Hot spot analysis of big data visualization research in China based on co-word network

Menglin Liang [1, a], Yu Liang [1, b]

[1] School of Shipping Economics and Management, Dalian Maritime University, Dalian, China,116026.

[a] liangmenglin2022@163.com, [b] liangyu@dlmu.edu.cn

Abstract. Big data visualization is a rapidly developing field, and various new technologies and methods continue to emerge, making big data visualization widely used in various fields. With the advancement of technology and the expansion of application scenarios, big data visualization will continue to play an important role in the future. This paper adopts the co-word network analysis method to conduct bibliometric analysis of the relevant research of big data visualization in China from 2012 to 2022, identifies the hot spots and connection points in the field of big data visualization in China by drawing a knowledge network map, shows the research structure of big data visualization through visual analysis, and predicts the future research development trend on this basis. The results show that seven knowledge points, such as "data analysis", " visualization analysis" and "big data technology", are the current research hotspots in the field of big data visualization, which determine the development direction of future research.

Keyword: Visualization analysis; big data visualization; co-word analysis; social network analysis.

1. Introduction

In recent years, big data visualization has been applied in many fields, since 2000, the rapid development of the Internet, big data has become high-value information, but due to the limitations of technology, lack of high-tech achievements [1]; for the study of visualization, scholars mostly use the methods of co-word analysis and social network analysis [3,4,] to explore the current situation of data visualization.

Although big data visualization is the hottest research topic at present, most of the domestic works on data visualization are translated foreign literature, and high-value academic papers are often aimed at data mining, algorithms and other aspects [14]. Therefore, it is necessary to use quantitative analysis methods to carry out systematic structured analysis of research related to big data visualization [5], analyze the internal relationship between knowledge topics and knowledge points in the field, and visualize the research results. Based on this, this study adopts the bibliometric method to analyze the keywords of the representative big data visualization research in China in the past ten years, and identifies research hotspots and discipline development trends by drawing a knowledge network map.

2. Research methods and data processing

2.1 Research methods

Bibliometrics is one of the most effective theoretical methods for studying discipline structure and predicting discipline development trends, mainly including co-word analysis [9] and co-citation analysis [10]. The method selected in this study is co-word analysis, which conducts bibliometric analysis of representative research on big data visualization in China, focusing on the research hotspots and structure of domestic big data visualization and the future development direction.

In terms of specific research design, factor analysis and multi-dimensional scale analysis are first used to conduct a preliminary analysis of the research structure of domestic big data visualization.

After that, the centrality and connectivity analysis of the big data visualization network is carried out to identify research hotspots and information connectivity points.

2.2 Keyword extraction and normalization

Step 1: Sample selection. In this study, core journals recognized by the National Natural Science Foundation of China were used as research data sources, and literature related to big data visualization research published in the past ten years was selected as research samples. With the theme of "big data visualization", a conditional search was carried out in CNKI database, and 698 documents that met the requirements were obtained.

Step 2: Keyword pre-statistics. First, the literature without keywords was removed, leaving 694 articles. Secondly, use NoteExpress software to count the number and frequency of keywords.

Step 3: Normalize relatively high-frequency keywords. This step is mainly divided into five stages: relatively high-frequency keyword interception, invalid word elimination, similar keyword merging, compound keyword splitting, and low-frequency keyword classification.

After the processing of the above steps, we finally obtained 44 keywords (Table 1), with a total frequency of 2303 occurrences, accounting for 71.30% of the total frequency of occurrences, and an average of 52.34 occurrences per keyword, basically covering the frontier and focus of domestic big data visualization research in the past decade. Therefore, these 44 keywords were selected as the research objects of co-word analysis.

