World Scientific Research Journal Volume 5 Issue 10, 2019
ISSN: 2472-3703 DOI: 10.6911/WSRJ.201910_5(10).0025

Study on Knowledge Mapping in Big Data

Xu Zhao[1, a], Guang Liu[2, b, *]

[1]Library, Shaanxi Normal University, Xi'an 710119, China

[2]Network Information Center, Shaanxi Normal University, Xi'an 710119, China.

[a]zhaoxu@snnu.edu.cn, [b]liuguang@snnu.edu.cn

Abstract

In order to understand the current research situation and development context in the field of big data, in this paper, knowledge mapping of big data is used for analysis. The data of the web of science core collection database (1986-2019) are analyzed in terms of research hotspots, trends, key literature in the field, researchers, organizations and countries of researchers, and the style of journals. The analysis results show that the models, methods and algorithms of big data analytics, the application of big data analytics in industry 4.0, healthcare and other industries and the challenges it faces, and privacy of big data, are the research topic of continuous concern.

Keywords

Knowledge mapping, big data, research hotspots, research trends.

1. INTRODUCTION

Big data is the general term of all technologies for processing large amounts of data, including capture, transfer, storage, curation, search, analysis, visualization, security, and privacy. Big data appeared in our lives in the 1980s, and recently the topic about it becomes more and more hot in healthcare, industry 4.0 and many other applications [1].

There are thousands of research papers on big data in web of science. How to get the research hotspots and trends in the big data generated by these big data papers, and find papers that needs to be read carefully, is the research foundation in the field of big data.

This paper aims to solve three problems: (1) Knowledge mapping analysis on the research hotspots and trends in big data. (2) Knowledge mapping analysis on the key literature in big data. (3) Knowledge mapping analysis on the researcher and their organizations and countries and the source journals.

2. DATA SOURCES FOR KNOWLEDGE MAPPING

Select the web of science core collection database [2], using TI="big data" as the retrieval condition in the advanced search, and a total of 7924 data were retrieved, The timespan is all years (1986-2019), and data retrieval deadline is 31 July 2019.

3. KNOWLEDGE MAPPING

Firstly, data preprocessing of the downloaded data, including file merging and keywords de-duplication and so on. Then analyzing data based on knowledge mapping in VOSviewer. VOSviewer is a software tool for constructing and visualizing bibliometric networks. These networks may for instance include journals, researchers, or individual publications, and they

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201910_5(10).0025

can be constructed based on citation, bibliographic coupling, co-citation, or co-authorship relations [3].

3.1. Knowledge Mapping of The Research Hotspots and Trends Based on Co-Occurrence Relations

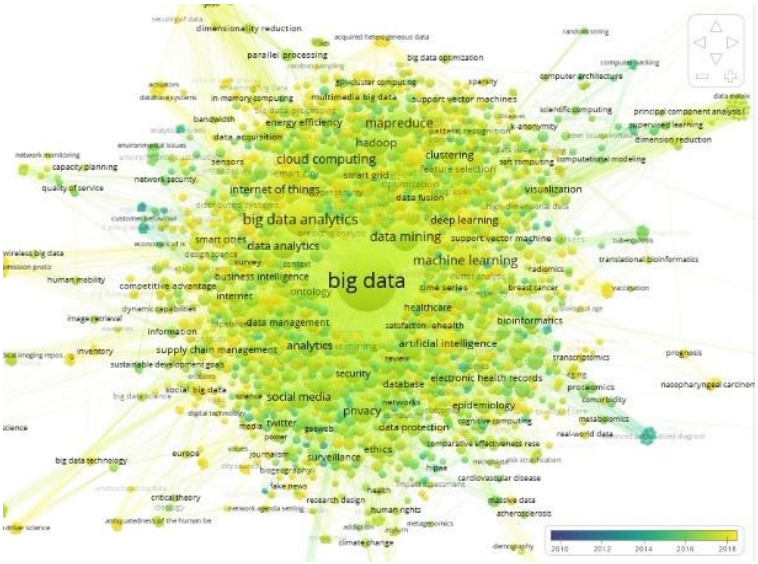

Fig 1. Overlay Visualization based on co-occurrence relations

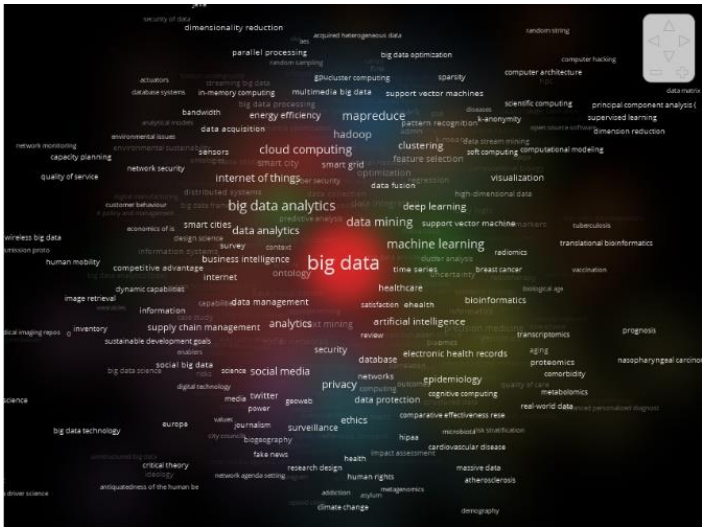

Fig 2. Density Visualization based on co-occurrence relations

Fig. 1 and fig. 2 are overlay visualization and density visualization which are constructed based on co-occurrence relations.

