Advance Journal of Food Science and Technology 11(10): 697-702, 2016
DOI:10.19026/ajfst.11.2766
ISSN: 2042-4868; e-ISSN: 2042-4876

Submitted: September 9, 2015 Accepted: October 17, 2015 Published: August 05, 2016

Research Article

Research on Visual Analysis of Agricultural Food Big Data Based on CiteSpace III

[1]Xuehong Zhang, [1]Yunjiang Xi, [1]Xiao Liao and [2]Shengze Li
[1]School of Business Administration, South China University of Technology, Guangzhou, 510641,
[2]Machinery Engineering Corporation of China, Beijing, 100055, China

Abstract: This study makes visual analysis on agricultural food big data retrieval literature by using the information visualization tool CiteSpace III and with the Web of Science™ core collection as data sources. The spatial and temporal distribution, research focus, major fields of study, research fronts and evolution paths on the research field of big data were analyzed by knowledge maps and literature research. The results of the research show that the research focus in future may include Hadoop Distributed File System, Hadoop Database, performance evaluation and medical research.

Keywords: Agricultural food big data, CiteSpace III, evolution paths, research focus, visualization

INTRODUCTION

"Big Data" special issue was published on Nature, in September 2008. Then the research and applications of big data became the focus of attention. Over the past years, research on big data showed a trend of explosive growth and has made great progress in many fields (Feng and Guo, 2013; Mcafee *et al*., 2012; Hashem *et al*., 2015). It is necessary for us to find out what are the hot research topics, the major fields of study and the research trends in future. With the above purposes, this study made visual analysis by knowledge maps based on CiteSpace III.

In this study, we took Web of Science ™ core collection as data source to insure the quality of literatures. Data were collected on May 15, 2015, by selecting the retrieval theme for "big data" and the time span for 2008-2015, including databases of SCI-EXPANDED, SSCI, CPCI-S and CPCI-SSH. The type of literature was refined to article or proceedings paper or reviewed with data download as "all records". Then a total of 2970 records were acquired for further analysis. These records come from 1744 institutions in 79 countries or regions, involving more than 100 research directions and nearly 800 kinds of journals and conference sets.

To make the visual analysis on the literature, we use CiteSpace III as knowledge mapping tools and the analyzing process are as follows: firstly the data were pre-processed, such as standardization of keywords, e.g., the keyword "Map-Reduce" was transformed into "MapReduce" and some homogeneous words were merged and so on. Then the data were input into CiteSpace tools for further analysis. The related settings are: the selected time period is “from 2008 to 2015”, time interval is 1 year, the high-frequency keywords are selected as: top 50, high-cited literature are selected as top 40. As for co-word network, keywords are set as nodes and for cited network, citing or cited literature are set as nodes. The visual analysis includes the spatial and temporal distribution based on bibliometrics, research focus, major fields of study and research front based on co-word network and the evolution paths in terms of cited reference co-appearance network (Chen, 2006; Wang, 2015).

This study makes visual analysis on agricultural food big data retrieval literature by using the information visualization tool CiteSpace III and with the Web of Science™ core collection as data sources. Through the analysis, we clearly know the development stage, research focus, major fields, research fronts and evaluation paths about the research of big data. On the regional distribution, USA, China and UK have made many achievements. Chinese Academy of Sciences is the most important institution on the research of big data. The research of big data has transformed into applications from theories. The technology of big data (MapReduce, Hadoop, cloud computing, etc.), applications (designing system and network data analysis), problems and challenges (the quality of big data) is the current research focus. On the future trend, HDFS, HBase, performance evaluation and medical research may represent the research front and develop into the hot spots in future. The spatial and temporal distribution, research focus, major fields of study, research fronts and evolution paths on the research field of big data were analyzed by knowledge maps and literature research. The results of

Corresponding Author: Xiao Liao, School of Business Administration, South China University of Technology, Guangzhou, 510641, China

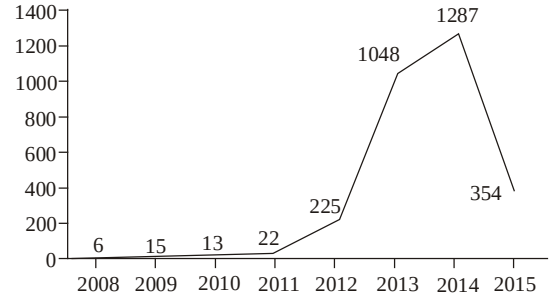

Fig. 1: Quantity of publications from 2008 to 2015

Table 1: Top 10 countries with most records

Country	Records	Country	Records
USA	1108	South Korea	111
China	589	Japan	105
England	167	Canada	104
Germany	162	Italy	83
Australia	131	France	70