Table 1. Summary table of important keywords in domestic big data visualization research

keyword	Frequency	keyword	Frequency	keyword	Frequency
Big data	407	Information	37	Citespace	23
Data visualization	543	visualization	35	Time-varying	21
Data analysis	114	Data model	34	data	
Visualization	91	Data security	33	Visualization	20
analysis	89	monitoring	32	Smart Library	20
Data journalism	82	Machine learning	31	Visualization of	
Big data	66	Data auditing	31	spatial	19
technology	63	Interactive	30	information	16
Data mining	59	Transportation	30	Urban	16
Intelligent	57	Data management		management	15
computing	45	Visualization of	28	Smart cities	13
Artificial	43	geographic	27	Virtual reality	11
intelligence	43	information	26	Data fusion	11
Data processing	41	Database		Data acquisition	10
Visual presentation	39	Business	25	Data literacy	9
Graph analysis		intelligence	24	Programming	8
Visualization		High-dimensional	23	language	
system		data Visualization		Data sharing	
Social media		Bibliometrics		Data storage	
Public services		Knowledge graph		Swarm	
		System		intelligence	
		architecture			

2.3 Data preparation

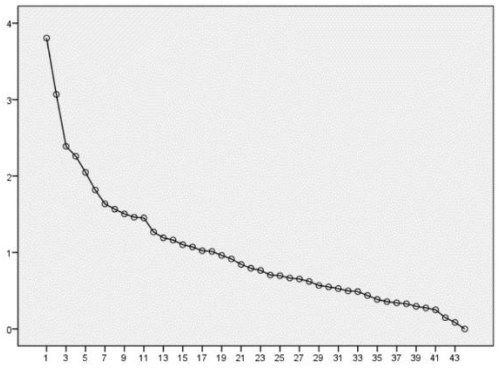
Apply the VBA programming technology of Excel 2003 to generate a co-word matrix of high-frequency keywords, the co-occurrence matrix is an undirected symmetry relationship matrix, and the value on the diagonal represents the word frequency of the keyword; On this basis, a logical matrix of high-frequency keywords is generated; According to the Ochiia coefficient [14], the co-word matrix is converted into a correlation matrix, and then the difference matrix is obtained. The co-word matrix, logical matrix, and difference matrix of high-frequency keywords are the data basis for big data visualization co-word network analysis.

$$O_{AB} = \frac{C_{AB}}{\sqrt{C_A} \times \sqrt{C_B}} \quad (1)$$

(O_{AB} is the correlation coefficient between A and B,C_{AB} is the number of simultaneous occurrences of A and B, ,C_A is the number of times A appears, C_B is the number of times B appears)

3. Factor analysis for big data visualization research

According to the 44 keywords selected in Table 1, SPSS software was used to factor analyze 44 high-frequency keywords, and the result was that 14 factors were formed, with a cumulative variance of 70.101% (Fig. 2). The cumulative variance result is more than 70%, which is acceptable, that is, 44 keywords are divided into 18 categories. However, if we look closely at the gravel plot (Fig. 2), it is not difficult to find that from the eighth factor, the change of factors tends to be flat, the variance explained by the first 7 factors is relatively high, and the least squares method used in factor analysis is prone to bias during the analysis process, so when this study does cluster analysis, 44 keywords are clustered into 7 categories.



ingredients	Initial eigenvalue			Extract the sum of squares to load		
	total	variance %	cumulation%	total	variance %	cumulation%
1	3.805	8.649	8.649	3.805	8.649	8.649
2	3.068	6.973	15.622	3.068	6.973	15.622
3	2.389	5.430	21.052	2.389	5.430	21.052
4	2.259	5.134	26.186	2.259	5.134	26.186
5	2.048	4.654	30.840	2.048	4.654	30.840
6	1.818	4.132	34.972	1.818	4.132	34.972
7	1.636	3.717	38.690	1.636	3.717	38.690
8	1.566	3.560	42.250	1.566	3.560	42.250
9	1.505	3.421	45.671	1.505	3.421	45.671
10	1.462	3.324	48.995	1.462	3.324	48.995
11	1.452	3.301	52.295	1.452	3.301	52.295
12	1.268	2.883	55.178	1.268	2.883	55.178
13	1.192	2.709	57.887	1.192	2.709	57.887
14	1.163	2.643	60.530	1.163	2.643	60.530
15	1.102	2.505	63.035	1.102	2.505	63.035
16	1.073	2.438	65.473	1.073	2.438	65.473
17	1.023	2.324	67.797	1.023	2.324	67.797
18	1.014	2.304	70.101	1.014	2.304	70.101

