Information Literacy

Meng Xiangbao et al. [4] proposed that data literacy is similar to information literacy, suggesting that data literacy education can be viewed as an expansion and refinement of information literacy education in the big data era [13]. Zhang Jinliang et al. [14] argued that data literacy has an undeniable close relationship with information literacy and constitutes an essential component of it. Data literacy represents an innovative development of information literacy against the backdrop of big data and serves as its core component [15]. J. C. Prado et al. [16] defined data literacy as a component of information literacy. Zhang Jingbo [17] considered data literacy a subset of information literacy, with data literacy education representing a further improvement and deepening of traditional information literacy education. Based on these perspectives, we believe that information includes various types, with data-based information being one type, and that information literacy should encompass data literacy. Therefore, information literacy courses may influence data literacy cultivation to some extent, and data literacy curriculum groups can be constructed with reference to information literacy courses.

1.4 Information Literacy-Related Curriculum Setup and Interpretation

China's Information Management and Information Systems major was established following the "Undergraduate Major Catalog and Professional Descriptions for Regular Institutions of Higher Education" issued by the Ministry of Education in July 1998 [18]. It is currently offered at Peking University, Wuhan University, Nanjing University, Sun Yat-sen University, East China Normal University, Heilongjiang University, Central China Normal University, Nanchang University, and other institutions. Taking Wuhan University as a representative example, its training objectives for the Information Management and Information Systems major indicate that the program cultivates senior interdisciplinary and innovative talents with foundations in information management, high information literacy, mastery of systematic thinking, and knowledge and abilities in information system design and management methods, enabling them to engage in information management and information system analysis, design, implementation, management, and evaluation in national departments. Information literacy is the core competency of students in this major. Relevant universities have incorporated information literacy courses into their professional training plans for the Information Management and Information Systems major. After consulting six teachers and experts in this field, we extracted courses closely related to information literacy from the curricula of several universities, as shown in Table 1 .

According to our analysis, although information literacy course offerings vary slightly across different universities, they are generally quite similar. To ensure convenience and reliability in data collection, our research team selected the Information Management and Information Systems major in the School of Management at our university as the research object. The major's training plan was formulated according to the spirit of the training guidelines from the Education Steering Committee and has undergone multiple revisions, with the current version being the 2016 edition. The professional courses in the training plan mainly include three modules: economics and management, computer science, and information science. Table 2 presents the main courses and details after removing economics and management courses and some pure computer language courses. We will use empirical research and correlation analysis to explore the impact of these relevant courses on data literacy cultivation and identify those closely related to data literacy, thereby constructing a suitable data literacy curriculum collection for data literacy education.

2. Empirical Research

2.1 Questionnaire Survey

This survey targeted undergraduate students from the 2015 to 2018 cohorts majoring in Information Management and Information Systems in the School of Management at Nanchang University. We distributed paper questionnaires offline between October 17 and November 15, 2018. A total of 210 questionnaires were distributed, 200 were collected, and 192 were valid.

2.1.1 Questionnaire Design The questionnaire consisted of three parts: (1) basic user information, including gender and grade level; (2) data literacy competency assessment, with items reflecting data literacy abilities; and (3) information literacy-related courses, with specific course items (including mastery level and the auxiliary role of lab sessions). The latter two parts used a matrix scale format with a 5-point Likert scale: for the second part, 1 = "very uncharacteristic," 2 = "somewhat uncharacteristic," 3 = "neutral," 4 = "somewhat characteristic," and 5 = "very characteristic"; for the third part, 1 = "very poor," 2 = "poor," 3 = "average," 4 = "good," and 5 = "very good." Specific content is shown in Table 3 .

2.1.2 Sample Profile Statistical analysis of personal information from the 192 valid questionnaires revealed 87 female respondents (45.31%) and 105 male respondents (54.69%), with relatively uniform distribution across grade levels, indicating strong representativeness. Details are shown in Table 4 .

2.1.3 Reliability and Validity Analysis (1) Reliability. This study involved five dimensions: data awareness, data ethics, data acquisition, data processing, and data analysis and utilization, labeled A-E respectively. Using SPSS 22 software, we measured data reliability quality with Cronbach's alpha coefficient. An alpha value above 0.8 indicates high reliability; between 0.7-0.8 indicates good reliability; between 0.6-0.7 indicates acceptable reliability; and below 0.6 indicates poor reliability. The comprehensive reliability analysis results are shown in Table 5 . The alpha coefficients for all six dimensions exceeded 0.6, with the minimum being 0.664, indicating good reliability quality and authentic, reliable research data.