Each node in fig. 1 represents a keyword. The higher the frequency of keywords appear in all keywords in the sample paper data, the larger the circle of nodes. If two keywords appear in a paper at the same time, that means the two keywords have the co-occurrence relation, then a link will appear in the two nodes. The more co-occurrence relations occur in two keywords, the greater the total link strength is, and the link between the two keywords become thicker. The color of the node depends on the publication time of paper including the keywords, the newer the publication date is, the lighter and yellower the color is.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201910_5(10).0025

In fig. 2, the redder the color of the area where the keyword is located, the greater the density of the keyword, that means the keyword represents the research hotspot. Visualization in VOSviewer can be enlarged to see more keywords and reduced to see the whole graph.

The node that represent the research hotspots are obvious: big data, big data analytics, machine learning, cloud computing, data mining, MapReduce, Hadoop, internet of things, privacy, social media, artificial intelligence, data science, deep learning, and so on.

All keywords are clustered into 23 clusters. Links and total link strength of each keyword can get from the co-occurrence relation network. For instance, fig. 3 is the co-occurrence relation network of the keyword “big data”, “big data” is the hottest keyword which belongs to cluster 1 and has many links.

Fig. 4 is the co-occurrence relation network of the keyword “spark” which belongs to cluster 3. The average publication time of papers including the keyword spark is 2017, therefore, the color of the node is yellower. The spark node links to some keywords including Big data, big data analytics, machine learning, cloud computing, MapReduce, Hadoop, deep learning and so on.

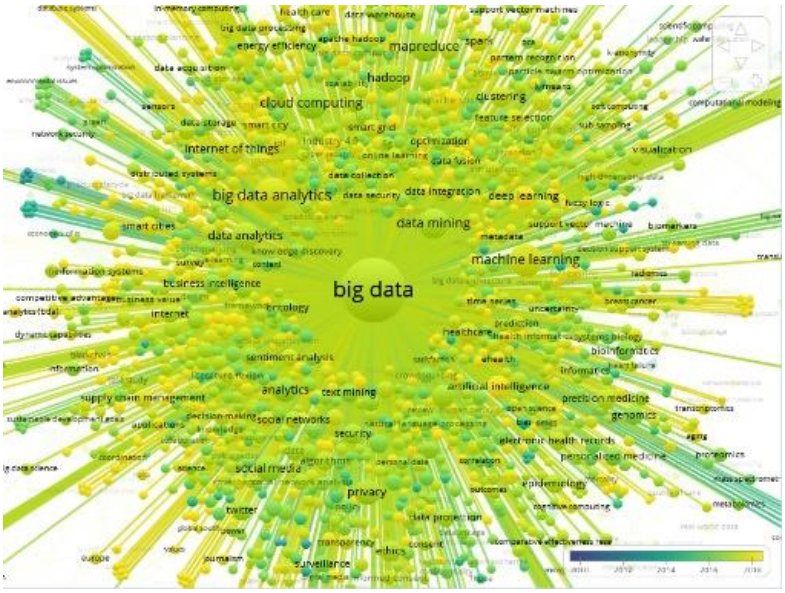

Fig 3. Co-occurrence relation network of the keyword “big data”

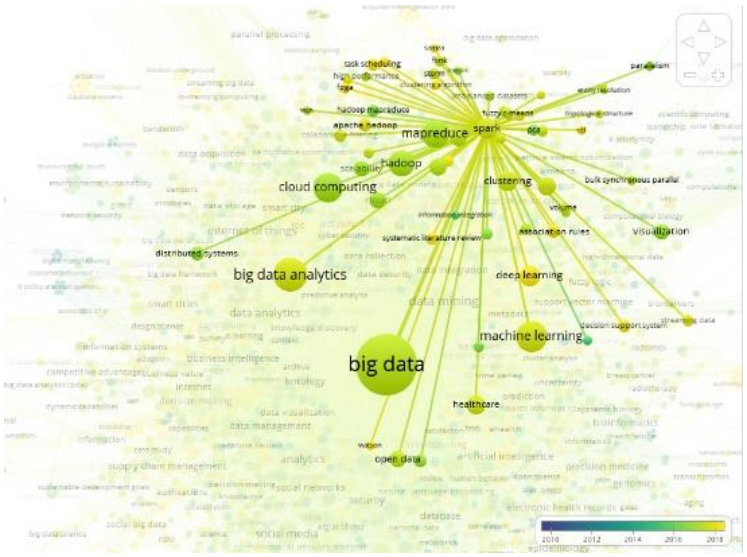

Fig 4. Co-occurrence relation network of the keyword “spark”

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201910_5(10).0025

Analyzing keywords in the time order of average publication can help researchers understand the general evolution of research hotspots in the field of big data. The order of keyword in big data field is sorted according to the number of occurrences, as shown in table 1.

Average publication time in table 1 is the average publication time of papers including the keyword instead of the initial appearance time of the keyword. Due to space constraints, only part of the keywords is listed in each year.