Table 2: Top 10 Institutions with most records

Institution	Records	Institution	Records
CHINESE ACAD SCI	64	UNIV SO CALIF	31
UNIV CALIF LOS ANGELES	36	HARVARD UNIV	31
TSINGHUA UNIV	34	UNIV CALIF SAN DIEGO	26
MIT	34	STANFORD UNIV	23
UNIV TECHNOL SYDNEY	31	UNIV ILLINOIS	21

the research show that the research focus in future may include Hadoop Distributed File System, Hadoop Database, performance evaluation and medical research.

LITERATURE DISTRIBUTION STATISTICS

Annual distribution: The annual distribution statistics are as shown in Fig. 1.

Figure 1, we can divide the research of "big data" into three stages:

- From 2008 to 2011, it belongs to the initial stage with only 56 records in total.
- From 2012 to 2013, it's in high-growth stage with doubled and redoubled achievements and the total records in 2013 are over one thousand.
- From 2014 to present, it walks into the stage of steady growth with 70.8 records per month in 2015. We can expect that the records in 2015 could reach 1300-1500. As it can be seen; the growth rate of literatures on big data has slowed down and goes into the stable period after explosive growth.

Regional disbution: From Table 1, we can see that USA has 1108 records among ten countries, while China (only including Chinese mainland and Hong Kong) has 589 records. These two countries generated more than 50% papers in the big data research. Then, it's UK, Germany, Australia, Korea, Japan, Canada, Italy and France. Totally, there are great gaps between USA and other countries.

Institution distribution: From Table 2, we can figure that there are seven institutions from USA, two from China, one from Australia. It is noteworthy that Chinese Academy of Sciences ranks first with 64 records.

It should be the most important institution of big data research in Chinese mainland. UCLA, Tsinghua University, MIT, UTS, USC, Harvard University followed closely. But there's no significant difference on quantity between them.

RESEARCH FOCUS, FIELDS AND FRONTS ANALYSIS

Research focus distribution: Table 3, there are the top 18 keywords whose frequency is greater than or equal to 50. According to the theory of knowledge map, centrality and high-frequency keywords represent the

Table 3: Top 18 high-frequency keywords (remove the search term)

Keyword	Frequency	Centrality	Keyword	Frequency	Centrality
MapReduce	264	0.2	Performance	72	0.17
Cloud computing	210	0.26	Visualization	67	0.04
Systems	165	0.36	Information	65	0.12
Hadoop	146	0.05	Analytics	64	0.03
Networks	126	0.17	Privacy	63	0
Algorithms	113	0.02	Management	60	0.15
Data mining	110	0.01	Data analytics	57	0.04
Model	110	0.18	Cloud	51	0.03
Classification	78	0.09	Social media	50	0.04

Table 4: Top 8 burst terms

Keyword	Strength	Begin	End	2011-2015
Component	7.0879	2012	2012	
Mapreduce	2.2458	2012	2013	
Hdfs	1.9958	2012	2012	
Evaluation	2.3649	2012	2012	
Hbase	1.8207	2012	2013	
Ehealth	3.8842	2013	2013	
Data analytics	2.9335	2013	2013	
Cancer	2.5031	2014	2015	

research focus at a time. Figure 2, it shows the research focus based on keyword co-appearance network. The bigger the node size, the higher the frequency of keyword; the connection between the nodes shows the co-appearance relationship; the nodes with purple circle mean high centrality.

Combining Table 4 with Fig. 2, we can learn the research focus about "big data": Cloud computing, MapReduce, systems, Hadoop, algorithms, data mining, model, performance, management.

