Journal of Mobile Multimedia, Vol. 1, No.1&2 (2015) 103-118

TRACING FAST-CHANGING LANDSCAPE OF STUDY ON BIG DATA

WEIGUANG WANG

Institute of Policy and Management, Chinese Academy of Sciences, Beijing 100190, P.R. China

XI ZHANG

College of Management and Economics, Tianjin University, Tianjin 300372, P.R. China

jackyzhang@tju.edu.cn

PATRICIA ORDONEZ de PABLOS

Department of Business Administration, University of Oviedo, Oviedo, Spain

JINGHUAI SHE

Capital University of Economics and Business, Beijing, P.R. China

This study traced the fast-changing landscape of study on big data. Three hottest aspects of big data study are: technologies, application to academic study and real value in big data. With the rising of mobile multimedia and social media, more and more data were produced in our daily life. Traditional methods and technologies are inefficient or even powerless facing such large scales of data. Also, academics gained greater abilities to produce much more data than before, but had few good methods of analysing big data. Therefore, new technologies targeting efficient utilization of big data are being proposed, like MapReduce, Hadoop and so on. Methods of applying big data to academic study are also being discussed intensively. Furthermore, the political and business value hidden in big data is very attractive. People believe that although some profits have been made, much more are hidden in big data to be dug out. In this paper, current stage, influential references, outstanding authors, important institutions, top-tier journals and evolution of hot topics are all analysed with help of CiteSpace to map big data study. Finally, forecasting on development of big data study was made to provide help for future study.

1 Introduction

Recently, more and more research focus on big data to promote technologies of handling large scale of data, to apply big data technologies into academic study or to realize business value hidden in big data. Big data “usually include data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process the data within a tolerable elapsed time” [25]. While Manyika, et al. [18] defined big data as “large pools of data that can be captured, communicated, aggregated, stored, and analysed”. Early in 2001, Laney [15] defined data growth challenges and opportunities as being three-dimensional. Now, tens of thousands of trillions bytes are being produced from time to time and large scale data become popular all around the world [8, 20, 24, 31]. Since these big data cannot be coped with easily using traditional methods [10, 26, 30], more attention was paid on big data study.

Firstly, big data brought some technical challenges for researchers to resolve. Dean and Ghemawat [6] published their introduction of "MapReduce" firstly in 2004. Then after four years' operation, they discussed again about their simplified data processing method. Their work set important technical base for big data study. After them, an open-source software framework named Apache Hadoop was provided [28]. Hadoop, which was created by [2], was derived from Google's MapReduce and Google File System papers. And Also, other technologies were discussed for big data study [9, 13]. And for application of big data study, big-data science was a pioneer [1, 17, 23, 27]. Schadt, et al. [22] tried to use cloud and heterogeneous computing methods to successfully tackle the problem of large scale of data in academic research. Howe, et al. [12] proposed three urgent actions to advance big data usage in academic study. Hey, et al. [11] spoke highly of big data and named data-intensive method as the forth paradigm in scientific discovery. Big data study also pursues real politic or business value hidden big data. White House made 84 different big data programs in six departments [14]. And big data played a great role in the victory of Obama's re-election campaign [19]. Business value in big data is much more highly agreed [3, 16, 21, 29]. Manyika, et al. [18] claimed that big data can bring great value in different domains: health care, public sector administration, retail, manufacturing and personal location data.

All these researches strengthened the huge potential of big data study. Technologies it needs [7], applications in academic research [22] and politic or business value hidden in it [18] are all important branches of big data study. However, it is not that clear about the way that big data study emerged. Trends of big data study are also in need of a good analysis. By providing a clear description on development of big data study at early the early stage of it, contributions can be made to accelerate its development. Furthermore, more and more academics are coming into big data study recently. While big data study keeps changing fast due to its early stage. Current status, influential references, outstanding authors, important institutions, top-tier journals and hot topics in big data study should be concluded to help researchers pick prosper topics up.

2 Current Status of Big Data Study

2.1. Learning at Universities

Web of Science (WoS) provides basic information of important academic literatures in many fields, including titles, authors, journals (or other source), abstracts as well as citation data. Also, WoS has some functions to do basic analysis on selected dataset. Using WoS, we mapped big data study's current status represented in literatures.

We used the term "big data" for topical searches on WoS. Timespan was set "all years", "Science Citation Index Expanded (SCI-EXPANDED) --1900-present", "Social Science Citation Index (SSCI)--1996-present" and "Conference Proceedings Citation Index- Science (CPCI-S) --1990-present" were selected for Database. Finally, 15744 records were found. After checking the result, we found too many papers in other fields were concluded in. Obviously, it is impossible for such a young research field to contain that many papers. Therefore, we made some refining to get target dataset. Firstly, search area was limited in title instead of topic (which covers the title, keywords and abstracts). And 332 records were got. Secondly, we substitute "big-data" for the term "big data" and 487 records were found. The last two results, "big data" in title (Dataset I) and "big-data" (Dataset II) were adopted for depiction of current status of big data study.

