Research on knowledge mapping construction of big data

Xiulian Jiang¹, Peng Ju¹, Yining Song²,³, Lingcheng Mao¹ and Guifang Xie¹

¹Xuzhou University of Technology, Xuzhou, Jiangsu, 221008, China
²University of Chinese Academy of Sciences, Beijing, 100000, China
³Xinjiang Astronomical Observatory, CAS, Urumqi, Xinjiang, 830011, China

Corresponding author and e-mail: Jiang Xiulian, 447502622@qq.com

Abstract. The diversified development of network applications and the informatization of social life are triggering explosive growth of data. The core journal articles based on big data research in the 2011-2018 database, which are included in the CNKI database, are used as research objects, and CiteSpace, Netdraw, and Ucinet are used as tools for visualizing knowledge maps. Visualize the growth trend of big data literature, authors and institutional cooperation networks; use co-occurrence analysis and cluster analysis to mine research hotspots of big data; use mutation word detection and time zone view to study evolution path of big data development. The research hotspots of big data are shown in three directions of big data technology, application field and challenges brought by big data. The evolution path of the field of big data shows that in early stage, technology research related to big data was mainly conducted, and later, it developed into the research trend of technology and application.

1. Introduction

From a macro perspective, big data is a bridge connecting the physical world, the information world and human society. From the perspective of information industry, big data is also a powerful push for the new generation information technology industry. The definition given by the McKinsey Global Institute is: a collection of data that is large enough to capture, store, manage, and analyze much beyond the capabilities of traditional database software tools, with massive data scales, fast data flow, Diverse data types and low value density are four major features [1]. The knowledge map method is a process of visualizing the relationship of the object to be analyzed. That is to say, in scientometrics, the knowledge mapping method can be used to combine citation analysis with data and information visualization to form various patterns with multidimensional structural laws [2]. China's research on the construction of scientific knowledge maps began in 2004. The research team is best known for the WISE laboratory led by Professor Liu Zeyuan of Dalian University of Technology. At present, knowledge mapping applications are mainly concentrated in the disciplines of library and information, education, medicine, computer science, and management [3].

2. Basic knowledge map analysis of big data

2.1. Analysis of the law of document growth

The law of the growth of the literature reveals to some extent the growth law of the amount of scientific knowledge, which is closely related. Therefore, observing and analyzing the volume and variation of journals in the field of big data helps us to understand the research prospects and
development trends in this field. The publication volume statistics of core journals related to big data from 2011 to 2018 are shown in Table 1.

Table 1. Statistics on the publication of big data research papers in 2011-2018

| Publication year | Issue volume | Publication year | Issue volume |
|------------------|--------------|------------------|--------------|
| 2011             | 2            | 2015             | 1067         |
| 2012             | 21           | 2016             | 1286         |
| 2013             | 263          | 2017             | 1337         |
| 2014             | 701          | 2018             | 1231         |

Among the 5,908 related articles collected, the research on big data in China started around 2011, and only 2 papers were published in that year, which was consistent with the time when the concept of big data was formally proposed. In the two years from 2011 to 2012, there were only 23 papers on big data research, accounting for 0.39% of the total length of the eight-year literature. However, since 2013, research papers have been developed in a blowout manner. By 2015, the number of documents has exceeded 1,000, and big data research is on the rise. In the three years from 2015 to 2017, the growth rate of the papers slowed down and showed a steady growth. By 2018, the relevant research papers had a slight downward trend, but in terms of volume, the amount of relevant research papers was still considerable. It can be seen that the research on big data is not reduced, and big data is promoted to national strategy in comprehensive time. Internet+, artificial intelligence and blockchain have become hot topics, and it is not difficult to predict the research prospects of big data.

2.2. Author's total knowledge map analysis

The scientologists Katz and Martin define scientific cooperation as: scientific cooperation is the joint work of research scholars to produce new scientific knowledge [4]. In the actual process, the forms of scientific cooperation are diverse. The scientific cooperation mentioned in this article refers to the emergence of different authors in a paper. Then, there is a cooperative relationship between the authors. Through the analysis of the co-occurrence knowledge of the co-authors, it helps us to understand the intensity of cooperation between scholars in the field of big data in 2011-2018, and whether there are obvious co-subgroups.