Major fields of big data research: Under the analysis of keyword network subgroups, we divided the current research into the following fields:

Research on the technology of big data:
Related keywords and main associations: Big Data--cloud computing, Big Data--MapReduce--Hadoop, Big Data--machine learning, Big Data--recognition--data mining--algorithms; Big Data--visualization. The fieldsmainly study various technology of big data including artificial intelligence, cloud computing, machine learning and data mining algorithms. The words, MapReduce and Hadoop, respectively ranked first and fourth in all keywords, it indicts that the research of technology is the major fields. In addition, it also includes heuristic analysis technology based on human-computer interaction, which intends to involve the person's cognitive capabilities that the machine is not good at into the analysis process, like visual data mining techniques and visual interactive analysis (Li and Gong, 2015; Staff, 2014; Shivhare *et al*., 2013; Li *et al*., 2014).

Design and application of systems based on big data: Related keywords and main associations: Big Data--systems--design--performance, Big Data--systems--model--management. "Systems" which ranked third in all keywords and the strong co-appearance relationships with other high-frequency keywords (model, design, performance, management) reflect the study of systems and model based on big data is becoming the hot spots in recent years, such as supply chain systems, performance management systems, self-quantification systems for personal health information (Almalki *et al*., 2015). Leveling *et al*. (2014) illustrated the important role of big data in the supply chain management. It may not only increase the visibility of supply chain, but also lead a new business model like the Amazon patent (Leveling *et al*., 2014).

Big data analysis based on network data: Related keywords and main associations: Big Data--Data analysis; Big Data--networks; Big Data--twitter--social media; Big Data-twitter--component. It makes sense to