Fig. 2 Stone plot and factor analysis

Multi-dimensional scale analysis is done on the basis of cluster analysis. Results of multidimensional scale analysis:

Stress =0.41398； RSQ =0.13166

In multidimensional scale analysis, the RSQ value is 0.13166, and the stress value is 0.41398. It is generally believed that RSQ greater than 0.6 indicates that the degree of fitting between the analysis results and the actual data is acceptable ; A stress less than or equal to 0.1 indicates that the analysis results are very good, and a stress greater than or equal to 0.15 is unacceptable . The results of multidimensional scale analysis show that although the 44 keywords meet the critical value in the spatial dimension, the difference between them is not obvious, that is, there are more overlapping parts. Therefore, in order to clarify the research context of big data visualization in China at this stage, in-depth knowledge network analysis is required.

4. Knowledge network analysis for big data visualization research

Based on social network analysis (SNA) [12], UCINET 6.0 was applied to analyze the co-word matrix of keywords and draw a knowledge network map for big data visualization research (Fig. 4). According to the location characteristics of each keyword node in the network, the importance of each knowledge point and the correlation between knowledge points are analyzed, and then the current research hotspots and information connection points are analyzed.

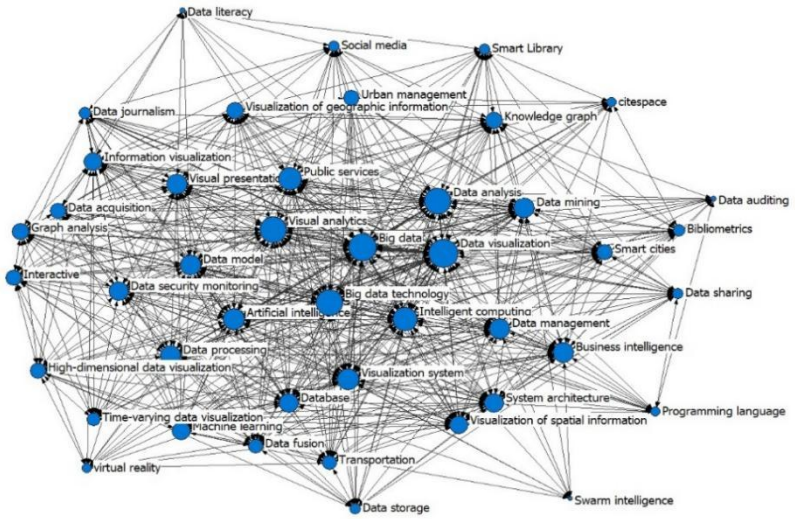


Fig. 4 Big Data Visual Research Knowledge Network Diagram

4.1 Network power distribution analysis

Social network analysis quantifies the power of knowledge points in big data visualization knowledge network according to the co-word strength and co-word category, and the main indicators are the degree centrality of nodes and the degree center potential of the network. By calculating the degree of degree centrality of each node, the knowledge points in the forefront of power are obtained (Table 5). Among them, the degree centrality of "Big data" in the first place and "Data visualization" in the second place are 950.000 and 927.000 respectively, which have strong control and influence in the network.

Table 5. Degree centrality of keywords

Ranking	keyword	Absolute degree centrality	Relative degree centrality
1	Big data	950.000	11.507
2	Data visualization	927.000	11.228
3	Big data technology	303.000	3.670
4	Data analysis	292.000	3.537
5	Visualization analysis	250.000	3.028
6	Intelligent computing	235.000	2.846
7	Visualization system	183.000	2.217
8	Data journalism	166.000	2.011
9	Data processing	162.000	1.962
10	Data mining	154.000	1.865
11	Artificial intelligence	153.000	1.853

At present, the central trend of knowledge network degree of big data visualization research is 3.2717%, and the network concentration trend is weak, indicating that the research on big data visualization in China is still relatively scattered at this stage. Combined with the network density, the current network density value is 8.8585%, and the network is sparse and the scale is low, indicating that there are a large number of knowledge points in the field of big data visualization that have not yet been excavated, and further research is needed.