This paper uses SATI (Documents and Statistics Information Analysis Tool) to calculate the author frequency in the sample literature, automatically generate a 100*100 author co-occurrence matrix, and then import the matrix into Ucinet software to generate a matrix format that Netdraw can recognize. Finally, it is drawn into the co-occurrence knowledge map of the author. The map is shown in Figure 1. The square nodes in the figure represent the key elements in the cooperation map. The larger the node, the higher the degree of criticality. Whether there is a cooperative relationship between nodes mainly depends on whether there is a connection between nodes, and the thicker the connection, the more cooperation between the two scholars, the closer the cooperation relationship. The links between are loose. Overall, the research team of 2-3 people is the main one, indicating that the researchers are not very concerned about the cooperation between groups, but there are also obvious co-groups. Academic research groups such as Wang Wei, Yu Feng, Zhang Hao, Lan Yuexin, Gao Hong, Liu Bingyue and other scholars. These scholars are mainly from Nanjing University, Harbin Institute of Technology, Jilin University, Tianjin University and other units. They are the backbone of the big data research community and guarantee the publication of high-quality academic research results.

2.3. Institutional co-existence knowledge map analysis

Research institutions can be understood to some extent as synonymous with academic groups. Analysis of core institutions helps to understand the distribution of key research forces in the big data arena. At the same time, the cooperation between institutions helps to realize the optimal allocation of scientific research resources, which is of great help to the improvement of scientific research production capacity. Therefore, the co-existing analysis of scientific research cooperation is also very necessary. This paper uses Excel and CiteSpace output files to calculate the institutional distribution of
the core journals in the 2011-2018 big data domain, which is a total of 5,778 domestic and foreign academic institutions. In order to better demonstrate the academic productivity of the big data field, the academic institutions of the statistics include first-level units and second-level units. This paper selects the top 10 academic institutions in the field of publications, which represent the core research and production institutions in the field of big data research. The distribution of core institutions is shown in Table 2.

### Table 2. Distribution of core institutions

| Serial number | Institution name                                      | Issue volume |
|---------------|-------------------------------------------------------|--------------|
| 1             | Wuhan University School of Information Management     | 56           |
| 2             | School of Journalism, Renmin University of China      | 53           |
| 3             | Chongqing University of Technology School of Accounting| 48           |
| 4             | Chinese Academy of Sciences University                | 43           |
| 5             | School of Information Management, Nanjing University  | 41           |
| 6             | Wuhan University Information Resources Research Center| 32           |
| 7             | School of Information Resource Management, Renmin University of China | 26 |
| 8             | Research Center for Journalism and Social Development of Renmin University of China | 19 |
| 9             | China Telecom Corporation Guangdong Research Institute | 18 |
| 10            | Chongqing Institute of Technology Cloud Accounting Institute | 18 |

The CiteSpace software was used to export the document information in the Reforks format as data preparation before visualization analysis. Before running the software, set the time zone to 2011-2018, the time slice is set to 1 year, the default “Threshold” is top 50, the node type is set to "institution", the relationship strength parameter selects the default value, click the "Go" button to debug Generate co-occurrence maps between agencies. There are 416 nodes in the map, 170 connections, and the network density is 0.002, indicating that there is less cooperation between organizations. This article selects the largest sub-network in the organization cooperation network, and its co-occurrence map is shown in Figure 2.

**Figure 1. Author co-occurrence map**

**Figure 2. The largest sub-network map of issuing organizations**

Through the largest sub-network map, it is not difficult to find that the group of the organization is mainly composed of the faculties and research institutes of well-known domestic universities such as Renmin University of China, Wuhan University, Tsinghua University, and Harbin Institute of Technology. The most prominent research institution is the School of Information Management of Wuhan University. From Table 2, we can also see that the cumulative number of publications reached 56, ranking first. Of course, Renmin University of China and Tsinghua University are also the backbone of research in the field of big data. Obviously, effective cooperation between academic
groups is the key to ensuring scientific research capacity, but on the whole, there is no particularly fixed research group in the domestic big data field. This is a problem that needs to be addressed and resolved in the future development of big data in China.