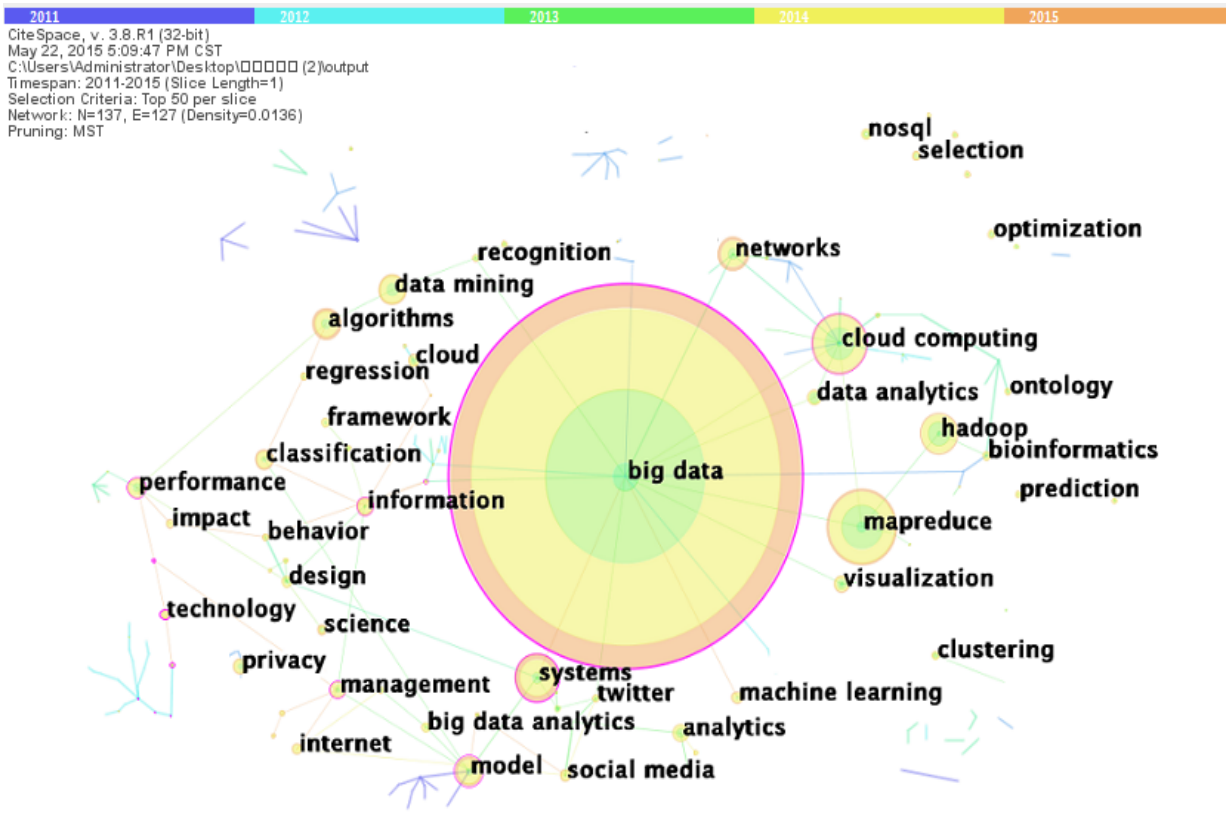

Fig. 2: Knowledge map of research focus in big data field

mine and to analyze various types of network data for discovering new rules (Tang and Chen, 2015). For example, the data of video-sharing site, shopping site, social media, like Twitter, Facebook and so on. Colleoni *et al*. (2014) investigated political homophily on Twitter. Using combination of machine learning and social network analysis they classified users as Democrats or as Republicans based on the political content shared (Colleoni *et al*., 2014). Yu and Wang (2015) collected real-time tweets from U.S. soccer fans during five 2014 FIFA World Cup. They used sentiment analysis to examine U.S. soccer fans' emotional responses in their tweets, particularly, the emotional changes after goals (Yu and Wang, 2015).

Research on the quality of big data: Related keywords and main associations: Big Data--quality--information--classification. The field mainly relates to the quality of big data and classification of information and data. To ensure the quality of big data is the premise for effective analysis. Small, easily overlooked data quality problems will be enlarged in the age of big data and even lead to unrecoverable disaster. It is estimated that the American corporations lose nearly $ 600 billion every year due to incorrect data. The company's rate of data error is about 1 to 5%, some companies may be up to 30% (Kwon *et al*., 2014; Hazen *et al*., 2014; Saha and Srivastava, 2014). The study in this field mainly focus on how to improve the quality of big data to reduce data error and ensure better analysis results.

Research fronts analysis based on burst terms: In Citespace III, burst terms are suitable for detecting the developing trends and the fronts. Therefore, we use word frequency detection technology to analyze the retrieved data to detect the words with high frequency rate (burst term) from a large number of keywords. Here list top 8 burst terms in Table 5 and you can see the time span of each word.

It is obvious that the number of burst terms in 2012 is more than any other year. It may have a greater relationship with the rapid growth of literatures "Mapreduce, HDFS, HBase, etc." has attracted scholars' attention; data analysis, component analysis, performance evaluation also came into view. It is noteworthy that the time span of "cancer" is from 2014 to 2015. The application of big data in cancer field may be the new front. There have been literatures about medical cases, cancer research and social health-care under the environment of big data. For example, Shneiderman *et al*. (2013) proposed that interactive information visualization and visual analytics methods will bring profound changes to personal health

Table 5: Top 6 cited references

Citations	Reference information	Reference resource
353	Mapreduce: Simplified data processing on large clusters; Dean and Ghemawat (2008)	Communication Of The Acm
130	Big data: The next frontier for innovation, competition and productivity; Manyika *et al*. (2011)	McKinsey Global Institute
118	Hadoop: The Definitive Guide; White (2009)	O'Reilly Media, Inc.
59	Big data: A Revolution That Will Transform How We Live, Work and Think; Mayer-Schönberger and Cukier (2013)	John Murray Publishers Ltd
51	Big data: The future of biocuration; Howe *et al*. (2008)	Nature
49	Big data: How do your data grow? ; Lynch (2008)	Nature