2.2. Depiction of Current Status of Big Data Study

Figure 1 and 2 are the numbers of publications and citations of Dataset I; while Figure 3 and 4 are the ones of Dataset II. According to both of them, big data study started far before, formed after 2005 and burst in 2012. Influence of big data study began before 2000, and became high around 2008.

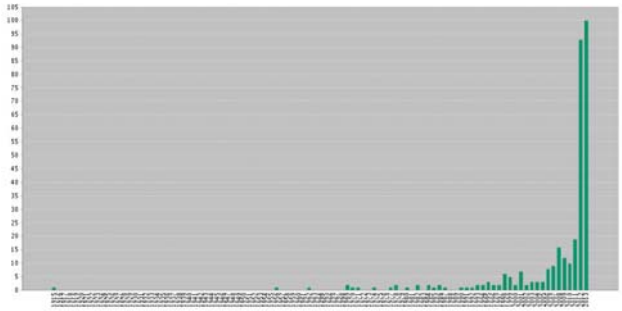

Figure 1 Published papers in each year in Dataset I

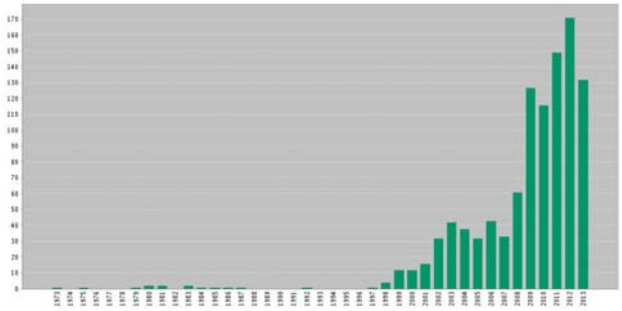

Figure 2 Citations in Each Year in Dataset I

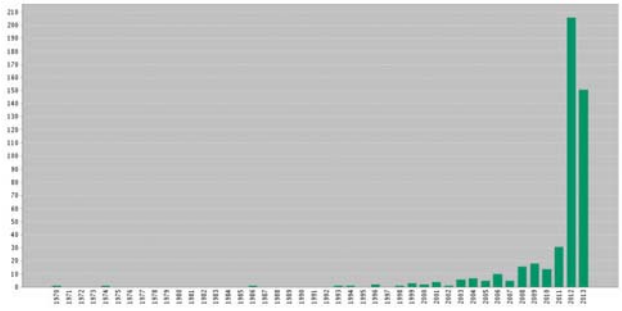

Figure 3 Published Papers in Each Year in Dataset II

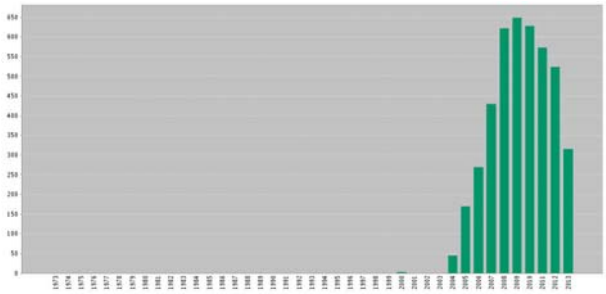

Figure 4 Citations in Each Year in Dataset II

Table 1 Top Authors in Dataset I

Field: Authors	Record Count	% of 332
ANONYMOUS	9	2.711 %
BERGH J	3	0.904 %
BOERI R	3	0.904 %
BONNEFOI H	3	0.904 %
KEIM D	3	0.904 %
MAURIAC L	3	0.904 %
TUBIANA-HULIN M	3	0.904 %

Table 2 Top Authors in Dataset II

Field: Authors	Record Count	% of 487
ANONYMOUS	5	1.027 %
HELBING D	4	0.821 %
DAI L	3	0.616 %
DAVENPORT TH	3	0.616 %
FENG L	3	0.616 %
INDYK W	3	0.616 %
KAJDANOWICZ T	3	0.616 %
KAZIENKO P	3	0.616 %
LIU H	3	0.616 %
MODELSKI J	3	0.616 %
ROMANIUK R	3	0.616 %
SCHADT EE	3	0.616 %
XIAO JF	3	0.616 %
ZHANG Z	3	0.616 %
ZHOU X	3	0.616 %

From the perspective of quantity, the top authors contributing to big data study are HELBING D and so on (Table 1 and 2). However, HELBING D does not have enough advantage than ones who has three publications. The top five institutions in big data study are "UNIVERSITY OF CALIFORNIA SYSTEM", "HARVARD UNIVERSITY", "MASSACHUSETTS INSTITUTE OF TECHNOLOGY MIT", "CHINESE ACADEMY OF SCIENCES" and "STANFORD UNIVERSITY" (Table 3 and Table 4). The journals or conferences paying most attention on big data study are NATURE and COMPUTER.