Studying the geographical distribution of academic institutions can help us judge the main positions in the field of big data research, sort out the benchmark areas, and conduct targeted learning exchanges. The main positions in the field of research big data are concentrated in the eastern coastal and economically developed areas such as Beijing-Tianjin-Hebei, Yangtze River Delta and Pearl River Delta, while the research on big data in the central and western regions of China is relatively rare. It can be seen that the distribution of academic institutions in the field of big data in China is relatively concentrated.

3. Research on Big Data Hotspot Based on Keyword Co-occurrence and Clustering Map

3.1 Keyword co-occurrence map
The co-occurrence analysis of the keywords is to analyze the keywords provided by the authors in the dataset, and to reflect them in a co-occurrence pattern, which can well reflect the current research hotspots in a certain research field and the research hotspots that have been generated in the past[5]. The author imported the 5908 data collected into the CiteSpace software. The node type selects keyword, the time slicing is set to 1 year, and the default “Threshold” is top 50. Using the path finding algorithm, click the “Go” button to debug. Generate a keyword co-occurrence map. There are 212 nodes in the map, 238 connections, and the network density is 0.0106. The software generated keyword co-occurrence map is shown in Figure 3.

In Figure 3, the larger the node, the more times the keyword appears, such as cloud computing, data mining, artificial intelligence, etc, which are relatively large keywords. To a certain extent, they embody the research in the field of big data. The centrality of the keyword is another indicator to measure the research focus. The higher the centrality of a keyword, the greater the probability that the keyword appears in the same document as other keywords. The key to higher centrality, the word is probably the topic that scholars focus on for some time [6]. Therefore, the combination of centrality and frequency is the best way to measure the importance of keywords.

The centrality of high-frequency keywords such as big data, cloud computing, data mining, Internet of Things, Hadoop, MapReduce, visualization, innovation, and smart city is relatively high, and comprehensive influence is relatively large, which can better reflect the research hotspots in the field of big data. Special attention is paid to cloud accounting, although its frequency is relatively high, but its centrality is 0, indicating that its comprehensive influence is lower than other high-frequency keywords.

3.2 Keyword clustering map
After using CiteSpace to obtain the co-occurrence map of the keywords, the author obtained the keyword clustering map through the software automatic clustering function. The clustering map is shown in Figure 4. The measurement of clustering maps mainly depends on two indicators: module value (Q value) and average contour value (S value). In general, Q value>0.3, S value>0.5, indicating the clustering effect is reasonable. In the clustering map generated, the Q value is 0.5999 and the S value is 0.8626. The clustering effect is reasonable. The map shows that this clustering divides keywords into 13 categories. The larger the number in the figure, the larger the cluster size, which is precision marketing, database, financial integration, challenge, information services, hadoop, cloud computing, blockchain, media convergence, data news, smart city, deep learning, Internet+. These cluster names can reflect the focus of scholars to a certain extent.
4. Research on Big Data Evolution Path Based on Mutant Words and Time Zone Diagram

4.1 Mutant words detection
Mutant words refer to “keywords or phrases whose word frequency contribution changes significantly in different time windows and show a sudden rise or fall [7].” The mutant words in the keyword and found it in 2011-2018. There were 33 mutant words in the year. The keyword with the highest mutation intensity is information service, reaching 8.8279. The frequency of use increased sharply between 2013 and 2015. It can be seen that information service has become a focus of scholars in the field of big data during this time. At the same time, the keywords with high mutation intensity include digital publishing, library, data mining, and internet finance. Mainly focused on the application of big data. At the same time, the keywords of online public opinion, small data, applications, and privacy. They reflect the research hotspots in the field of big data. The words mutated from 2016 to 2018. There is a mutant tendency for further. From the macro analysis of the time span of mutant words, in the early days of big data, the research hotspots mainly focus on big data computing frameworks such as Mapreduce, Hadoop, cloud computing and related technology levels. There are also nouns, such as data news, big data technology, big data era and so on. After that, the research focus in the field of big data quickly expanded to application areas, such as libraries, enterprises, commercial banks, and finance. In recent years, the application of big data has continued to receive attention, but the focus has also begun to change, such as more emphasis on data security, privacy and so on.