(2) Validity. This study used KMO test and Bartlett's sphericity test for validity analysis. When the KMO test coefficient exceeds 0.5 and the chi-square statistic significance level P-value of Bartlett's sphericity test is less than 0.05, each latent variable has convergent and discriminant validity, indicating good construct validity. The initial analysis of six dimensions and 17 scale items yielded the results shown in Table 5. The overall KMO test coefficient was 0.808, and the chi-square statistic significance level P-value of Bartlett's sphericity test was strictly less than 0.05, within the ideal range, confirming the validity of the data sample for this research model.

2.2 Results Analysis

2.2.1 Analysis of Data Literacy Level Differences Across Grades To determine data literacy levels across grades, we first compared the mean values of each data literacy dimension by grade level, as shown in the line graph in Figure 1 [Figure 1: see original paper]. Analyzing the overall trend across dimensions, we found that students' data literacy levels showed an upward trend with increasing grade level, particularly for data awareness and data acquisition dimensions. This suggests that as grade level increases, students take more relevant courses and accumulate more data literacy knowledge. However, the limited improvement across dimensions indicates that while current courses have some effect on cultivating data literacy, the impact is limited, suggesting that curriculum construction should strengthen courses related to data ethics, data processing, and data utilization capabilities.

Figure 1 reveals that although higher-grade students have taken more relevant courses than lower-grade students, there is no significant difference in the data ethics dimension across grades. The average data ethics scores remain between 3.2-3.4 for all grades, generally at an upper-medium level, indicating that data ethics does not improve through relevant coursework. We believe this result suggests that data ethics formation may come from learning or training outside the curriculum, leaving considerable room for improvement in how courses cultivate data ethics. In contrast, data awareness and data acquisition capabilities clearly strengthen with grade level, reaching level 4, indicating that relevant courses effectively cultivate students' data awareness and improve data acquisition skills. However, data processing and data analysis capabilities show no significant improvement with grade level, with scores not reaching level 4, suggesting that curriculum construction should consider improving courses related to data processing and analysis.

2.2.2 Laboratory Course Support Based on questionnaire results, we obtained student ratings of the degree to which lab sessions support mastery of corresponding course knowledge, with average scores shown in Figure 2 [Figure 2: see original paper]. The results indicate that supporting lab sessions significantly facilitate students' mastery of course knowledge, with all ratings at above-average levels. This suggests that data literacy-related courses should also include corresponding lab sessions. However, the facilitation effects did not reach level 4, indicating that the current lab session setup needs improvement, such as increasing investment in experimental facilities, improving the lab environment, and adding lab hours.

2.2.3 Correlation Analysis Between Data Literacy and Related Courses Pearson correlation coefficient is widely used to measure the degree of correlation between two variables. This study used SPSS 22 software to conduct Pearson correlation analysis to examine the relationship between information literacy-related courses and data literacy dimensions. Pearson correlation coefficient values reflect correlation strength, as detailed in Table 6 [19]. The correlation analysis results are shown in Table 7 . The findings reveal that most courses have relatively high correlations with relevant data literacy dimensions, except for "Modern Information Technology Foundation" and "Information Resource Management," whose maximum correlation coefficients with data literacy dimensions only reached 0.179 and 0.192, respectively, below 0.2. To refine the data literacy curriculum, we excluded these weakly correlated courses, retaining the remaining nine courses as the data literacy curriculum.

2.2.4 Cluster Analysis of Data Literacy-Related Courses To further clarify the closeness between the identified data literacy courses and various data literacy dimensions, we used the correlation coefficients calculated above as independent variables and performed cluster analysis on the 11 course samples using SPSS 22 software. This analysis grouped courses with similar impacts on data literacy dimensions together for modular classification. Cluster analysis only groups courses with relatively close distances and does not indicate that a course cannot simultaneously affect two dimensions. For example, "Information Storage and Retrieval" may impact data acquisition but is more closely related to data awareness, so it was classified as a data awareness course. We used the between-groups linkage method for clustering and Euclidean Distance for distance measurement, with results shown in Figure 3 [Figure 3: see original paper].

