The major research themes of big data literature:

From 2001 to 2016

Louis Y. Y Lu
College of Management
Innovation Center for Big Data and Digital Convergence
Yuan Ze University
Taoyuan, Taiwan
louislu@saturn.yzu.edu.tw

John, S. Liu
Graduate Institute of Technology Management
National Taiwan University of Science and Technology
Taipei, Taiwan
johnliu@mail.ntust.edu.tw

***Abstract*—The paper adopts a citation-based analysis to survey the big data relevant literature from 2001 to 2016 to identify the major research themes over the past decade. Databases of Web of Science (WOS) and Scopus are used to retrieve the relevant articles. The edge-betweenness clustering technique is applied to the citation network that is built from the retrieved data. The major research groups in the big data field are identified via the clustering. A word clouding is used to extract the major research themes of each group by calculating the frequencies of words used in the titles of articles. Eleven major and minor research themes are identified: “ cloud computing”, “data mining”, “MapReduce”, “privacy”, “social media”, “healthcare”, “ethics and legal”, “surveillance”, “geography”, “business intelligence”, and “bioinformatics”. The approaches presented in this paper can also be applied to any scientific or technological field.**

Keywords—big data; data mining; cloud computing; data analytics; citation analysis

I. INTRODUCTION

Big data is one of the hottest topics for both researchers and practitioners over the past years. Corporates and governments investigate structured and unstructured forms of data to seek aids for product and process innovation and policymaking. The processes of big data analysis include acquisition, extraction, integration, modeling, analysis, interpretation, and deployment. Many research studies focus on part of these steps and could not see the whole picture of the big data literature. This paper adopts a citation-based bibliometric approach to conducting a complete review of more than three thousand big data related papers, identifying the major research themes over the past 10 years. This review also makes a methodological contribution to the big data field. The results are very valuable for those who want to understand the whole picture of the development of big data and providing the directions for future research.

II. LITERATURE REVIEW

Some researchers have reviewed the literature of big data but only from a specific perspective. Doulkeridis and Nørvåg [1] conducted a survey of large-scale analytical query processing in MapReduce. They categorized existing research on MapReduce improvements according to the specific problem to be solved. Costa [2] reviewed the challenges of using big data in biomedicine and translational science and reported the major breakthroughs achieved. Polato, Goldman and Kon [3] conducted a systematic literature review to Hadoop and classified the main topics addressed in the literature. Fan, Han and Liu [4] examined the features of big data and reported the impacts on statistical and computational methods as well as computing architectures. They suggested various new ideas on the big data analysis and computation. Chen and Lin [5] provided a survey of deep learning and pointed out current research efforts, the challenges and the future trends to big data.

The above-mentioned review papers enrich our understanding and provide insights into various issues of big data. However, when conducting a qualitative research, the authors' knowledge constraints or their subjective judgments limit the contribution. Moreover, only covering a small number of dataset restricted by the method adopted also confines the findings. This study adopts the citation-based edge-betweenness clustering technique to review the 3,898 papers published over 2006 to 2016 and identifies the major research themes in the big data field. We report some descriptive statistics, such as the most influential journals, the growth of papers published.

III. METHOD

The most popular academic citation databases Web of Science (WoS) and Scopus are used to retrieve the relevant papers on 2016/8/15. Time span ranges from 2001 to 2016. The term ‘big data’ is used as the keyword for the query. After removing the articles with no authors and missing data, we got 7,881 and 14,479 articles from WoS and Scopus respectively. When conducting the preliminary analysis, we found that only 4,160 and 6,400 papers construct the citation network. We combine these two datasets by analyzing the data of authors, year, title, journal, volume, issue, page start, page end, and references. The combined dataset consists of 7,520 papers in which 1,120 from WOS, 3,360 from Scopus, and 3,040 from both. Because the method we used is a citation-based, hence the 7,520 data are used for the further analyses.