Table 1. Average publication time of keywords

time	keywords
2013	social network sites, communication studies
2014	Bias, quantitative analysis, multivariate, national governance, social data
2015	systems biology, storage, uncertainty, policy, digital humanities, anonymization, editorials, semantics, data scientist, mobile internet, data-intensive computing, virtualization, data compression
2016	big data, cloud computing, data mining, mapreduce, hadoop, data analytics, privacy, social media, analytics, data science, ethics, twitter, security, predictive analytics, business intelligence, genomics, algorithms, bioinformatics, epidemiology, text mining, data, data protection, personalized medicine, ontology, data quality, surveillance, database, cloud, performance, statistics, innovation, business analytics, social networks, data management, optimization, computational social science, visualization, parallel computing, knowledge management, epistemology, scalability, technology, data storage, sampling, methodology, information technology, informatics, natural language processing, random forest, data fusion, parallel processing, pattern recognition, medical informatics, crowdsourcing, knowledge discovery, data warehouse, data privacy, anomaly detection, education, health informatics, information, unstructured data, challenges, big data computing
2017	big data analytics, machine learning, internet of things, artificial intelligence, clustering, spark, iot, healthcare, precision medicine, apache spark, classification, smart cities, smart city, feature selection, electronic health records, data integration, smart grid, supply chain management, distributed computing, energy efficiency, sustainability, internet, open data, decision making, hdfs, nosql, sentiment analysis, industry 4.0, prediction, remote sensing, learning analytics, big data processing, multimedia big data, decision-making, china, social network analysis, support vector machine, public health, network analysis, simulation, supply chain, social big data, forecasting, data sharing, neural networks, big data applications, literature review, geospatial big data, fuzzy logic, survey, governance, wireless sensor networks, data security, management, decision tree, gis,, particle swarm optimization, data collection, access control, big data mining, time series, transparency, feature extraction, framework, online learning, value creation, resource allocation
2018	deep learning, internet of things (iot), fog computing, regression, smart manufacturing, blockchain, mortality, systematic literature review, task scheduling, nasopharyngeal carcinoma, data preprocessing, higher education, radiomics, gene expression, sdn, smart data, edge computing
2019	climate policy, database systems, digital platforms, ensemble methods, graphics processing units (gpus), integration systems, interpolation, long-short term memory, security of data, spatio-temporal analysis, stock market, tensorflow

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201910_5(10).0025

Fig. 5 is the co-occurrence network with time axis which adjusted by the order of average publication time of keywords. A total of 23 clusters are arranged in parallel, and each cluster is formed by a group of documents based on the same research paradigm, and links between documents represent the co-occurrence relations. The closer the node is to the right of the image, the newer the average publication time is, the newer the research direction is.

Analyzing the co-occurrence network with time axis can help researcher understand the main research topics of each cluster in big data field and the general evolution of research hotspots of each research topic.

Cluster 1 is the class with the largest number of keywords, and then the number of keywords in each cluster decreases in turn. 34.7% of clusters (the first eight) account for 78.2% of the total number of keywords, 52.2% of clusters (the first twelve) account for 91.2% of the total number of keywords. It shows that more than 90% of the research in the field of big data focuses on 12 research topics.