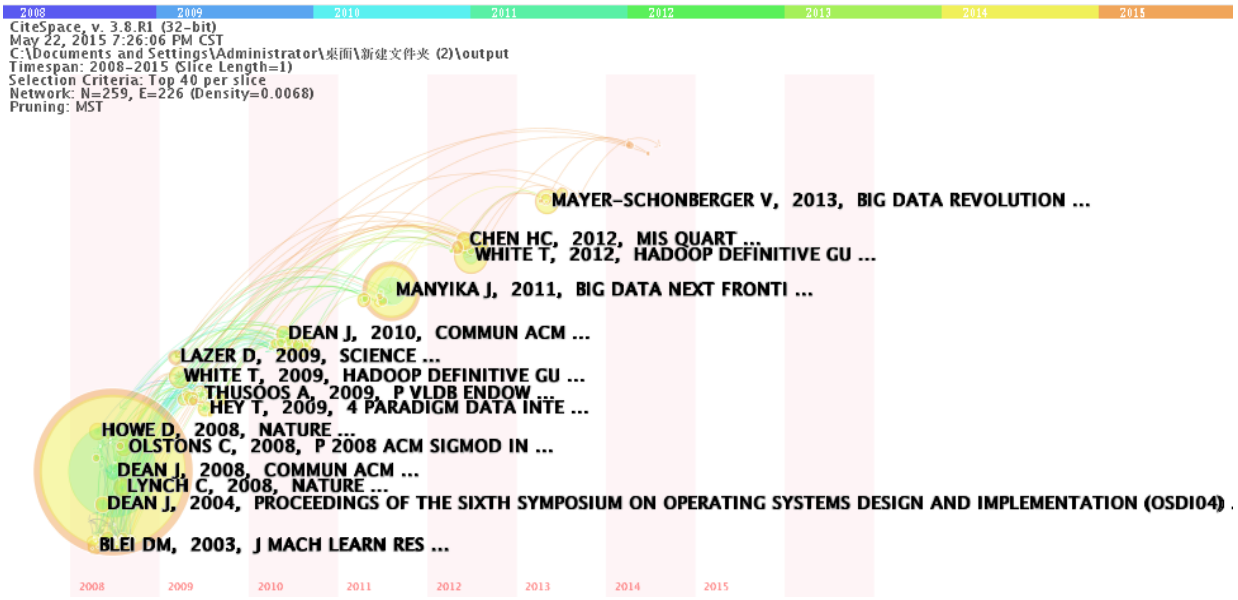

Fig. 3: Time-zone view of cited reference co-appearance network (2008-2015)

programs, clinical healthcare delivery and public health policymaking. Fridley *et al*. (2014) thought that each woman and her cancer are unique, successful cures and outcomes will only come from informative biomarkers and treatments that target specific cells within each person's tumor. Therefore, it may be a good way to provide medical services through personalized big data (Shneiderman *et al*., 2013; Fridley *et al*., 2014; Raghupathi and Raghupathi, 2014; Anderson and Chang, 2015).

EVOLUTION PATHS ANALYSIS BASED ON TIME-ZONE VIEW OF CITED REFERENCE CO-APPEARANCE

Time-zone view is a kind of knowledge map which places emphasis on time dimension to show the evolution paths. Figure 3, each circular node represents a cited reference, bigger nodes with higher total citations and greater value. In Table 5, it lists the top six cited references which are the basement of big data research.

From Table 5, it can be seen the most frequently cited reference is *Mapreduce*: Simplified data processing on large clusters published by Dean and Ghemawat (2008). The article learned from functional programming language and applied the MapReduce model to the parallel computing of big data sets. It shows that improving the ability of using big data by virtue of key technology became the focus of big data research (Dean and Ghemawat, 2008). The report from McKinsey Global Institute in 2011 ranked second, it systematically expounded the concept of big data, key technology and applications. At the same time, it revealed that data were becoming intangible assets., the age of big data, published by Mayer- Schönberger and Cukier (2013), presented three rules of dealing data, that is, all not sampling, efficiency not accuracy, correlation not causation. It challenged the traditional way of human cognition and thought. The Key Nodes is an important symbol of the applications in the age of big data (Mayer-Schönberger and Cukier, 2013).

In summary, we can sort out the evolution paths of big data research. In 2008, proposed the concept, technology applications and stressed using MapReduce on parallel operation of big data set, while began to extend to biology subject. In 2009, mainly explored Hadoop, MapReduce algorithm and building model. Data analysis became the foundation of Scientific discoveries. After 2011, described the concept and core technology systematically, analyzed the application deeply. For nearly two years, big data research has translated into social science and practical diffusion from computer science and data science. Scholars are concerned about public opinion analysis, sentiment analysis, behavior analysis and the quality of big data. In the meanwhile, the research of applications related to products and services innovation, marketing innovation under the environment of big data has come into the scholars' view (Howe *et al*., 2008; Lynch, 2008).

CONCLUSION AND DISCUSSION

In this study, we took Web of Science™ core collection as data source and made quantitative and visual analysis by CiteSpace III. Through the analysis, we clearly know the development stage, research focus, major fields, research fronts and evaluation paths about the research of big data. On the regional distribution, USA, China and UK have made many achievements. Chinese Academy of Sciences is the most important institution on the research of big data. The research of big data has transformed into applications from theories. The technology of big data (MapReduce, Hadoop, cloud computing, etc.), applications (designing system and network data analysis), problems and challenges (the quality of big data) is the current research focus. On the future trend, HDFS, HBase, performance evaluation and medical research may represent the research front and develop into the hot spots in future.