4.2 Network connectivity node analysis

In a big data visualization knowledge network, the degree value of a node represents the number of relationships it has, that is, the kind of common word that the keyword has. By analyzing the keyword logic matrix, 44 keywords coexisted in 575 undirected relationships, and the average node value was 13.068. The results of the test of its distribution (Fig. 6) show that the degree distribution of 44 high-frequency keywords conforms to the power-law distribution [9].

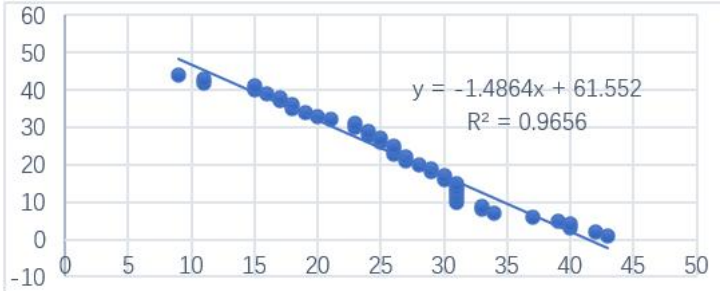


Fig. 6 Distribution of degree values

The degree value of high-frequency keywords obeys the power law distribution, indicating that the current domestic big data visualization knowledge network is a scaleless network, and there are some core knowledge points with extensive connections in the network, and their existence makes indirect connections between some unrelated knowledge points , which are the information connection points of the development and evolution of big data visualization research and have important research value. In order to identify the connectivity points in the analysis network, the intermediate centrality and near-centrality analysis of the logical matrix of keywords were performed (Table 7).

Table 7: Intermediate and Near Centrality of Keywords (excerpt)

serial number	keyword	Intermediate centrality	Proximity to centrality
1	Data visualization	67.925	43.000
2	Big data	49.150	45.000
3	Big data	36.652	50.000
4	technology	32.041	50.000
5	Data analysis	22.258	53.000
6	Visualization	17.562	57.000
7	analysis	15.915	58.000
8	Intelligent	14.540	61.000
9	computing	13.429	58.000
10	Data management	12.088	52.000
	Business intelligence		
	Visualization system		
	Public services		

The existence of connectivity points provides the possibility for the integration and development of various branches of research in the field of big data visualization research, and the interdisciplinary zone is the high incidence range formed by new theories and new methods, so these connection points are very likely to develop into hot topics in the future of big data visualization research.

5. Knowledge subnetwork analysis

Although the current cross-overlapping of various branches of the field of big data visualization is more common, and the research structure with clear boundaries cannot be formed, according to the nature of network nodes, nodes with strong power and information control capabilities are taken as the "key points", and "subnets" are established with them as the center, which reflect the aggregation trend of knowledge points and represent the current domestic research theme group of big data visualization to a certain extent.

Compared with Table 6-7, 7 key nodes with greater power and strong information control capabilities in the network were screened out, Since “Big Data” and “Data Visualization” coincide with the research theme, the concepts of “data analysis” and “data mining” were proposed earlier, so they did not have the value of in-depth analysis, so other keywords were selected for analysis., so

other keywords were selected for analysis, namely: “Visualization analysis”, “Big data technology”, “Intelligent computing”, “Artificial intelligence”, “Business intelligence”(Fig. 9)