Table 3 Top Organizations in Dataset I

Field: Organizations-Enhanced	Record Count	% of 332
UNIVERSITY OF CALIFORNIA SYSTEM	11	3.313 %
HARVARD UNIVERSITY	9	2.711 %
MASSACHUSETTS INSTITUTE OF TECHNOLOGY MIT	9	2.711 %
CHINESE ACADEMY OF SCIENCES	6	1.807 %
STANFORD UNIVERSITY	6	1.807 %

Table 4 Top Organizations in Dataset II

Field: Organizations-Enhanced	Record Count	% of 487
UNIVERSITY OF CALIFORNIA SYSTEM	20	4.107 %
HARVARD UNIVERSITY	14	2.875 %
CHINESE ACADEMY OF SCIENCES	10	2.053 %
MIT	10	2.053 %
STANFORD UNIVERSITY	9	1.848 %

Table 5 Top Sources in Dataset I

Field: Source Titles	Record Count	% of 332
NATURE	16	4.819 %
COMPUTER	10	3.012 %
COMMUNICATIONS OF THE ACM	8	2.410 %
COMMUNICATIONS IN COMPUTER AND INFORMATION SCIENCE	6	1.807 %
ECONTENT	6	1.807 %

Table 6 Top Sources in Dataset II

Field: Source Titles	Record Count	% of 487
NATURE	16	3.285 %
COMPUTER	13	2.669 %
SECOND INTERNATIONAL CONFERENCE ON CLOUD AND GREEN COMPUTING SECOND INTERNATIONAL CONFERENCE ON SOCIAL COMPUTING AND ITS APPLICATIONS CGC SCA 2012	10	2.053 %
PROCEDIA COMPUTER SCIENCE	9	1.848 %
2012 IEEE CONFERENCE ON VISUAL ANALYTICS SCIENCE AND TECHNOLOGY VAST	8	1.643 %
COMMUNICATIONS IN COMPUTER AND INFORMATION SCIENCE	8	1.643 %
HARVARD BUSINESS REVIEW	8	1.643 %
IEEE CONFERENCE ON VISUAL ANALYTICS SCIENCE AND TECHNOLOGY	8	1.643 %
PROCEEDINGS OF SPIE	8	1.643 %

3 Analysis on Trends of Big Data Study

To map trends of big data study, citation analysis should be included in. Citation data of the used dataset can provide information in a much larger number of relevant publications. According to citation analysis, influential references, outstanding authors, important institutions, top-tier journals and hot topics in big data study can be picked out.

Citation data in the timespan from 1990 to 2013 were taken in our analysis. And in order to make the result more accurate, we combined the two dataset in our study together using CiteSpace to show the whole trend of big data study clearly. CiteSpace is software used to analyse co-citation networks. It can give out statistic data and visualization of input dataset [4, 5]. During the analysis, we took one year as a slice and top 100 cited terms in each slice were used for visualization. Some important indicators were provided by CiteSpace. “Centrality” is “a metric of a node in a network that measures how likely an arbitrary shortest path in the network will go through the node”; “Burst” is “single or multi-word phrases extracted from the title, abstract, or other fields of a bibliographic record and the frequency of the term bursts, i.e. sharply increases, over a period of time” and “Half-life” is “the number of years that a publication receives half of its citations since its publication”.

In order to perform the study records, all students' smartphones need to connect to the network. In order to connect the smartphones to the network, the information outlet for each seat is required. It is difficult to install the information outlets in all lecture rooms. Moreover, maintenance and deployment of many information outlets require a lot of expenses.

However, now it is easy to make Wi-Fi connection with mobile terminals such as smartphones. We installed the terminal adopter for Wi-Fi connection to improve the network environment for the lecture room. It should be noted that many devices can make Wi-Fi connection simultaneous during the lecture. Figure 3 shows the network environment for the learning recorder system.

3.1. Lecturer's smartphone application

Figure 5 showed the main cluster in full spot of references. From the figure, a conclusion can be made that big data study is at a very early stage. All the most cited references are not together. This phenomenon indicates core references of big data study are all set up on base of other fields. They have not yet constructed a system to contain some branches around a powerful center. They must spend some time to establish a core first, which will combine the most influential works about big data together. Then more and more branches will emerge around this core.