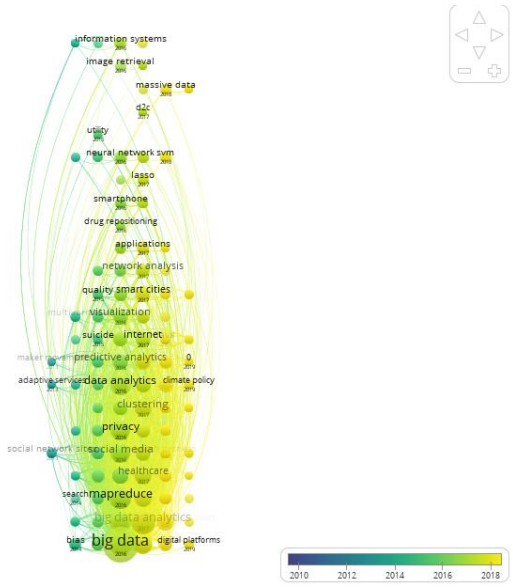

Fig 5. Co-occurrence network with time axis

3.2. Knowledge Mapping of The Key Literature Based on Citation Relations

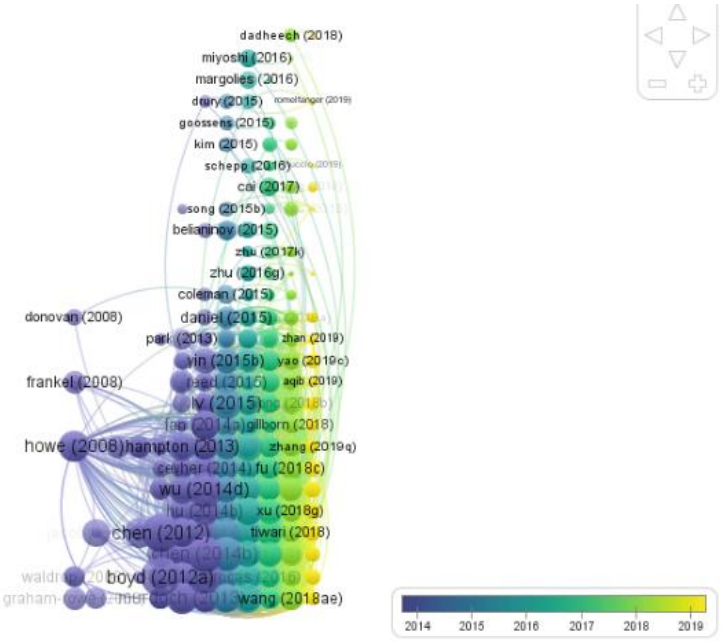

Fig 6. Citation network with time axis

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201910_5(10).0025

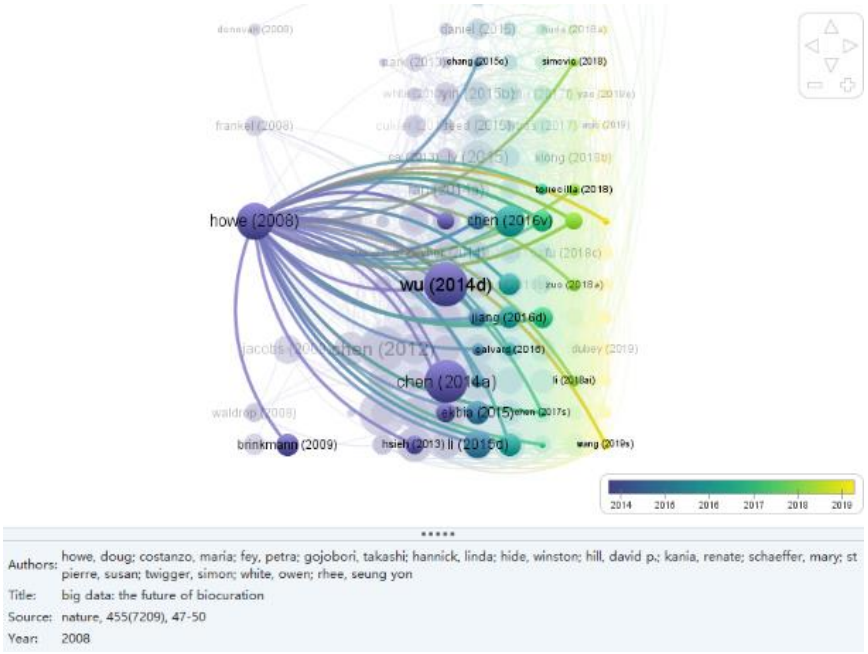

Fig 7. Citation network of document Howe (2008)

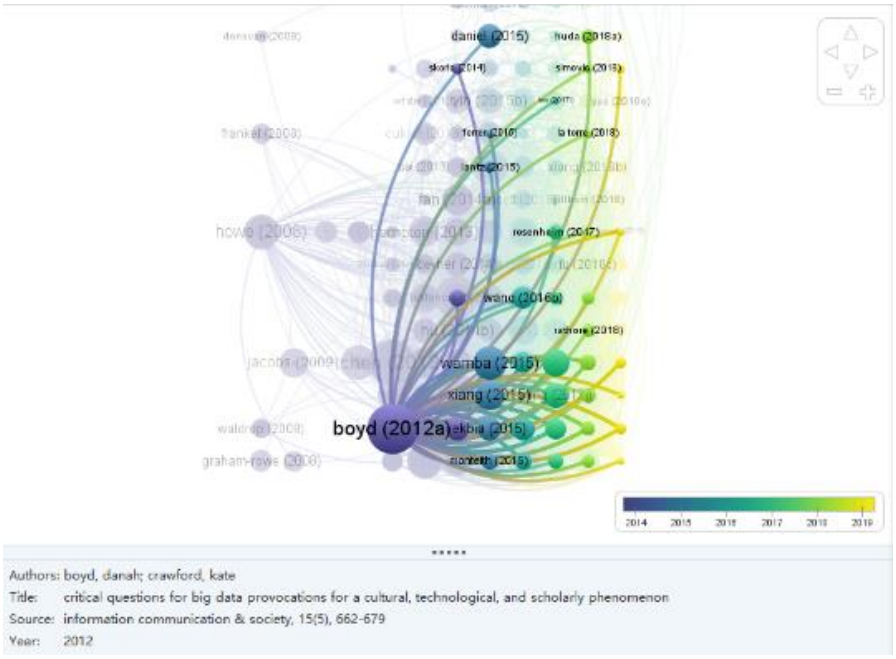

Fig 8. Citation network of document boyd(2012)

Fig. 6 is overlay visualization with time axis which is constructed based on citation relations, which adjusted by the order of average publication time of papers. A total of 27 clusters were clustered. Nodes in the graph represent documents in the field of big data. The size of nodes takes citation as weight. The larger the citation value, the larger the node is. Two nodes connected by link represent two references with citation relationship, and the left node is the reference of the right node.

Every year there are some highly cited documents. As shown in fig. 7, document Howe(2008) is the largest citation in earlier articles. And document boyd(2012) in fig. 8 is the most cited document in the field of big data by far.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201910_5(10).0025

Documents that are cited more frequently each year may be the important literatures in big data field. Researchers can get a general understanding of the overall research development context by analyzing the research content of these documents and combining with the hotspots and trends of keywords obtained from the analysis in section 3.1.

The research on big data started with the origin of big data, gradually involving more and more content, including the impact of big data, how to manage big data, the characteristics of big data, and the challenge [4-14], etc. The advances in Hadoop, cloud computing, deep learning and other technologies, and the emergence of the internet of things, have promoted and expanded the research content of big data analytics[15-24]. Nowadays, the models, methods and algorithms of big data analytics and computation have entered a new stage of development [25-28]. In industry 4.0, healthcare and other industries, the application of big data analytics and the challenges it faces remain the focus of attention [29, 30]. Privacy of big data has also been the research topic of continuous concern [31-34].

3.3. Knowledge Mapping Based on Co-Authorship On Authors

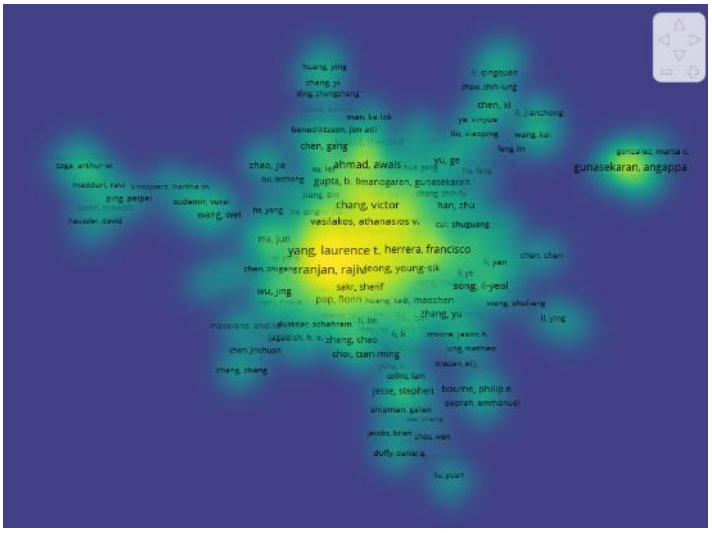

Fig 9. Density visualization with weight of documents based on co-authorship relations and authors as the unit of analysis