Some statistics of the final dataset are provided to exhibit which journal contributes more on publishing big data relevant

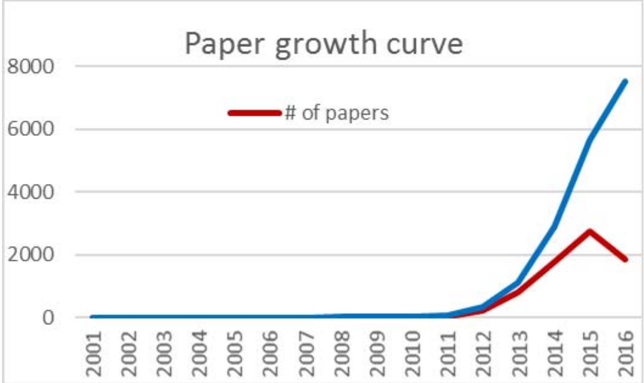
The authors thank the partial financial support from the MOST of Taiwan grant 102-2632-H-155-001-MY3, 102-2410-H-011-008-MY3, and 105-2634-F-155-001.

978-1-5090-4314-9/16 $25.00 © 2016 IEEE
DOI 10.1109/CIT.2016.46

articles and the growth curve of publications. Table 1 shows the data of the number of papers published, g-index, h-index, citations per paper, the active years of publication, and title of the journal. One can judge the contribution of each journal from different viewpoints. Figure 1 and 2 show the growth curve of the number of papers published and the number of authors contributed to big data per year. It presents that the number of publications in big data boomed in 2011 and increased drastically recently.

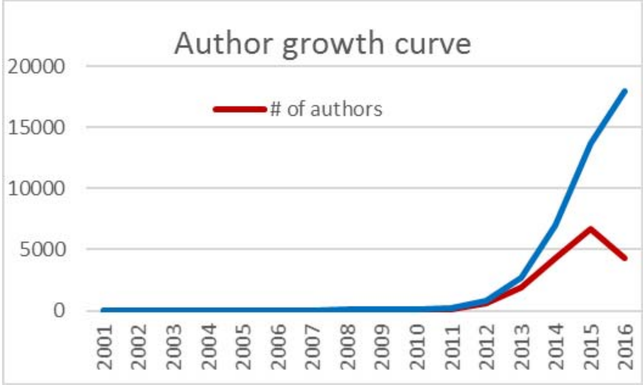
TABLE 1. JOURNAL STATISTICS

Total papers	g-index	h-index	Citations/ paper	Active years	Journal title
36	36	12	38.11	2008~2015	Nature
69	31	17	15.67	2009~2016	PVE
21	21	8	27.33	2012~2016	PNASUSA
23	20	8	19.09	2008~2016	CA
19	19	7	19.05	2012~2016	Science
20	18	5	16.35	2013~2016	ITOKADE
17	17	8	38.82	2012~2015	HBR
17	17	4	254.41	2003~2016	Bioinformatics
68	15	9	4.35	1998~2016	Computer
56	15	9	4.95	2011~2016	PO
30	15	7	8.27	2011~2016	IIS
52	14	9	5.54	2013~2016	FGCS
17	14	8	12.59	2014~2016	HA
14	14	8	14.21	2014~2016	ISPM
29	14	6	7.10	2014~2016	IS
38	13	8	5.05	2014~2016	IN
25	13	5	7.72	2012~2016	IIC
16	13	5	10.56	2013~2016	IS
68	12	6	2.41	2013~2016	BD
41	11	6	3.07	2014~2016	IA



Note: 2016 only counted to Aug.

Fig. 1. Growth curve of papers published



Note: 2016 only counted to Aug.

Fig. 2. Growth curve of authors

The edge-betweenness clustering is used to split the citation network into groups. Girvan and Norman [6] proposed the edge-betweenness clustering technique to cluster a social network. To speed up the process of clustering, they introduced a fast algorithm [7][8] and the concept of modularity to identify the optimal structure of a network [9]. The betweenness of an edge is "the number of the shortest paths between pairs of vertices that run along it". According to this definition, the edge(s) between groups own a larger betweenness. Figure 2 represents the concept of edge-betweenness clustering. If edges a, b, c, and d are removed, three groups are clearly isolated. To assure the quality of clustering, the index of modularity is defined as "the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random". The optimal structure of a network is the one with the maximum modularity. The procedure of edge-betweenness clustering is as follows: First, calculate the betweenness for all edges in a citation network; Second, remove the edge(s) with the highest betweenness; Third, recalculate the betweenness for all edges affected by the removal; Fourth, repeat step 2 and 3 until no edge remains. Thereupon trace back the above processes and select the division with the largest modularity as the result.