REFERENCES

Almalki, M., K. Gray and F.M. Sanchez, 2015. The use of self-quantification systems for personal health information: Big data management activities and prospects. Health Inform. Sci. Syst., 3: S1.

Anderson, J.E. and D.C. Chang, 2015. Using electronic health records for surgical quality improvement in the era of big data. JAMA Surg., 150(1): 24-29.

Chen, C.M., 2006. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. J. Am. Soc. Inf. Sci. Tec., 57(3): 359-377.

Colleoni, E., A. Rozza and A. Arvidsson, 2014. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. J. Commun., 64(2): 317-332.

Dean, J. and S. Ghemawat, 2008. MapReduce: Simplified data processing on large clusters. Commun. ACM, 51(1): 107-113.

Feng, Z.Y. and X.H. Guo, 2013. On the research frontiers of business management in the context of Big Data. J. Manage. Sci. China, 01: 1-9.

Fridley, B.L., D.C. Koestler and A.K. Godwin, 2014. Individualizing care for ovarian cancer patients using big data. J. Natl. Cancer I., 106(5): dju080.

Hashem, I.A.T., I. Yaqoob, N.B. Anuar, S. Mokhtar, A. Ghani *et al*., 2015. The rise of "big data" on cloud computing: Review and open research issues. Inform. Syst., 47: 98-115.

Hazen, B.T., C.A. Boone, J.D. Ezell and L.A. Jones-Farmer, 2014. Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. Int. J. Prod. Econ., 154: 72-80.

Howe, D., M. Costanzo, P. Fey, T. Gojobori, L. Hannick *et al*., 2008. Big data: The future of biocuration. Nature, 455(7209): 47-50.

Kwon, O., N. Lee and B. Shin, 2014. Data quality management, data usage experience and acquisition intention of big data analytics. Int. J. Inform. Manage., 34(3): 387-394.

Leveling, J., M. Edelbrock and B. Otto, 2014. Big data analytics for supply chain management. Proceeding of the IEEE International Conference on Industrial Engineering and Engineering Management (IEEM, 2014), pp: 918-922.

Li, F., B.C. Ooi, M.T. Özsu and S. Wu, 2014. Distributed data management using MapReduce. ACM Comput. Surv. (CSUR), 46(3), Article No. 31.

Li, X.L. and H.G. Gong, 2015. A survey on big data systems. Sci. China Inform. Sci., 45(1): 1-44.

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

Manyika, J., M. Chi, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A.H. Byers, 2011. Big data: The next frontier for innovation, competition and productivity. Report, McKinsey Global Institute.

Mayer-Schönberger, V. and K. Cukier, 2013. Big Data: A Revolution that will transform How We Live, Work, and Think [M]. Zhejiang People's Publishing House, Press, Hangzhou, pp: 4.

Mcafee, A., E. Brynjolfsson T.H. Davenport, D.J. Patil and D. Barton, 2012. Big data. The management revolution. Harvard Bus. Rev., 90(10): 61-67.

Raghupathi, W. and V. Raghupathi, 2014. Big data analytics in healthcare: Promise and potential. Health Inform. Sci. Syst., 2(1): 3.

Saha, B. and D. Srivastava, 2014. Data quality: The other face of big data. Proceeding of the IEEE 30th International Conference on Data Engineering (ICDE, 2014), pp: 1294-1297.

Shivhare, H., N. Mishra and S. Sharma, 2013. Cloud computing and big data. Proceeding of 2013 International Conference on Cloud, Big Data and Trust, pp: 222-225.

Shneiderman, B., C. Plaisant, B.W. Hesse, 2013. Improving health and healthcare with interactive visualization methods. HCIL Technical Report, 2013.

Staff, C., 2014. Visualizations make big data meaningful. Commun. ACM, 57(6): 19-21.

Tang, J. and W.G. Chen, 2015. Deep analytics and mining for big social data. Chinese Sci. Bull., 60(5/6): 509-519.

Wang, B.L., 2015. Research on big data based on scientometrics and visualization analysis. J. Intell., 34(2): 131-136.

White, T., 2009. Hadoop: The Definitive Guide O'Reilly Media, Inc., Sebastopol, Calif.

Yu, Y. and X. Wang, 2015. World cup 2014 in the twitter world: A big data analysis of sentiments in U.S. sports fans' tweets. Comput. Human Behav., 48: 392-400.