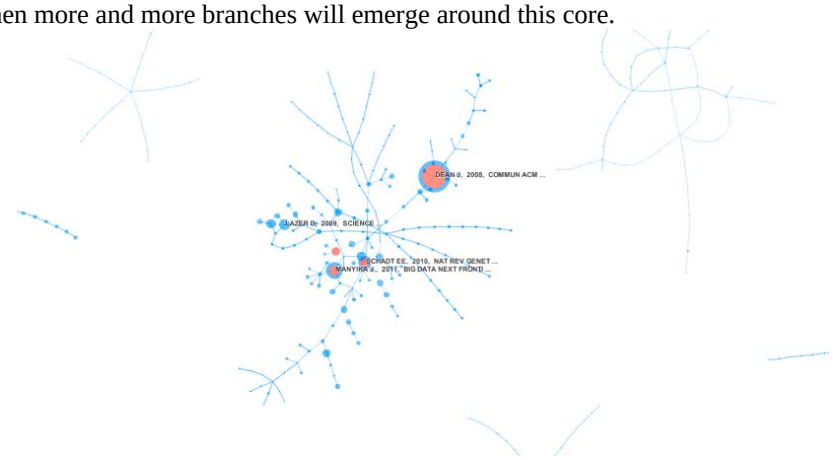

Figure 5 Main References Cluster on Big Data Study

Table 7 listed top 5 references in frequency. “MapReduce: simplified data processing on large clusters” is published by Jeffrey Dean and Sanjay Ghemawat in 2008 on “Communications of the ACM” is the first one. In this paper, an application named MapReduce used by Google was introduced again about its function to simplify big data processing. In fact, MapReduce was designed and implemented by Jeffrey Dean and Sanjay Ghemawat in 2004. They introduced this application in there paper on the “Sixth Symposium on Operating System Design and Implementation” in December 2004. This work is critical for big data study. It has very far-reaching influence in this field. Also, it is a very hot topic recently according to its “Burst”.

“Big data: The next frontier for innovation, competition, and productivity” is a report done by James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh and Angela Hung Byers who worked in McKinsey Global Institute in 2011. In this report, they “examine the potential value that big data can create for organizations and sectors of the economy and seeks to illustrate and quantify that value.” Although this work has been published for only less than two years, it attracted quite many attentions and is going to attract more and more in recent future.

"Computational solutions to large-scale data management and analysis" was composed by Eric E. Schadt, Michael D. Linderman, Jon Sorenson, Lawrence Lee, and Garry P. Nolan on "Nature Reviews Genetics" in 2010. In this article, authors tried to use cloud and heterogeneous computing methods to successfully tackle the problem of handling with big data. "Big data: The future of biocuration" is also a work about the lag of curation to the big data generation in biological study on "Nature" in 2008. Doug Howe and fourteen other authors proposed three urgent actions to advance this key field in this paper. They are all important works to provide solutions about usage of big data in academic research. But the first one is relatively more influential according to the number of citation.

"Computational Social Science" is a research published on "Science" in 2009 which was done by David Lazer and fourteen other authors. This paper claimed that big data can uncover patterns of individual and group behaviours. This work has quite far-reaching influence in this field, too.

Table 7 Top 5 References in Big Data Study

Freq	Burst	Centrality	Google Scholar Citation	Author	Year	Title	Source	Half-life
23	10.53	0	7964	Dean J	2008	MapReduce: simplified data processing on large clusters	COMMUN ACM	4
12	5.45	0	237	Manyika J.	2011	Big data: The next frontier for innovation, competition, and productivity	Report	1
10	3.77	0	138	Schadt EE	2010	Computational solutions to large-scale data management and analysis	NAT REV GENET	1
8		0	23	Lazer D	2009	Computational Social Science	SCIENCE	3
8		0	173	Howe D	2008	Big data: The future of biocuration	NATURE	1

3.2. Cited Authors

Contrasts to top references, outstanding authors dominate the core of big data study (figure 7 and 8). Most important authors are listed in Table 8 and 9. "DEAN J" and "Dean J." refer to the same one, Jeff Dean. He is a computer scientist and software engineer. He is a Google Fellow in the Systems and Infrastructure Group. He is very eminent in designing systems, computing infrastructure and other applications. He has contributed to projects in Google, including "Spanner", "Google Translate", "Big Table", "MapReduce" and "Google Brain". He is the most influential one in big data study now.

Eric Emil Schadt is a mathematician and computational biologist who founded the Icahn Institute for Genomics and Multiscale Biology. He is also chair of the Department of Genetics and Genomics Sciences at Mount Sinai School of Medicine. His work focused on supercomputing and its application in biological areas.

James Manyika works in McKinsey Global Institute and he is the first author of “Big data: The next frontier for innovation, competition, and productivity”, which made great impact on big data study.

