CSCanada
Management Science and Engineering
Vol. 10, No. 4, 2016, pp. 62-67
DOI:10.3968/9000

ISSN 1913-0341 [Print]
ISSN 1913-035X [Online]
www.cscanada.net
www.cscanada.org

Research on Visual Analysis of Big Data Based on CiteSpace III

ZHANG Xuehong[a],*

[a]School of Business Administration, South China University of Technology, Guangzhou, China.
*Corresponding author.

Received 5 September 2016; accepted 2 November 2016
Published online 26 December 2016

Abstract
This paper makes visual analysis on big data retrieval literature by using the information visualization tool CiteSpace III and 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 the future may include Hadoop Distributed File System, Hadoop Database, performance evaluation and medical research.

Key words: Big data; CiteSpace III; Research focus; Evolution paths; Visualization

Zhang, X. H. (2016). Research on Visual Analysis of Big Data Based on CiteSpace III. *Management Science and Engineering*, *10*(4), 62-67. Available from: URL: http://www.cscanada.net/index.php/mse/article/view/9000
DOI: http://dx.doi.org/10.3968/9000

INTRODUCTION
In September 2008, "Big Data" special issue was published in *Nature*, then the research and applications of big data quickly 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 (Lohr, 2012; Feng et al., 2013; Mcafee et al., 2012). It is necessary for us to find out what are the hot research topics, the major fields of study, and research trends in future. With the above purposes, this paper made visual analysis by knowledge maps based on CiteSpace III.

1. DATA SOURCES AND RESEARCH METHODS
In this paper, 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 2,970 records were acquired for further analysis. These records come from 1,744 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 data was 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 is 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).

2. LITERATURE DISTRIBUTION STATISTICS

2.1 Annual Distribution

The annual distribution statistics are as shown in Figure 1.

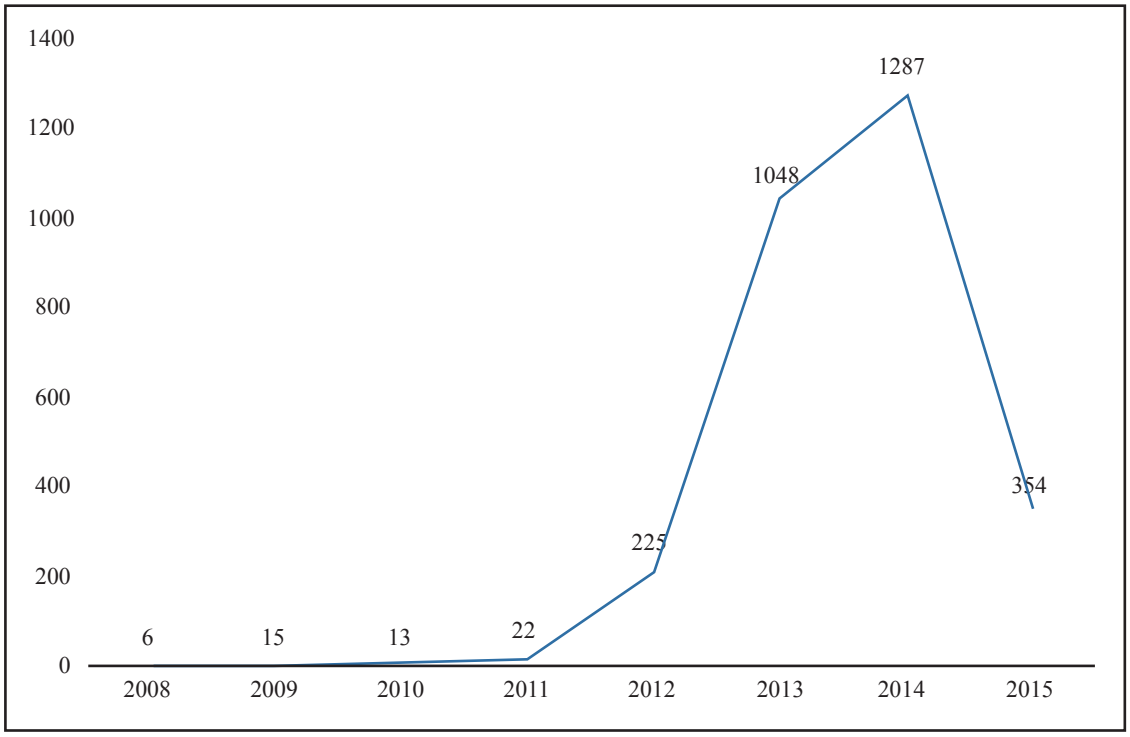

Figure 1
Quantity of Publications From 2008 to 2015

From Figure 1, we can divide the research of "big data" into three stages: a) From 2008 to 2011, it belongs to the initial stage with only 56 records in total. b) 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. c) 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.