Based on the clustering results, we modularized courses according to data literacy dimensions. We believe that data awareness and data ethics are foundational dimensions of data literacy and prerequisites for data acquisition, processing, and analysis. Moreover, courses related to these two competencies share considerable commonality. Therefore, we unified courses related to these two dimensions into one category: the Data Awareness and Ethics Training Module. Figure 3 shows that courses related to data acquisition and data processing are relatively close in distance, indicating their interconnectedness, so we combined these two dimensions into one category named the Data Operation Module. In summary, we divided the clustering results into three modules: Data Awareness and Ethics Module, Data Operation Module, and Data Analysis and Utilization Module. Based on the courses included in each category, the module characteristics are summarized as follows:

First Category: Data Awareness and Ethics Module. This module aims to cultivate students' data awareness, data management, data ethics, data security, and big data fundamentals. Main courses include "Information Storage and Retrieval," "Intellectual Property Protection," and "Network Information Security."

Second Category: Data Operation Module. This module aims to develop students' data collection and storage capabilities, data processing and presentation skills, including learning relevant data collection and processing software. Main courses include "Data Structure," "Database Principles," "Informetrics," "Data Mining," "Information Visualization," and "Statistics."

Third Category: Data Analysis and Utilization Module. This module aims to cultivate students' data analysis and value discovery abilities, helping them understand various method application scenarios and draw conclusions based on analytical results. Courses include "Information Analysis and Prediction" and "Information System Analysis and Design."

3. Construction of Data Literacy Curriculum Group

3.1 Supplement of Data Literacy Courses

Given that Figure 1 shows no significant difference in data ethics dimension scores across grades, and that improvements in students' data processing and data analysis abilities are not substantial, there remains considerable room for improvement and perfection in the current curriculum setup. Therefore, based on previous research and referencing the information literacy-related course offerings in other universities' information management programs (see Table 1), and after consulting six teachers and experts in our university's relevant programs, we supplemented and improved the curriculum modules constructed above by adding data literacy-related courses, as shown in Table 8 .

Additionally, based on the analysis of lab session support, which shows that lab sessions significantly facilitate course knowledge mastery, we believe that, without affecting basic class hours, corresponding lab sessions should be appropriately offered in conjunction with theoretical courses to improve students' data operation abilities. Furthermore, students' data management practical skills should be trained through laboratory operation skills training. Universities should establish advanced data literacy operation training experimental centers to provide supporting experimental resources for theoretical teaching, enabling students' data literacy to continuously improve and steadily advance through practical teaching.

3.2 Structural Model of Data Literacy Curriculum Group

The data lifecycle refers to the process from data collection (or formation), processing, preservation, dissemination, retrieval, access, and utilization to disappearance or disuse [17]. In the data lifecycle, certain logical relationships exist from data acquisition to data utilization: data processing is driven by data collection, data processing serves data analysis, and the purpose of data analysis is to discover data value. Data awareness and ethics are the basic prerequisites for the data acquisition-processing-analysis process. That is, data acquisition must first be based on data awareness and must comply with data ethics—data must be obtained through legitimate channels, data cannot be arbitrarily altered during processing, and data sources must be cited during utilization. Otherwise, subsequent data processing, analysis, and utilization cannot proceed normally or will be affected. Therefore, the construction of data literacy curriculum groups should follow the logical relationships of the data lifecycle for orderly course arrangement.

The setup, content, teaching models, and sequencing of data literacy courses are shown in Table 9 . In the big data era, people's demand for data is more urgent. As the main front for education, universities should strengthen data literacy education and cultivate students' ability to recognize and 挖掘数据价值. This study takes information literacy-related courses in the information management major as the foundation, uses correlation analysis to examine the relationship between information literacy courses and data literacy, extracts courses related to data literacy, and employs cluster analysis to modularize data literacy courses and construct a data literacy curriculum group. Based on current curriculum deficiencies, we improved the data literacy curriculum group and ultimately constructed it according to data lifecycle theory. Therefore, we offer the following recommendations for the construction and implementation of university data literacy curriculum groups: (1) When implementing data literacy education using this curriculum group, programs should consider students' disciplinary backgrounds and foundational conditions. For example, given significant differences among liberal arts, science, and engineering disciplines, each discipline can appropriately adjust this curriculum group according to professional needs to construct a data literacy curriculum group suitable for their students. (2) The structure of data literacy curricula should also be updated in real-time according to changing era demands.