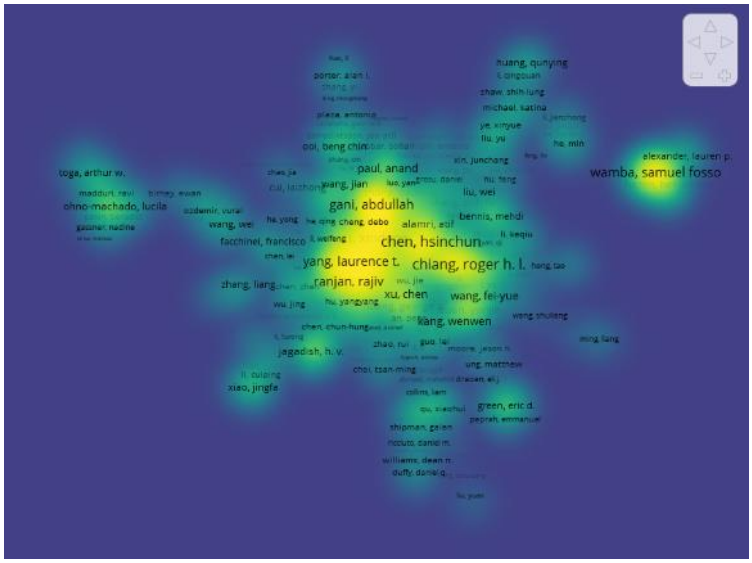

Fig 10. Density visualization with weight of citations based on co-authorship relations and authors as the unit of analysis

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201910_5(10).0025

Fig. 9 is density visualization with weight of documents which are constructed based on co-authorship relations and authors as the unit of analysis. Fig. 10 is density visualization with weight of citations which are constructed based on co-authorship relations and authors as the unit of analysis.

As shown in fig. 9 and fig. 10, the more the number of published articles, the closer the author is to the yellow hot spot of the density map. The fig. 9 and fig. 10 show which authors have published more papers in the field of big data.

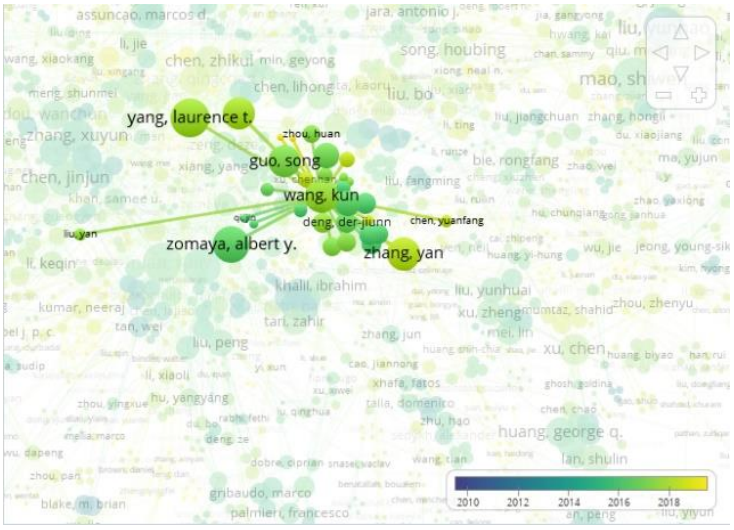

Fig 11. Co-authorship network of Wang, Kun

Each author cooperative relationships with different weights can be obtained by links between nodes in overlay visualization. Take author Wang, Kun as an example, fig. 11 is the co-authorship network of Wang, Kun. Authors who have links to Wank, Kun have partnerships with him.

Table 2 list the names of authors in the top 10 sorted by citations and in the top 10 sorted by documents.

Table 2. Names of authors in the top 10 sorted by citations and by documents

Names of authors in the top 10 sorted by citations	Names of authors in the top 10 sorted by documents
chiang, roger h. l.	yang, laurence t.
storey, veda c.	ranjan, rajiv
chen, hsinchun	herrera, francisco
chen, min	wang, lizhe
gani, abdullah	chang, victor
hashem, ibrahim abaker targio	gunasekaran, angappa
mao, shiwen	zomaya, albert y.
wu, xindong	wamba, samuel fosso
wu, gong-qing	paul, anand
ding, wei	ahmad, awais

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201910_5(10).0025

3.4. Knowledge Mapping Based on Co-Authorship On Organizations

Fig. 12 is overlay visualization with weight of documents which are constructed based on co-authorship relations and organizations as the unit of analysis. Fig. 12 shows which organizations have more papers in the field of big data, and organization cooperative relationships with weight of documents can be obtained by links between nodes in overlay visualization.

Fig. 13 is density visualization with weight of citations which are constructed based on co-authorship relations and organizations as the unit of analysis. Fig. 13 shows which organizations have more citations in the field of big data, and organization cooperative relationships with weight of citations can be obtained by links between nodes in overlay visualization.

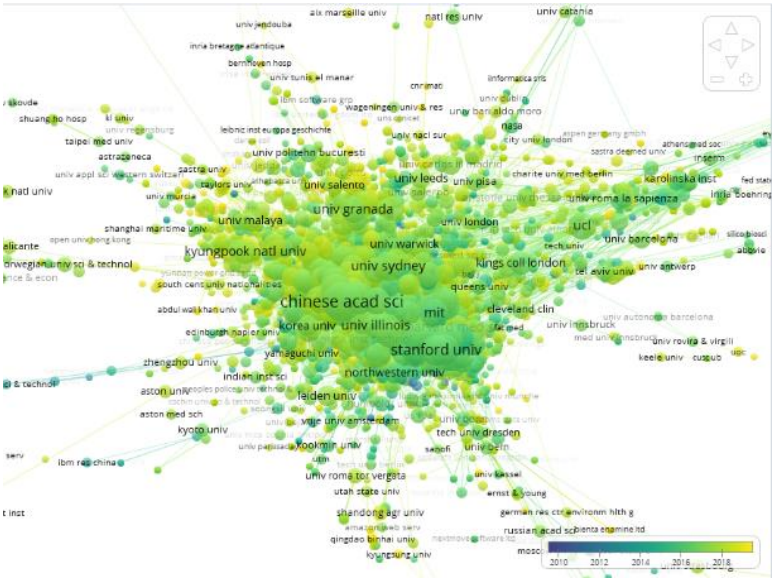

Fig 12. Density visualization with weight of documents based on co-authorship relations and organizations as the unit of analysis