4.2 Time zone view
In order to further study the law of the evolution of key words in the big data field over time, to explore the evolution path of the big data research topic, the author uses the time zone view provided by CiteSpace for path analysis. The time zone view collects the keywords in the same time in the same time zone, where the same time refers to the time when the keyword first appears, and the time series are arranged in order from far to near. The time zone view obtained by this study is shown in Figure 5.

From Figure 5, we can see that in the initial stage of big data research, the research content mainly includes data mining, data analysis, mapreduce, cloud computing and other technical foundations. During 2012-2013, the research theme has grown dramatically, except with the introduction of new technical terms such as visualization, association rules, mobile Internet, etc., we also found terminology at the technical application level represented by Internet finance, libraries, information services, digital libraries, etc. In 2014, the application scope of big data continued to expand, and application fields represented by cloud accounting, online public opinion, digital publishing, and university libraries appeared. At the same time, we can also see that the research theme also involves
challenges, privacy protection and other aspects. The research scope of big data has been further broadened, and relevant scholars have realized the challenges and problems in the era of big data. In 2015, with the introduction of the national big data strategy and the Internet+, national governance appeared in the field of big data research, and scholars began to explore the integration of Internet+ and big data. At the same time, the research scope of big data has been extended to the ideological and political education in colleges and universities, and the innovative reform of ideological and political education in colleges and universities has begun to be explored. During 2016-2018, the frequency of new nodes began to decrease. In 2017, the integration of artificial intelligence and big data technology has been widely concerned by scholars, and the combination of big data and current affairs policies has become more and more obvious. At the same time, new technical level terms such as parallel computing and deep learning have also been proposed. From the evolution of the above topics, it is not difficult to see that the main content of big data research will continue to focus on the technical level and application level. The integration of big data and other technologies will also generate new kinetic energy for the economy and society.

5. Conclusions
Using Netdraw and CiteSpace as representative tools, the knowledge map construction of sample data is studied from perspectives of literature growth law, literature authors and research institutions. The conclusions are as follows. First, the volume of publications in the field of big data in 2011-2018 is generally on the rise, and the relationship between the volume of publications and time is linearly fitted. The fitting equation is $F(t) = -4.51E5 + 2.24E2 \times t$. Second, scholars in the field of big data in China are relatively loosely connected with the overall research team of 2-3 people. Third, domestic big data research groups are mainly concentrated in colleges and universities (including computer, information management and other departments) as well as university research centers (institutes, research institutes). The most prominent research institution is the School of Information Management of Wuhan University. There is less cooperation and the organization's distribution is relatively concentrated.

Acknowledgement
This research was supported by the Research on the Application of Big Data Technology in College Management under Grant 2017-R-59479.

References
[1] Xiao Ming, Kong Achievements. Big Data: Where to Go—Based on the Perspective of Bibliometrics[J]. Journal of Library Science, 2014, 36(11): 110-117.
[2] Wang Cenlan, You Jianxin. Review of Big Data Literature: Visualization Research Based on Software Citespace[J]. Science and Technology Management Research, 2017, 37(21): 180-189.
[3] Li Xin. Research on the construction of knowledge maps in management science and engineering disciplines [D]. Hubei University of Technology, 2016.
[4] Lei Shuiwang. Research Progress of Big Data at Home and Abroad Based on Knowledge Mapping[J]. Library and Information Research, 2017, 10(02): 62-69.
[5] Yang Ruixian. Bibliometric Analysis of Big Data Research[J]. Information Science, 2015, 33(08): 152-156.
[6] Hou Yuanyuan, Huang Yurong, Zhang Hong, Lu Huaqiao. Analysis of the Progress of Big Data Research in China Based on Bibliometrics[J]. Library and Information Service, 2014, 58(S2): 204-208.
[7] LIU Minjuan, ZHANG Xuefu, YAN Yun. Research on the subject evolution method based on core words, mutated words and new words[J]. Journal of Information, 2016, 35(12): 175-180.