2.2 Regional Distribution

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.

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

2.3 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.

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

3. RESEARCH FOCUS, FIELDS AND FRONTS ANALYSIS

3.1 Research Focus Distribution

In 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 research focus at a time. In 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.

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

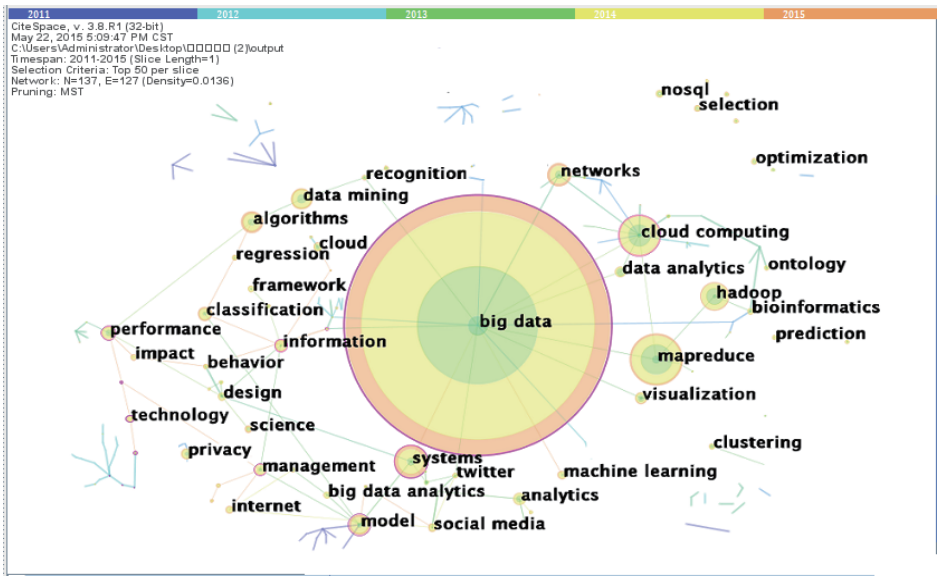

Figure 2
Knowledge Map of Research Focus in Big Data Field

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

3.2 Major Fields of Big Data Research

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

(a) 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 fields mainly study various technologies 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, 2015; Hashem et al., 2015; Ge et al., 2014; Shivhare, Mishra, & Sharma, 2013; Li et al., 2014).

(b) 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, Gray, & Sanchez, 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 to a new business model like the Amazon patent (Leveling, Edelbrock, & Otto, 2014).

(c) 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 mine and to analyze various types of network data for discovering new rules (Tang & 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). Yang 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 to their tweets, particularly, the emotional changes after goals (Yang & Wang, 2015).

(d) 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 & Srivastava, 2014). The study in this field mainly focusses on how to improve the quality of big data to reduce data error and ensure better analysis results.

3.3 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 the 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, Ben Shneiderman et al. (2013) proposed that interactive information visualization and visual analytics methods will bring profound changes to personal health programs, clinical healthcare delivery, and public health policymaking. Brooke l. Fridley et al. (2014) thought that each woman and her cancer are unique, successful cures and outcomes will only come from informative biomarkers/signatures 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, Plaisant, & Hesse, 2013; Fridley, Koeslter, & Godwin, 2014; Raghupathi & Raghupathi, 2014; Anderson & Chang, 2015).

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	

3.4 Evolution Paths Analysis Based on Time-Zone View of Cited Reference Co-Appearance Network

Time-zone view is a kind of knowledge map which places emphasis on the time dimension to show the evolution paths. In 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.