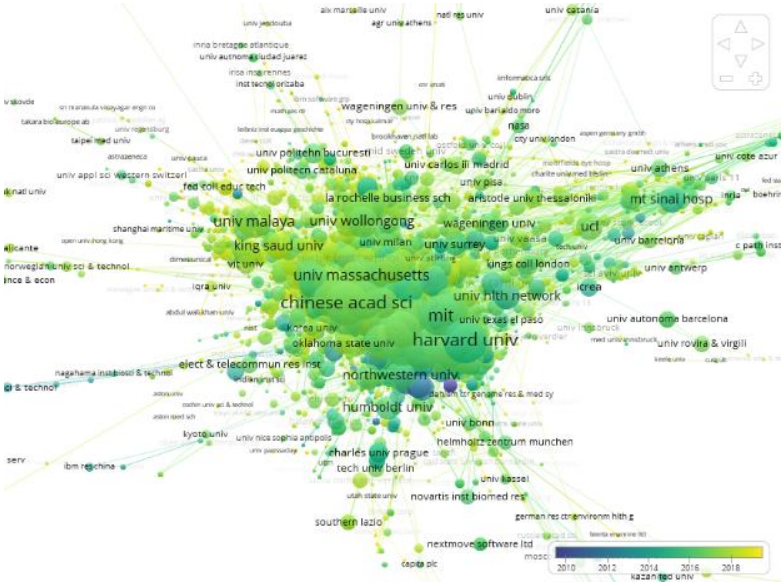

Fig 13. Density visualization with weight of citations based on co-authorship relations and organizations as the unit of analysis

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201910_5(10).0025

Table 3 list the names of organizations in the top 10 sorted by citations and in the top 10 sorted by documents.

Table 3. Names of organizations in the top 10 sorted by citations and by documents

Names of organizations in the top 10 sorted by citations	Names of organizations in the top 10 sorted by documents
harvard univ	chinese acad sci
chinese acad sci	tsinghua univ
mit	stanford univ
huazhong univ sci & technol	huazhong univ sci & technol
tsinghua univ	nyu
stanford univ	univ michigan
microsoft res	univ sydney
univ arizona	harvard univ
georgia state univ	univ minnesota
univ cincinnati	univ washington

3.5. Knowledge Mapping Based on Co-Authorship On Countries

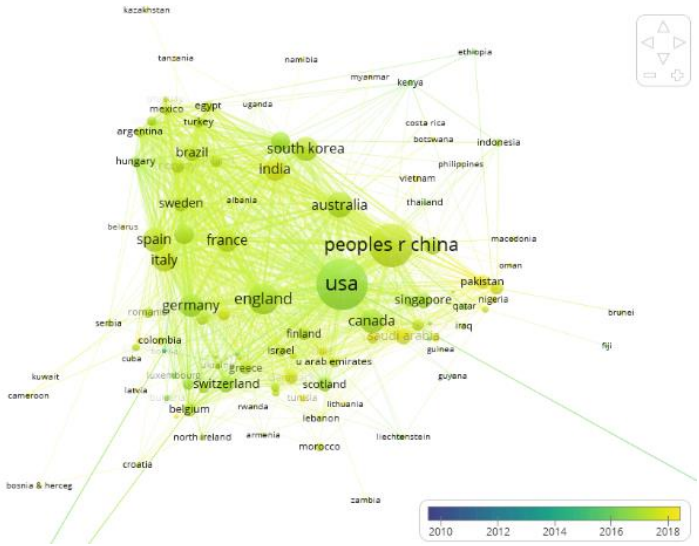

Fig 14. Density visualization with weight of documents based on co-authorship relations and countries as the unit of analysis

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201910_5(10).0025

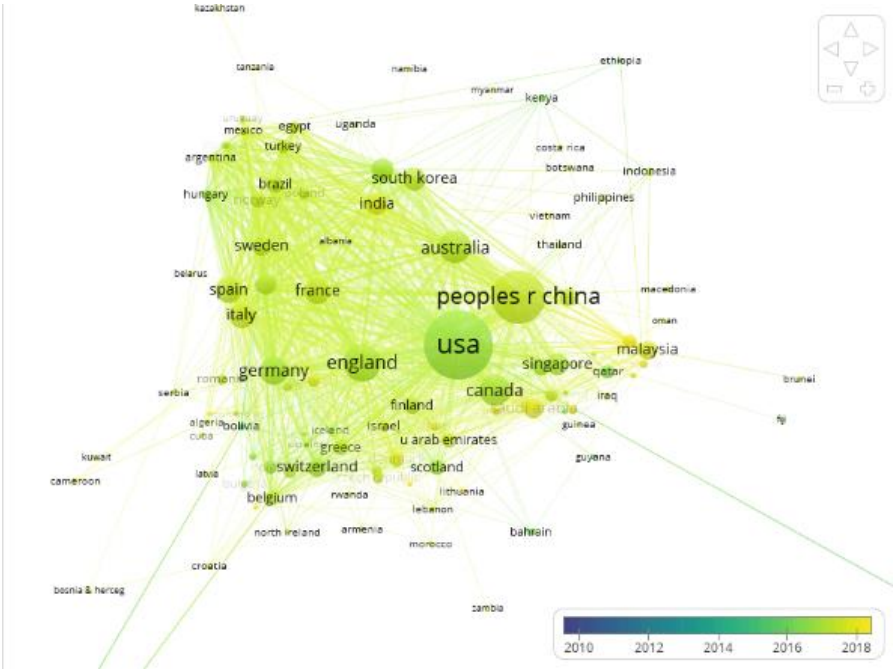

Fig 15. Density visualization with weight of citations based on co-authorship relations and countries as the unit of analysis

Fig. 14 is overlay visualization with weight of documents which are constructed based on co-authorship relations and countries as the unit of analysis. Fig. 14 shows which countries have more papers in the field of big data, and country cooperative relationships with weight of documents can be obtained by links between nodes in overlay visualization.