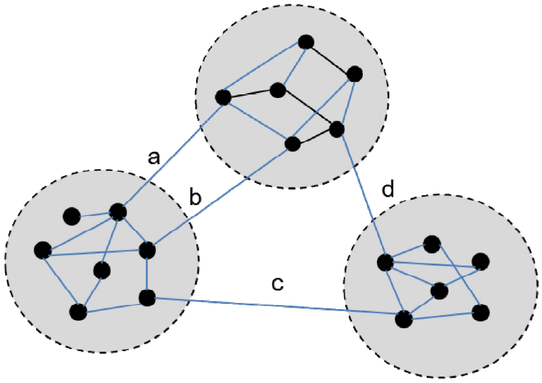


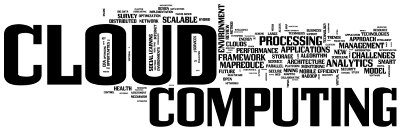
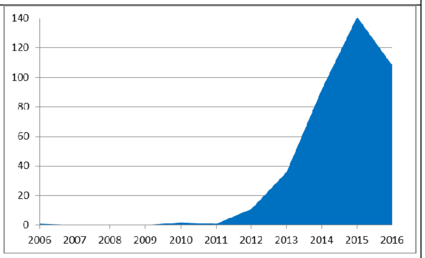
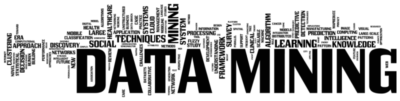
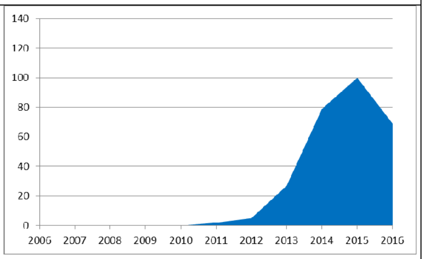
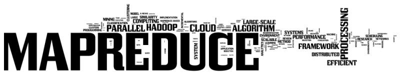
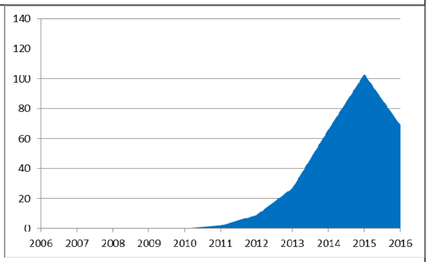
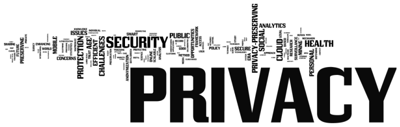
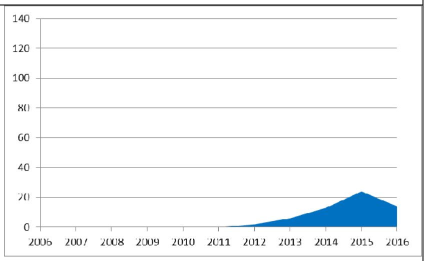
Fig. 2. The concept of edge-betweenness clustering