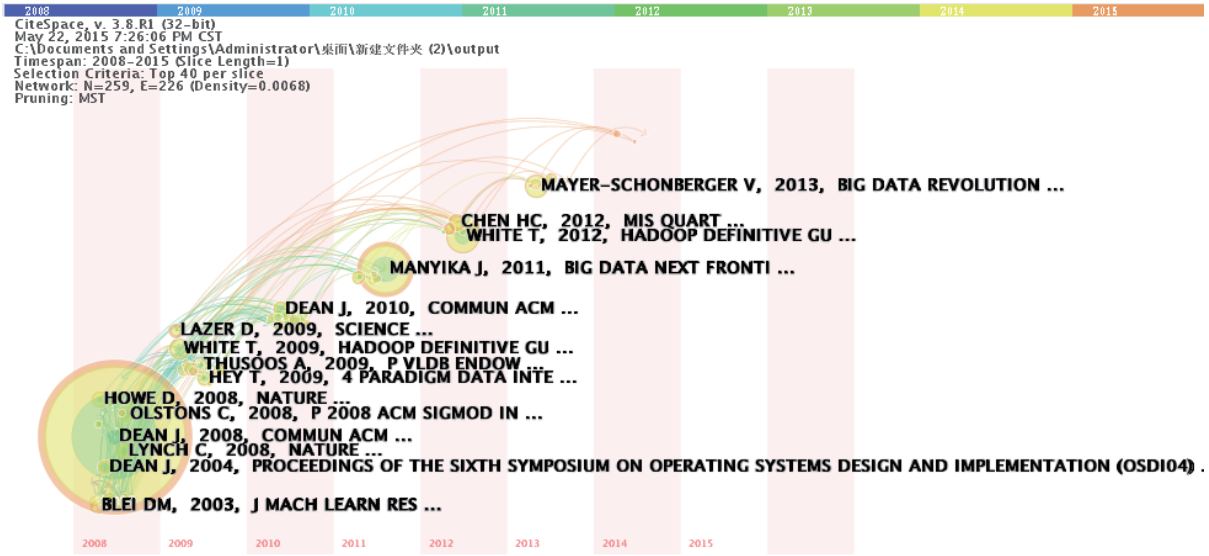

Figure 3
Time-Zone View of Cited Reference Co-Appearance Network (2008-2015)

Table 5
Top 6 Cited References

Citations	Reference information	Reference resource
353	Mapreduce: Simplified data processing on large clusters; 2008; Dean J	Communication of the Acm
130	Big data: the next frontier for innovation, competition, and productivity; 2011; Manyika J	McKinsey Global Institute
118	Hadoop: The Definitive Guide; 2009 ;White T	O'Reilly Media, Inc.
59	Big data: A Revolution That Will Transform How We Live, Work and Think; 2013; Mayer-schonberger	John Murray Publishers Ltd
51	Big data: The future of biocuration; 2008 ; Howe, Doug	Nature
49	Big data: How do your data grow? 2008; Lynch, Clifford	Nature

From Table 5, it can be seen the most frequently cited reference is *Mapreduce: Simplified data processing on large clusters* published by Jeff Dean in 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 & 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-schonberger in 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 & 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

In this paper, 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 the future.

REFERENCES

Almalki, M., Gray, K., & Sanchez, F. M. (2015). The use of self-quantification systems for personal health information: big data management activities and prospects. *Health Inf Sci Syst, 3*(Suppl 1 HISA Big Data in Biomedicine and Healthcare 2013 Con): S1.

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

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology, 57*(3), 359-377.

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication, 64*(2), 317-332.

Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM, 51*(1), 107-113.

Feng, Z. Y., & Guo, X. H., et al. (2013). On the research frontiers of business management in the context of big data. *Journal of Management Sciences in China,* (01), 1-9.

Fridley, B. L., Koeslter, D. C., & Godwin, A. K. (2014). Individualizing care for ovarian cancer patients using big data. *Journal of the National Cancer Institute, 106*(5), dju080.

Ge, B., Ge, S., & Minard, T. (2014). Visualizations make big data meaningful. *Communications of the ACM, 57*(6).

Hashem, I. A. T., Yaqoob, I., & Anuar, N. B., et al. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems, 47,* 98-115.

Hazen, B. T., Boone, C. A., & Ezell, J. D., et al. (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. *International Journal of Production Economics, 154,* 72-80.

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

Kwon, O., et al. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. *International Journal of Information Management, 34*(3), 387-394.

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

Li, X. L. (2015). A survey on big data systems. *SCIENTIA SINICA Informationis,* (01), 1-44.

Li, F., Ooi, B. C. O., & Zsu, M. T., et al. (2014). Distributed data management using MapReduce. *ACM Computing Surveys (CSUR), 46*(3), 31.

Lohr, S. (2012). The age of big data. *New York Times,* p.11.

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

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think.* Houghton Mifflin Harcourt.

Mcafee, A., Brynjolfsson, E., & Davenport, T. H., et al. (2012). Big data: The management revolution. *Harvard Bus Rev., 90*(10), 61-67.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. *Health Information Science and Systems, 2*(1), 3.

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

Shivhare, H., Mishra, N., & Sharma, S. (2013). *Cloud computing and big data* (pp.222-225). Proceedings of 2013 International Conference on Cloud, Big Data and Trust.

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

Tang, J., & Chen, W. G. (2015). Deep analytics and mining for big social data. *Chinese Science Bulletin, 60*(5/6), 509-519.

Wang, B. L. (2015). Research on big data based on scientometrics and visualization analysis. *Journal of Intelligence, 34*(2), 131-136.

Yu, Y., & Wang, X. (2015). World cup 2014 in the twitter world: A big data analysis of sentiments in U.S. sports fans' tweets. *Computers in Human Behavior, 48,* 392-400.