Fig. 15 is density visualization with weight of citations which are constructed based on co-authorship relations and countries as the unit of analysis. Fig. 15 shows which countries have more citations in the field of big data, and country cooperative relationships with weight of citations can be obtained by links between nodes in overlay visualization.

Table 4 list the names of countries in the top 10 sorted by citations and in the top 10 sorted by documents.

Table 4. Names of counties in the top 10 sorted by citations and by documents

Names of counties in the top 10 sorted by documents	Names of counties in the top 10 sorted by citations
usa	usa
peoples r china	peoples r china
england	england
australia	australia
spain	canada
canada	germany
south korea	france
germany	spain
italy	italy
india	south korea

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201910_5(10).0025

3.6. Knowledge Mapping Based on Bibliographic Coupling

3.6.1 Bibliographic coupling of documents

If both articles cite the same paper in the references, the two articles are called coupled papers. The more papers with similar subject or professional contents and the more references they contain, the stronger the coupling strength of the coupled papers will be.

A total of 22 clusters were clustered by bibliographic coupling of documents. Papers in a cluster have similar research contents. By analyzing the literature in the same cluster, researchers can analyze the research topics. Combined with research hotspot in section 3.1 and key literature analysis in section 3.2, researchers can get a clearer understanding of the research status.

The links between nodes which take citations as weights indicates that there is a bibliographic coupling relationship between nodes. Fig. 16 is overlay visualization of salas-vega(2015) which are constructed based on bibliographic coupling relations.

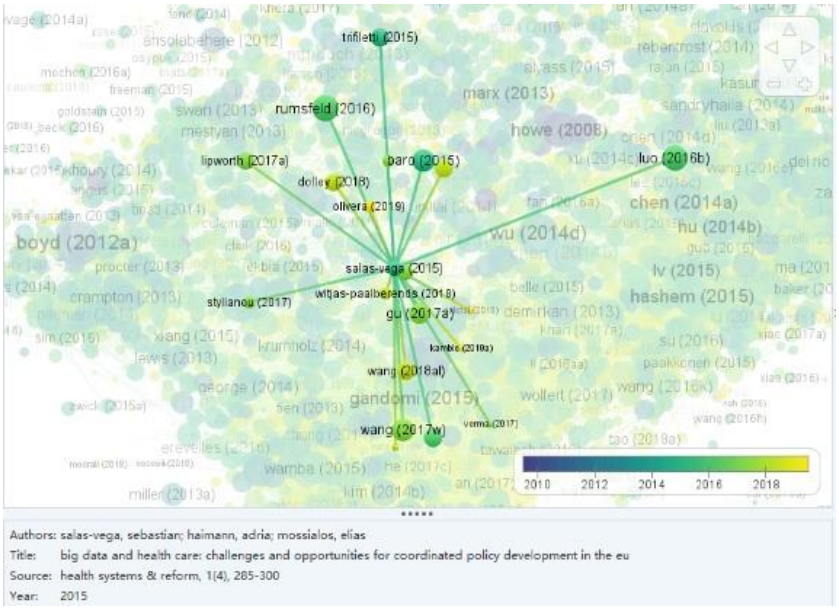

Fig 16. Overlay visualization of salas-vega(2015) based on bibliographic coupling relations

3.6.2 Bibliographic coupling of sources

A total of 9 clusters were clustered by bibliographic coupling of sources. From the analysis results, most journals are clustered in cluster 1 to 4, cluster 5 has only three journals, cluster 6 to 9 has only two journal each.

Fig. 17 shows that from cluster 1 to 4, different clusters are displayed by different colors, with clear boundaries and compact interior. Papers in a cluster have similar research contents. Therefore, the style of journals can be understood through the bibliographic coupling.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201910_5(10).0025

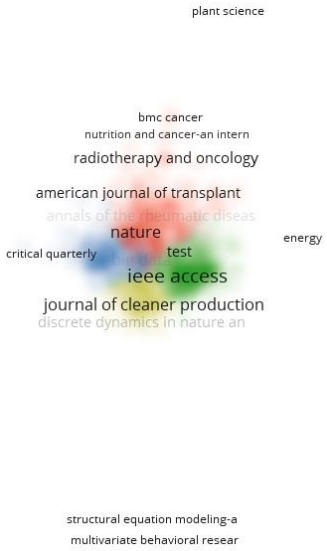

Fig 17. Density visualization based on bibliographic coupling of sources

4. CONCLUSION

In this paper, knowledge mapping of big data based on web of science core collection database is used for analysis. Through analysis, researchers can see that the research hotspots and trends, the general evolution of research hotspots, and the main research topics in the field of big data. Researchers can also learn about the publication and cooperation of researchers, organizations and countries. The research in the field of big data covers many aspects of big data. Analyzing knowledge mapping in big data can get a general understanding of the development of big data.