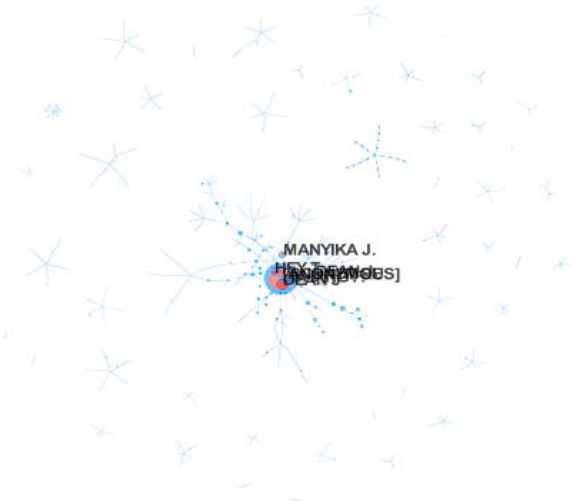

Figure 7 Full Shot of Most Cited Authors

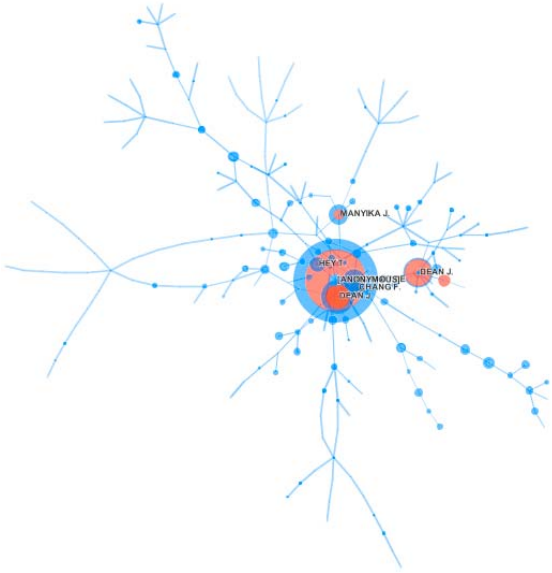

Figure 8 Main Authors Cluster on Big Data Study

Fay W. Chang is another researcher in Google. Her research area is Distributed Systems and Parallel Computing. She is the first author of a very influential publication - “Bigtable: A Distributed Storage System for Structured Data” in 2006. In this publication, Jeff Dean was listed as the second author.

White Tom is the author of “Installing Apache Hadoop - Hadoop: The Definitive Guide”. In this publication, an open-source software framework, Apache Hadoop is introduced. Hadoop can support

data-intensive distributed applications and Hadoop was derived from Google's MapReduce and Google File System papers. White Tom is becoming more and more popular now revealing great impact of Apache Hadoop.

Table 8 Top 5 Authors in Big Data Study in Frequency

Freq	Burst	Centrality	Author	Half-life
57	15.93	0	[Anonymous]	18
24	8	0	DEAN J	2
20	5.61	0	Dean J.	6
16	4.44	0	Schadt EE	0
13	5.06	0	Manyika J.	1
11	3.58	0	Chang F.	6

Table 9 Top 5 Authors in Big Data Study in Burst

Freq	Burst	Centrality	Author	Half-life
57	15.93	0	[Anonymous]	18
24	8	0	DEAN J	2
20	5.61	0	Dean J.	6
13	5.06	0	Manyika J.	1
16	4.44	0	Schadt EE	0
8	4.17	0	White Tom	3

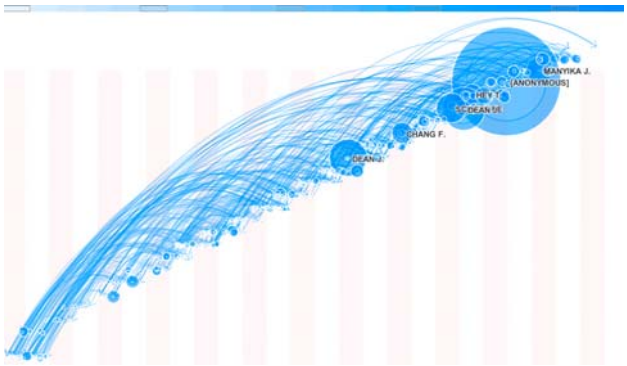

Figure 9 Time Zone Mode of Most Cited Authors

Besides, in the time zone mode of visualization (Figure 9), all the important authors appeared only after 2004. Combining with figure 7, this claims that central cluster of authors has already formed. Development of big data study is very quick. It is predictable that exploration of this field will become much stronger.