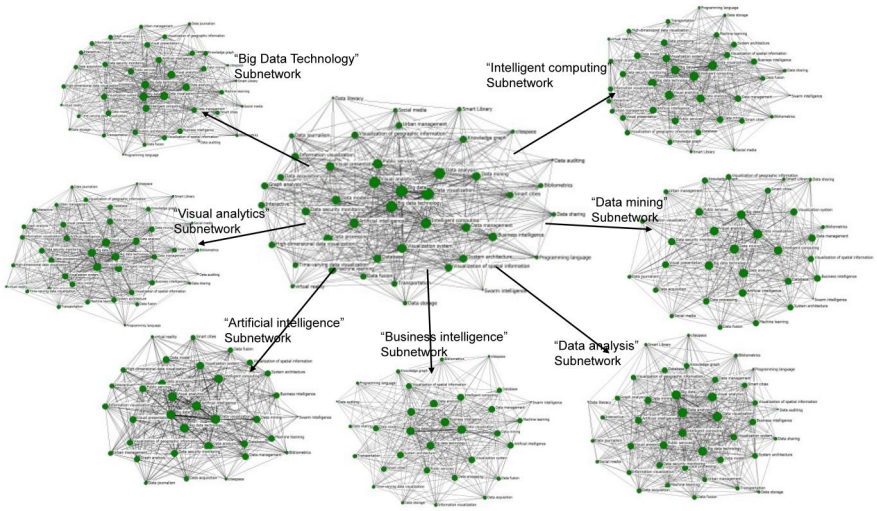


Fig. 9 Schematic diagram of the "keypoint" subnetwork

5.1 Visualization analysis study topic Subnetwork Analysis

From the perspective of subnet analysis, the visual analysis of big data is still the most concerned hot spot in the field of visual analysis. Scholars apply more visual analysis to the field of transportation, including the planning, selection, optimization of transportation trajectories, traffic visualization is not only a chart, but also the guarantee of urban transportation.

Table 10. “Visualization analysis” Co-word Situation (excerpt)

keyword	frequency	keyword	frequency	keyword	frequency
Big data	65	Transportation	13	Information	8
Data visualization	16	Citespace	10	visualization	8
Data analysis	13	Intelligent computing	8	Bibliometrics	6
				Business intelligence	

5.2 Big data technology research topic subnetwork analysis

From the perspective of co-words, data analysis and data processing are currently important areas of interest in the research theme of big data technology. The application of big data technology covers a variety of fields, especially in the current era of Industry 4.0, intelligent computing, artificial intelligence, data journalism and other hot fields are also inseparable from big data technology.

Table 11. “Big data technology" Co-word Situation (excerpt)

keyword	frequency	keyword	frequency	keyword	frequency
Big data	51	Data processing	17	Data auditing	9
Data visualization	47	Intelligent	9	Artificial intelligence	8
Data analysis	20	computing	9	Data journalism	8
		Graph analysis			

5.3 Intelligent computing research topic subnetwork analysis

Computing has become a key factor in formulating and promoting social development. In the new era of digital civilization of the Internet of Everything, traditional data computing can no longer meet the current needs, and the combination of computing and artificial intelligence has become a frontier field. Technology is the foundation of intelligent computing, so the study of

intelligent computing is inseparable from big data technology; Transforming data into intelligent information can effectively help enterprises make strategies and decisions, and business intelligence has also become a hot topic in intelligent computing research.

Table 12. “Intelligent computing” Co-word Situation (excerpt)

keyword	frequency	keyword	frequency	keyword	frequency
Big data	37	Data mining	12	Visualization system	10
Data visualization	37	Artificial	11	Big data technology	9
Data analysis	15	intelligence	10	Business intelligence	8
		Data processing			

5.4 Business intelligence research topic subnetwork analysis

From the distribution of high-frequency words, business intelligence research is mainly focused on artificial intelligence and visualization systems, artificial intelligence application fields involve a wide range, with the further development of technology artificial intelligence is also gradually commercialized; Business intelligence is mainly the use of software and services to informatize data, visualization system is an important carrier of business intelligence, so the research on visualization system can promote the development of business intelligence.

Table 13. “Business intelligence” Co-word Situation (excerpt)

keyword	frequency	keyword	frequency	keyword	frequency
Data visualization	19	Visualization	7	Intelligent	6
Big data	18	system	6	computing	6
Artificial	9	Data mining	6	Big data	4
intelligence		Data analysis		technology	
				Data auditing	

5.5 Artificial intelligence research topic subnetwork analysis

The research of artificial intelligence is mainly algorithms, data processing applications, how to combine artificial intelligence into business is an important research topic; In addition, intelligent algorithms in swarm intelligence research are also the focus of artificial intelligence research.

Table 14. “ Artificial intelligence” Co-word Situation (excerpt)

keyword	frequency	keyword	frequency	keyword	frequency
Data visualization	25	Visualization	11	Data mining	7
Big data	18	system	9	Swarm	7
Intelligent	11	Business	8	intelligence	4
computing		intelligence		Interactive	
		Big data			
		technology			

6. Conclusion

In the research on big data visualization in domestic journal literature, keywords such as big data, data visualization, visual analysis, data mining, data mining, data analysis will appear in many years from 2012 to 2022. Among them, the concepts of data visualization, visual analysis, data analysis and data mining are put forward earlier, and relevant researches tend to be mature. The advent of big data combines data analysis, data mining and visual analysis to form the concept of data visualization. This also lays a foundation for the research of big data visualization in other fields, such as artificial intelligence, data security monitoring, business intelligence, etc.