With the progress of the models, methods and algorithms of big data analytics and computation, such as Hadoop, cloud computing, deep learning and other related technologies, and the emergence of the internet of things which have promoted and expanded the research content of big data analytics and its application in all walks of life, the management, analytics and calculation of big data have been stepping into a new stage of development. The models, methods and algorithms of big data analytics, the application of big data analytics in industry 4.0, healthcare and other industries and the challenges it faces, and privacy of big data, are the research topic of continuous concern.

REFERENCES

[1] Xu, L.D. and L. Duan, Big data for cyber physical systems in industry 4.0: a survey. Enterprise Information Systems, 2019. 13(2): p. 148-169.

[2] WebofScience, http://www.webofknowledge.com/.

[3] VOSviewer, https://www.vosviewer.com/.

[4] Howe, D., et al., Big data: The future of biocuration. Nature, 2008. 455(7209): p. 47-50.

[5] Lynch, C., Big data: How do your data grow? Nature, 2008. 455(7209): p. 28-29.

[6] Jacobs, A., The Pathologies of Big Data. Communications of the Acm, 2009. 52(8): p. 36-44.

[7] Lavalle, S., et al., Big Data, Analytics and the Path From Insights to Value. Mit Sloan Management Review, 2011. 52(2): p. 21-32.

[8] Chen, H.C., R.H.L. Chiang, and V.C. Storey, Business Intelligence and Analytics: From Big Data to Big Impact. Mis Quarterly, 2012. 36(4): p. 1165-1188.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201910_5(10).0025

[9] McAfee, A. and E. Brynjolfsson, STRATEGY & COMPETITION Big Data: The Management Revolution. Harvard Business Review, 2012. 90(10): p. 60-+.

[10] Madden, S., From Databases to Big Data. Ieee Internet Computing, 2012. 16(3): p. 4-6.

[11] Davenport, T.H., P. Barth, and R. Bean, How 'Big Data' Is Different. Mit Sloan Management Review, 2012. 54(1): p. 43-+.

[12] Ansolabehere, S. and E. Hersh, Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate. Political Analysis, 2012. 20(4): p. 437-459.

[13] Murdoch, T.B. and A.S. Detsky, The Inevitable Application of Big Data to Health Care. Jama-Journal of the American Medical Association, 2013. 309(13): p. 1351-1352.

[14] Marx, V., The Big Challenges of Big Data. Nature, 2013. 498(7453): p. 255-260.

[15] Demirkan, H. and D. Delen, Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. Decision Support Systems, 2013. 55(1): p. 412-421.

[16] O'Driscoll, A., J. Daugelaite, and R.D. Sleator, 'Big data', Hadoop and cloud computing in genomics. Journal of Biomedical Informatics, 2013. 46(5): p. 774-781.

[17] Wu, X.D., et al., Data Mining with Big Data. Ieee Transactions on Knowledge and Data Engineering, 2014. 26(1): p. 97-107.

[18] Fan, J.Q., F. Han, and H. Liu, Challenges of Big Data analysis. National Science Review, 2014. 1(2): p. 293-314.

[19] George, G., M.R. Haas, and A. Pentland, Big Data and Management. Academy of Management Journal, 2014. 57(2): p. 321-326.

[20] Chen, X.W. and X.T. Lin, Big Data Deep Learning: Challenges and Perspectives. Ieee Access, 2014. 2: p. 514-525.

[21] Lv, Y.S., et al., Traffic Flow Prediction With Big Data: A Deep Learning Approach. Ieee Transactions on Intelligent Transportation Systems, 2015. 16(2): p. 865-873.

[22] Triguero, I., et al., MRPR: A Map Reduce solution for prototype reduction in big data classification. Neurocomputing, 2015. 150: p. 331-345.

[23] Obermeyer, Z. and E.J. Emanuel, Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. New England Journal of Medicine, 2016. 375(13): p. 1216-1219.

[24] Zaharia, M., et al., Apache Spark: A Unified Engine for Big Data Processing. Communications of the Acm, 2016. 59(11): p. 56-65.

[25] Zhang, Q.C., et al., An Adaptive Dropout Deep Computation Model for Industrial IoT Big Data Learning With Crowdsourcing to Cloud Computing. Ieee Transactions on Industrial Informatics, 2019. 15(4): p. 2330-2337.

[26] Romanowski, A., Big Data-Driven Contextual Processing Methods for Electrical Capacitance Tomography. Ieee Transactions on Industrial Informatics, 2019. 15(3): p. 1609-1618.

[27] Wang, X.K., et al., A Tensor-Based Big-Data-Driven Routing Recommendation Approach for Heterogeneous Networks. Ieee Network, 2019. 33(1): p. 64-69.

[28] Singh, S.P., et al., Fog computing: from architecture to edge computing and big data processing. Journal of Supercomputing, 2019. 75(4): p. 2070-2105.

[29] Sun, Y.C., et al., Internet of Things and Big Data Analytics for Smart and Connected Communities. Ieee Access, 2016. 4: p. 766-773.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201910_5(10).0025

[30] Wang, Y.C., L. Kung, and T.A. Byrd, Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change, 2018. 126: p. 3-13.

[31] Xu, L., et al., Information Security in Big Data: Privacy and Data Mining. Ieee Access, 2014. 2: p. 1149-1176.

[32] Zhang, Q.C., et al., Privacy-Preserving Double-Projection Deep Computation Model With Crowdsourcing on Cloud for Big Data Feature Learning. Ieee Internet of Things Journal, 2018. 5(4): p. 2896-2903.

[33] Yin, C.Y., et al., Location Privacy Protection Based on Differential Privacy Strategy for Big Data in Industrial Internet of Things. Ieee Transactions on Industrial Informatics, 2018. 14(8): p. 3628-3636.

[34] Li, S., et al., Searchable Encryption Scheme for Personalized Privacy in IoT-Based Big Data (vol 19, 1059, 2019). Sensors, 2019. 19(10).