3.3. Lecture flow

The top-tier journals in big data study are NATURE, SCIENCE, COMMUN ACM, LECT NOTES COMPUT SC and so on (Table 10). They were cited most. While some ones are very hot to be cited more and more. COMMUN ACM, LECT NOTES COMPUT SC and HADOOP DEFINITIVE GU are attracting more attentions now (Table 11). Among these ones, HADOOP DEFINITIVE GU is a book which has been introduced in 3.2. According to the "Half-life", COMMUN ACM is the most far-reaching journal in big data study. Furthermore, journal cluster of big data study developed quite well, however, more time is needed to make journals like COMMUN ACM and LECT NOTES COMPUT SC to be more dominant in big data study.

Figure 10 Main Journals Cluster on Big Data Study

Table 10 Top 10 Authors in Big Data Study in Frequency

Freq	Burst	Cent-rality	Year	Source	Half-life
69		0	1993	NATURE	4
65		0	1995	SCIENCE	2
57	10.54	0	1983	COMMUN ACM	12
50	6.61	0	2004	LECT NOTES COMPUT SC	1
46		0	1997	P NATL ACAD SCI USA	2
28		0	1999	BIOINFORMATICS	4
25		0	2003	NAT REV GENET	0
24		0	1995	MACH LEARN	5
24		0	2011	PLOS ONE	0
23		0	2007	BMC BIOINFORMATICS	1

Table 11 Top 9 Authors in Big Data Study in Burst

Freq	Burst	Cent-rality	Year	Source	Half-life
57	10.54	0	1983	COMMUN ACM	12
50	6.61	0	2004	LECT NOTES COMPUT SC	1
13	5.47	0	2010	HADOOP DEFINITIVE GU	1
15	4.33	0	1997	PROC INT CONF DATA	6
14	4.23	0	2009	VLDB	1
14	4.23	0	2005	SIGMOD	6
13	3.83	0	2009	PVLDB	2
9	3.61	0	2003	VLDB J	9
13	3.38	0	2002	IEEE INTERNET COMPUT	9

3.4. Hot Topics in Big Data Study

Table 12 Top 10 Authors in Big Data Study in Frequency

Freq	Burst	Cent-rality	Keyword	Year
95	18.21	0	big data	2009
26	3.61	0	cloud computing	2009
25	8.26	0	mapreduce	2010
13		0	visualization	2001
12	3.26	0	hadoop	2012
11		0	model	1999
10		0	systems	2011
9		0	performance	2011
9		0	design	2006
9		0	networks	2010

The important topics in big data study are “big data”, “cloud computing”, “mapreduce”, visualization”, “hadoop” and so on (Table 12). Obviously, “mapreduce”, “cloud computing” and “hadoop” are hottest ones. This can be explained by the early stage of big data study. This young field obtained some important technologies now but usage of big data still has no united way. Structure of these topics can found in figure 11. In order to trace development of big data study, we all drew the evolutionary map showing status of big data study in each year (Figure 12). Topic of big data occurred in 2008. Then in 2009, 2010 and 2011, topics about big data study developed slowly. That is because researchers were focusing on uncovering basic characters of big data. After that, big data study burst in 2012 and 2013. It will keep increasing rapidly in recent future.