References

[1] Xi Jinping: Further Emancipate the Mind, Deepen Reform, and Implement Work to Promote New Breakthroughs in Comprehensively Deepening Reform at a New Starting Point [EB/OL]. [2018-08-09]. http://www.xinhuanet.com/politics/2018-01/23/c_{1122303539}.htm.

[2] Koltay T. Facing the Challenge of Data-Intensive Research: Research Data Services and Data Literacy in Academic Libraries [M]. Eger: Innovation in Libraries and Information Services, 2016.

[3] Stephenson E, Caravello P S. Incorporating Data Literacy into Undergraduate Information Literacy Programs in the Social Sciences: A Pilot Project [J]. Reference Services Review, 2007, 5(4): 525-540.

[4] Meng Xiangbao, Li Aiguo. Research on Scientific Data Literacy Education in Foreign University Libraries [J]. Journal of Academic Libraries, 2014, 32(3): 11-16.

[5] Huang Ruhua, Li Baiyang. Data Literacy Education: An Extension of Information Literacy Education in the Big Data Era [J]. Library and Information Science Knowledge, 2016(1): 21-29.

[6] Wei Lai, Wang Siming. Research on the Construction of Data Literacy

Course Content in Chinese University Libraries [J]. Information and Documentation Services, 2018(6): 90-95.

[7] Hogenboom K, Hayslett M. Pioneers in the Wild West: Managing Data Collection [J]. Library and the Academy, 2017, 17(4): 629-657.

[8] Carlson J R, Fosmire M, Miller C, et al. Determining Data Information Literacy Needs: A Study of Students and Research Faculty [J]. Libraries and the Academy, 2011(2): 629-657.

[9] Hu Hui, Wu Ming. Analysis of the Enlightenment of Information Literacy Education for Data Literacy Education [J]. Library Theory and Practice, 2016(11): 9-14.

[10] Long Qian. Construction of Data Literacy Competency Indicator System and Investigation and Analysis of Data Literacy Competency Status of University Faculty and Students [J]. Library, 2015(12): 51-56.

[11] Zhu Sha. Research on the Planning and Setup of Information Technology Curriculum Group for E-Commerce Major [J]. Computer Knowledge and Technology, 2018, 14(24): 271-272.

[12] Yang Zhibiao, Zhang Dan. Research on the Construction of University Information Literacy Education Curriculum System Based on Curriculum Groups [J]. China Management Informationization, 2017, 20(24): 247-248.

[13] Meng Xiangbao, Chang E, Ye Lan. Data Literacy Research: Origin, Status and Prospects [J]. Journal of Library Science in China, 2016, 42(2): 109-126.

[14] Zhang Jinliang, Li Baozhen. The Connotation, Value and Development Path of Teachers' Data Literacy in the Big Data Era [J]. e-Education Research, 2015, 36(7): 14-19.

[15] Mandinach E B, Cummer E S. A Systemic View of Implementing Data Literacy in Educator Preparation [J]. Educational Research, 2013, 42(1): 32-37.

[16] Prado J C, Marzal M A. Incorporating Data Literacy into Information Literacy Programs: Core Competencies and Contents [J]. Libri, 2013, 63(2): 123-134.

[17] Zhang Jingbo. Data Literacy Education in the Big Data Era [J]. Science, 2013, 65(4): 29-32.

[18] Wang Xiaodong, Wu Yaqin, Wang Husheng. Exploration of Information Management and Information Systems Major Construction Centered on Information Talent Cultivation [J]. Journal of Inner Mongolia Medical College, 2012(8): 539-544.

[19] Fu Zhigang. Statistics [M]. Shanghai: Economic Management Publishing House, 2015.

Author Contributions

Shen Jiujiu: Responsible for topic planning and paper revision;
Xu Ping: Paper writing and revision;
Zhang Qin: Questionnaire collection and statistics;
Gong Huaping: Paper revision.

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