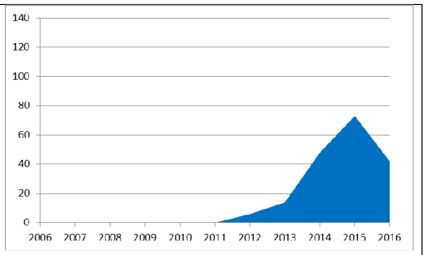
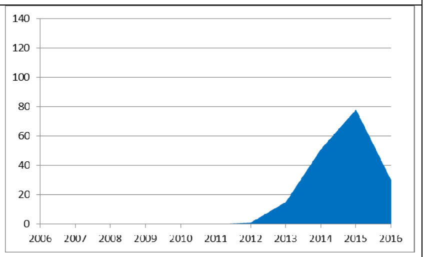
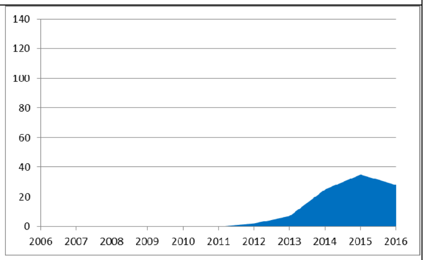
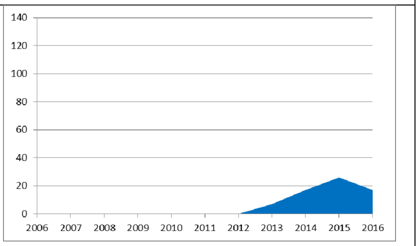
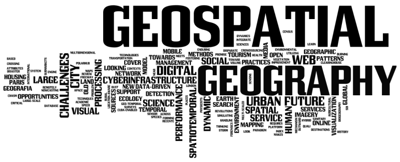
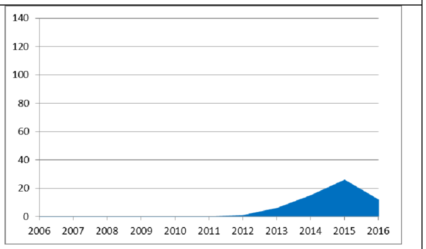
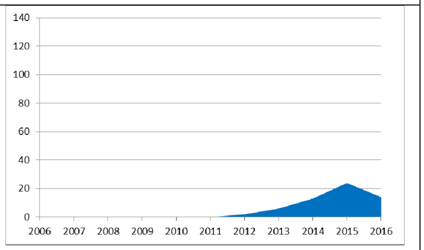
Many research studies have applied the edge-betweenness clustering to identify the major research themes or research fronts of a targeted scientific and technological field and demonstrated that this is an efficient and effective way [10-12].

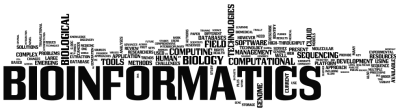
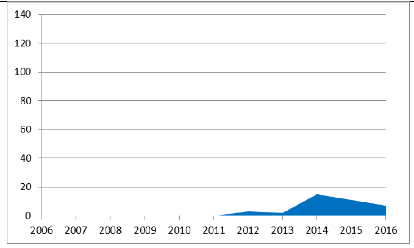
IV. Results

We apply the edge-betweenness clustering technique on the citation network to grouping the 3,898 papers. Six major research groups with more than 100 papers are identified. Five minor groups include 38 to 97 papers. Others are small-sized. We focus on these 11 major research groups to conduct further analysis. We apply a word clouding technique on the titles of all the papers in each group to identify the major keywords and name the theme of each group accordingly. Table 2 exhibits the theme, the number of papers, word clouding result, and growth curve of these 11 groups. We select the keyword(s) with the highest appearance in each group and name them as "cloud computing", "data mining", "MapReduce", "privacy", "social media", "healthcare", "ethics and legal", "surveillance", "geography", "business intelligence", and "bioinformatics". From the growth curve, one can find that researchers have started to investigate the themes of healthcare, surveillance, and geography since 2012, and hence, these three themes can be viewed as the emerging research topics in the big data field.

Table 2. Major attributes of eight research groups

Theme	No. of papers	Keywords via word clouding	Growth curve
Cloud computing	392	 CLOUD COMPUTING PROCESSING APPLICATIONS FRAMEWORK MAPREDUCE ANALYTICS CHALLENGES MANAGEMENT SCALABLE ENVIRONMENT PERFORMANCE ARCHITECTURE STORAGE HEALTH MODEL SURVEY DISTRIBUTED NETWORK SOCIAL LEARNING HADOOP SERVICES	
Data mining	282	 DATA MINING TECHNIQUES SOCIAL LEARNING KNOWLEDGE MINING HEALTHCARE SYSTEMS DISCOVERY APPROACH CLUSTERING PROCESSING ALGORITHM PREDICTION DETECTION INTELLIGENCE FRAMEWORK SURVEY	
MapReduce	276	 MAPREDUCE PARALLEL HADOOP CLOUD ALGORITHM PROCESSING LARGE-SCALE FRAMEWORK PERFORMANCE SYSTEMS DISTRIBUTED EFFICIENT COMPUTING SIMILARITY SYSTEM	
Privacy	222	 PRIVACY SECURITY PRIVACY-PRESERVING HEALTH SOCIAL CLOUD PROTECTION CHALLENGES EFFICIENT ISSUES PUBLIC ANALYTICS PERSONAL MOBILE CONCERNS SECURE AGE ANONYMIZATION	