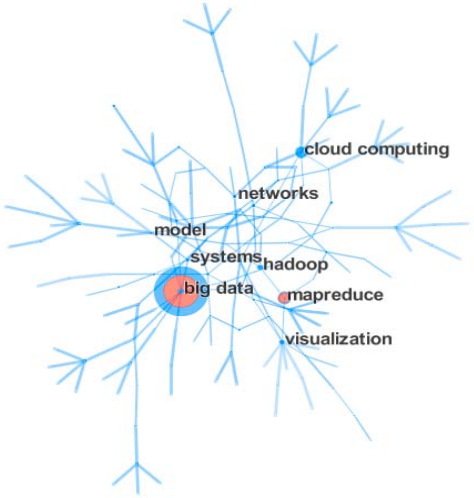

Figure 11 Main Topics Cluster on Big Data Study

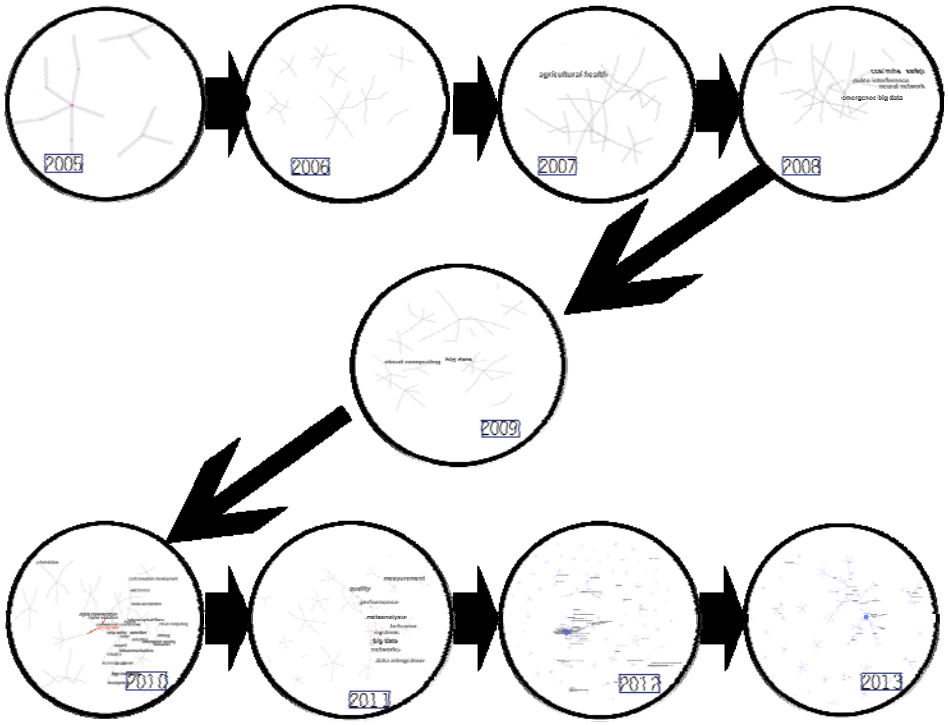

Figure 12 Evolution of Big Data Study

4 Discussion

1. Big data study is at its early stage of development. Current big data study is too new now to form a complete system. Basic theories in other fields need to be used in big data study. But authors have formed a powerful core to support big data study. This is a good sign of future rapid development. Also, after several years' incubation, topics connected with big data are exploding in the past two years. All the centrality indicators of references, authors, journals and topics are 0 in the analysis. This states the young stage of big data study, too.

2. The most influential references in big data study are "MapReduce: simplified data processing on large clusters", "Big data: The next frontier for innovation, competition, and productivity", "Computational solutions to large-scale data management and analysis" and so on. They attracted most attention in this field. In fact, they represent the three important branches of big data study: the technologies, application in academic research and politic or business value hidden in big data.

3. The most outstanding authors are Jeff Dean, Eric Emil Schadt, James Manyika and so on. They are settled in center of authors' main cluster of big data study. They lead trend of big data study. Jeff Dean contributed to basic technologies on big data; Eric Emil Schadt proposed possible ways to apply big data into academic study; James Manyika discussed business value of big data and the way to realize it.

4. The top-tier journals in big data study are NATURE, SCIENCE, COMMUN ACM, LECT NOTES COMPUT SC and so on. Among these journals, COMMUN ACM and LECT NOTES COMPUT SC should be treated as core ones in big data study in the future, while NATURE and SCIENCE do not focus on big data study only. Some time is needed to make COMMUN ACM and LECT NOTES COMPUT SC enjoy much more impact.

5. "Mapreduce", "cloud computing" and "hadoop" are hottest topics connected to big data. It is obvious that study of big data focus on techniques now. Big data study did not occurred as an important keyword until 2008. Then, after three years' incubation, topics connected to big data burst. It is fair to say that current status of big data study was shaped in 2008 to 2011.

Overall, big data study started far before, formed after 2005 and burst in 2012. Influence of big data study began before 2000, and became high around 2008. As a young academic field, big data study formed some core technologies. A system of big data references still needs some time to be established. Core cluster of authors is already there studying topics about big data. It is predictable that big data will keep busting in recent years. Three most important branches of big data study were found in this paper. Technologies of big data, which aims to provide more powerful tools for manage large scale of data, are becoming more and more. Scientists in academic research also play an outstanding role in applying theories of big data into academic study. In this branch, some important work has been done. Finally, another potential branch, exploring business and political value hidden in big data, which is still lack of critical theories, faces many challenges which is also grate opportunities for both academic research and business benefit.

5 Conclusions

This paper traced the fast-changing landscape of big data study with help of WoS and CiteSpace. Current status of big data study was depicted by drawing trends of publications and citations in big

data study. And top authors, institutions, journals were quantitatively listed to show its young stage. Big data study is relatively new and fast-changing. It became visible as an independent term in important literature after 2004 and began to burst in 2012. Then trends of big data study development were analysed. Influential references, outstanding authors, top-tier journals and hot topics in big data study were carefully studied. Three branches were found in recent big data study. Finally, evolution of topics in big data study was mapped to show its process.

Acknowledgements

The work described in this paper was supported by the National Natural Science Foundation of China (No: 71201155), and the Science Foundation of Ministry of Education of China (Grant No. 13YJA630073). Additional Funding was received from the Selected Seed Foundation of Tianjin University (No. 60306088).