The results show that although scholars have carried out a lot of research in the field of big data visualization, many aspects are still in their infancy. Big data visual analysis aims to use computer

automation analysis capabilities at the same time, fully tap the cognitive advantages of people for visual information, organically integrate the respective strengths of man and machine, and assist people to understand the information, knowledge and wisdom behind big data more intuitively and efficiently with the help of human-computer interactive analysis methods and interactive technology, and then related data processing foundations such as data mining and data analysis have become hot words; China has increased the visual analysis of big data on social media, and paid more attention to the application of big data visualization in artificial intelligence and business intelligence.

In the current knowledge network, seven knowledge points of “big data technology”, “data analysis”, “intelligent computing”, “artificial intelligence”, “business intelligence”, “data mining” and “visualization analysis” are the current hot spots in big data visualization research. Among them, the influence and control of “big data technology” and “data analysis” in the network far exceed the rest of the nodes, and are the hot spots in the current research hotspots of big data visualization.

Funding: This research was funded by [National Natural Science Foundation of China] grant number [72104043], and by [China Postdoctoral Science Foundation] grant number [2022M710571].

References

[1] CHEN Jun,XIE Weihong,CHEN Yangsen,et al. Comparative Study on Big Data Visualization Academic Papers at Home and Abroad——Based on Bibliometric and SNA Methods[J].Science and Technology Management Research,2017,37(08):44-53.

[2] REN Lei,DU Yi,MA Shuai,et al. Review of visual analysis of big data[J].Journal of Software,2014,25(09):1909-1936.

[3] WANG Aizhu,MA Yan,XIANG Zhu. Visual analysis of domestic education big data based on knowledge graph[J].digital education,2020,6(06):28-32.

[4] MIAO Linlin,CHEN Jingjing. Research on Investigation Theory Based on Visual Analysis of CiteSpace Knowledge Graph[J].Journal of Guangxi Police College,2021,34(05):74-85.

[5] LYU Yibo,CHENG Lu. Structural Analysis of Innovation Management Research in China Based on Co-Word Network[J]. Journal of Management, 2011, 8(10): 1541-1548.

[6] ZHANG Qin,MA Feicheng. Foreign Knowledge Management Research Paradigm: Co-word Analysis as a Method[J]. Journal of Management Science, 2007,12(6)：65-75.

[7] J Law , S Bauin , et al. Policy and the mapping of scientific change: A co-word analysis of research into environmental acidification [J]. Scientometrics, 1988, 14(3/4)： 251~264.

[8] S Bauin, B Michelet, et al. Using Bibliometrics in Strategic Analysis：Understanding chemical reactions at the CNRS [J]. Scientometrics, 1991,22(1)：113~137.

[9] M Callon, JP Courtial, etal. Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry [J]. Scientmetrics, 1991, 22(1)：155~205.

[10] JM Meyer, AC Heath, et al. Using multidimensional scaling on data from pairs of relatives to explore the dimensionally of categorical multifactorial traits[J]. Genetic Epidemiology，1992(9)：87-107.

[11] JB. Kruskal, M Wish. Multidimensional scaling[M]. Newbury Park, 1978.

[12] S Wasserman, K Faust. Social Network Analysis[M]. The Press Syndicate of Cambridge，1994.

[13] WANG Aizhu,MA Yan,XIANG Zhu. Visual analysis of domestic education big data based on knowledge graph[J].digital education,2020,6(06):28-32.

[14] TAN Juan,YANG Yanan. Research Status of Data Visualization Technology——Based on the Perspective of Social Network Analysis[J].Value Engineering,2018,37(36):215-216.

[15] GENG Zhijie,WANG Wennai. Reasons for Power Law Distribution Characteristics of Citation Network[J]. Intelligence Magazine 2009(28)：15-17.