Social media	183	SOCIAL MEDIA TWITTER ANALYTICS SENTIMENT ONLINE DIGITAL NETWORKS HEALTH MARKETING URBAN INFORMATION PUBLIC	
Healthcare	175	HEALTHCARE HEALTH CARE MEDICAL ANALYTICS CLINICAL PATIENT PATIENTS METHODS BASED SERVICES QUALITY FRAMEWORK KNOWLEDGE MANAGEMENT LARGE	
Ethics and legal	97	ETHICS ETHICAL LEGAL PRIVACY SOCIAL ISSUES ASPECTS MEDIA DILEMMAS HEALTH PUBLIC OPPORTUNITIES SECURITY SHARING ANALYTICS MEDICINE	
Surveillance	67	SURVEILLANCE PRIVACY PUBLIC VIDEO GLOBAL KNOWLEDGE ETHICS	
Geography	60	 GEOSPATIAL GEOGRAPHY DIGITAL CYBERINFRASTRUCTURE SOCIAL WEB URBAN FUTURE SPATIAL SERVICE CHALLENGES CITY PROCESSING LARGE SCIENCE SPATIOTEMPORAL	
Business intelligence	59	BUSINESS INTELLIGENCE ANALYTICS HEALTHCARE SYSTEMS MODEL FRAMEWORK INFORMATION DECISION	

Bioinformatics	38		

V. Conclusion

This paper presents a unique approach to identifying the major research themes within big data literature. We retrieved 3898 articles from Scopus and Web of Science and constructed the citation network among these articles and then applied edge-betweenness clustering to identify the major research themes in the big data literature. The major research themes reveal the active research topics and the growth curves hint at the directions for future research. The results provide a valuable reference for those who are interested in understanding the development of big data and are very helpful for filling the gap in the literature review.

The methodology adopted in this study is also applicable to other academic field and any technological field. One can retrieve relevant data of a target scientific or technological field, and then apply the edge-betweenness clustering method to identify the major research trends or technological trends.

References

[1] C. Doulkeridis and K. Nørvåg, "A survey of large-scale analytical query processing in MapReduce," VLDB Journal, vol. 23, pp. 355-380, 2014

[2] F.F. Costa, "Big data in biomedicine," Drug Discovery Today, vol. 19, pp. 433-440, 2014.

[3] I. Polato, R. Re´, A. Goldman and F. Kon, "A comprehensive view of Hadoop research - A systematic literature review," Journal of Network and Computer Applications, vol. 46, pp. 1-25, 2014.

[4] J. Fan, F. Han and H. Liu, "Challenges of Big Data analysis," National Science Review, vol. 1, pp. 293-314.

[5] X.-W. Chen and X. Lin, "Big data deep learning: Challenges and perspectives," IEEE Access, vol. 2, pp. 514-525.

[6] M. Girvan and M.E. Newman, "Community structure in social and biological networks," Proceedings of the National Academy of Sciences of the United States of America, vol. 99, pp. 7821-7826, 2002.

[7] M.E.J. Newman, "Fast algorithm for detecting community structure in networks," Physical Review, E 69, 066133, 2004.

[8] M.E.J. Newman, "Modularity and community structure in networks," Proceedings of the National Academy of Sciences of the United States of America, vol. 103, pp. 8577-8582, 2006.

[9] M.E.J. Newman and M. Girvan, "Finding and evaluating community structure in networks," Physical Review, E 69, 026113, 2004.

[10] L.Y.Y. Lu and J.S. Liu, "Development trajectory and research themes of foresight," Technological Forecasting and Social Change, vol. 112, pp. 347-356, 2016.

[11] L.Y.Y. Lu and J.S. Liu, "A novel approach to identify the major research themes and development trajectory: The case of patenting research," Technological Forecasting and Social Change, vol. 103, pp. 71-82, 2016.

[12] J.S. Liu, L.Y.Y. Lu and W-M. Lu, "Research Fronts in Data Envelopment Analysis," Omega - The International Journal of Management Science, vol. 58, no. 1, pp. 33-45, 2016.