References

[1] C. Snijders, U. Matzat, and U. Reips, "Big data': Big gaps of knowledge in the field of Internet science," *International Journal of Internet Science,* vol. 1, pp. 1-5, 2012.
[2] J. Manyika *et al.*, *Big data: The next frontier for innovation, competition, and productivity*, The McKinsey Global Institute, 2011.
[3] D. Laney, *3D Data Management: Controlling Data Volume, Velocity, and Variety*, META Group, 2001.
[4] C. Doctorow, "Big data: Welcome to the petacentre," *Nature,* vol. 455, no. 7209, pp. 16-21, 2008.
[5] A. Setyono, M. J. Alam, and C. Eswaran, "Study and Development of the Transmission Method for Large Multimedia File Size Using MMS Technology Study and Development of the Transmission Method for Large Multimedia File Size Using MMS Technology," *Journal of Mobile Multimedia,* vol. 8, no. 1, pp. 001-024, 2012.
[6] A. Noman, and C. Adams, "Providing A Data Location Assurance Service for Cloud Storage Environments," *Journal of Mobile Multimedia,* vol. 8, no. 4, pp. 265-286, 2013.
[7] X. Zhang, D. Vogel, and Z. Zhou, "Effects of Information Technologies, Department Characteristics and Individual Roles on Improving Knowledge Sharing Visibility: A Qualitative Case Study," *Behaviour & Information Technology,* vol. 31, pp. 1117-1131, 2012.
[8] O. Trelles *et al.*, "Big data, but are we ready?," *Nat Rev Genet,* vol. 12, no. 3, pp. 224-224, 2011.
[9] X. T. Guo *et al.*, "Chaos Theory as a Lens for Interpreting Blogging," *Journal of Management Information Systems,* vol. 26, no. 1, pp. 101-127, Sum, 2009.
[10] X. Zhang, P. de Pablos, and Y. Zhang, "The Relationship between Incentives, Explicit and Tacit Knowledge Contribution in Online Engineering Education Project," *International Journal of Engineering Education,* vol. 28, pp. 1341-1346, 2012.
[11] J. Dean, and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," *To appear in OSDI*, pp. 1, 2004.
[12] T. White, *Hadoop: the definitive guide*: O'Reilly, 2012.
[13] A. Bialecki *et al.*, "Hadoop: a framework for running applications on large clusters built of commodity hardware," *Wiki at http://lucene. apache. org/hadoop,* vol. 11, 2005.
[14] J. T. Dudley, and A. J. Butte, "In silico research in the era of cloud computing," *Nat Biotech,* vol. 28, no. 11, pp. 1181-1185, 2010.
[15] A. Jacobs, "The pathologies of big data," *Commun. ACM,* vol. 52, no. 8, pp. 36-44, 2009.
[16] C. Lynch, "Big data: How do your data grow?," *Nature,* vol. 455, no. 7209, pp. 28-29, 2008.
[17] M. Waldrop, "Big data: Wikiomics," *Nature,* vol. 455, no. 7209, pp. 22-25, 2008.

[18] B. Allen *et al.*, "Software as a service for data scientists," *Commun. ACM*, vol. 55, no. 2, pp. 81-88, 2012.
[19] M. C. Schatz, "CloudBurst: highly sensitive read mapping with MapReduce," *Bioinformatics (Oxford, England)*, vol. 25, 2009.
[20] E. E. Schadt *et al.*, "Computational solutions to large-scale data management and analysis," *Nature Reviews Genetics*, vol. 11, no. 9, pp. 647-657, 2010.
[21] D. Howe *et al.*, "Big data: The future of biocuration," *Nature*, vol. 455, no. 7209, pp. 47-50, 2008.
[22] A. J. Hey, S. Tansley, and K. M. Tolle, "The fourth paradigm: data-intensive scientific discovery," 2009.
[23] T. Kalil. "Big Data is a Big Deal."
[24] B. S. M. News. "How big data analysis helped President Obama defeat Romney in 2012 Elections."
[25] G. Lotan *et al.*, "The revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions," *International Journal of Communication*, vol. 5, pp. 1375-1405, 2011.
[26] M. Savage, and R. Burrows, "The coming crisis of empirical sociology," *Sociology*, vol. 41, no. 5, pp. 885-899, 2007.
[27] D. Boyd, and K. Crawford, "Critical Questions for Big Data," *Information, Communication & Society*, vol. 15, no. 5, pp. 662-679, 2012/06/01, 2012.
[28] X. Zhang, P. de Pablos, and Q. Xu, "Knowledge Sharing Visibility in Electronic Knowledge Management Systems: An Empirical Investigation," *Computers in Human Behavior*, vol. 29, pp. 307-313, 2013.
[29] J. Dean, and S. Ghemawat, "MapReduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107-113, 2008.
[30] C. Chen, "Searching for intellectual turning points: Progressive knowledge domain visualization," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, pp. 5303-5310, 2004.
[31] C. Chen, "CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 3, pp. 359-377, 2006.