ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/354715787

Research Trends in Social Media/Big Data with the Emphasis on Data Collection and Data Management: A Bibliometric Analysis

Chapter · September 2021

DOI: 10.1007/978-3-030-83010-6_4

CITATION
1

READS
31

2 authors:

Qiong Peng
University of Maryland, College Park
9 PUBLICATIONS **15** CITATIONS
SEE PROFILE

Xinyue ye
Texas A&M University
329 PUBLICATIONS **5,329** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Spatial-Social Networks View project

Monitoring the Spatial Spread of COVID-19 and Effectiveness of the Control Measures using Big Movement Data View project

All content following this page was uploaded by Xinyue ye on 23 September 2021.

The user has requested enhancement of the downloaded file.

Research Trends in Social Media/Big Data with the Emphasis on Data Collection and Data Management: A Bibliometric Analysis

Qiong Peng and Xinyue Ye

Abstract: Data collection and data management research is a multidisciplinary field that covers a wide range of subjects, each of which serves as a fundamental prerequisite for Human Dynamics Research utilizing Social Media/Big Data. The ever-changing academic landscape of this field has been characterized by rapid expansion of various applications and dynamic collaboration across multiple disciplines, yielding an increasing number of publications. This chapter reviews the research trend of such techniques and methodologies over the past ten years (2010-2019). Specifically, we conducted Bibliometric analysis to examine growth of output during 2010-2019, distribution of output in subject categories and journals, most cited documents, geographic and institutional distribution of publications, institution collaboration network, and keywords. The keyword analysis reveals that "big data", "social media", "data collection", "Twitter", "Facebook", and "privacy", were popular throughout the past 10 years. Additional keywords such as "data management", "cloud computing", "machine learning", "data mining", "Internet of Things", "big data analytics", "crowdsourcing", "data analytics", "data science", "big data management", "Hadoop", "MapReduce", "sentiment analysis", "surveillance", "business intelligence", and "IoT" have attracted increasing attention, further reflecting research trends.

4.1 Introduction

Advances in social sensing and data acquisition technologies have led to an enormous amount of human dynamics data. Social media data, in particular, have been used extensively in human dynamics research [1]. In many ways, analysis of social media data could prove especially helpful in detecting anomalous events such as panic resulting from natural disasters [2, 3], a pandemic [4], and measuring interaction among cities and regionals [5, 6]. The social media data could also be used to observe shopping behavior [7], travel recommendations [8], political activism [9],

Q. Peng.
School of Architecture, Planning and Preservation, University of Maryland
College Park, MD
X. Ye. (✉) (corresponding author)
Department of Landscape Architecture and Urban Planning,
Texas A&M University,
College Station, TX
Email: xinyue.ye@tamu.edu

and business intelligence [10]. However, there are potential drawbacks from relying on these data, such as unrepresentative sample, rumors, and location spoofing [11].

Focusing on the technical aspects of data collection and data management, social media data including text messages, photos, and videos posted on social media platforms can be collected and managed using several techniques. Furthermore, social media may be associated with location information, such as through a check-in function or geo-tagging ; this information could offer a solid foundation for geographical analysis Still, there are some strategic decisions that need to be made before data collection methods are established, such as strategic decisions about the period of data collection and search criteria for collecting data (i.e. based on lists of user accounts or filtered by topics and corresponding hashtags) [12]. To access social media data, researchers can fetch the data using Python scrapy and coding-free web tools. Even more, there are free research web tools that enable data collection and visualization, such as the Social Data Analytics Tool (SODATO) and Netlytic [13]. Developed by the Copenhagen Business School, SODATO can fetch data from Facebook and Twitter. Those interested in utilizing this tool may contact the research group, but public access to SODATO is currently limited. Netlytic can capture data from YouTube and Twitter, and the data are geo-coded.

Efficient data management is also critical to handling large quantities of social media data. There are increasing numbers of established data management techniques and techniques vary across data structure. For example, Zheng et al. [14] conclude that there are three common data structures- stream, trajectory, and graph data. Regarding various data structure, Zheng et al. introduce corresponding data management techniques, including data reduction techniques for trajectories[1], noise filtering techniques for trajectories, techniques for indexing and query trajectories, managing spatiotemporal graphs techniques, hybrid indexing structures techniques that can well organize different data sources [14].

Social data collection and data management research comprise a multidisciplinary field that covers a wide range of subjects including information science, computer science, and geography. It is necessary to identify the cutting-edge trends related to this research. Although there is increasing interest in big data and social media data-particularly data collection and data management-there is still limited research on the “big picture” of social media data collection and data management. In addition, human dynamics research spans multiple disciplines of study. Hence, it is necessary to examine global development and research trends comprehensively when discussing social media data collection and data management. A systematic review of social media data collection and management techniques could help readers gain a better understanding of research achievements, directions, and development of research methods.

Bibliometric analysis incorporates both visual and quantitative analytics that can be used to summarize trends in selected research fields [15, 16]. This type

[1] A spatial trajectory is a trace indicating a moving vehicle or induvial in geographical spaces.

of analysis can reveal temporal dynamics of scholarly outputs, spatial and institutional distributions of publications, scientific collaborations, and major research trends [17]. Furthermore, bibliometric network analysis, such as co-word analysis [18], co-citation analysis [19], co-authorship analysis [20], and co-publication analysis [21]), can be conducted to shed light on relationships between keywords, as well as other identifiers such as country, research institute, and author.

In this study, we used a bibliometric method to examine global research trends of big data and social media data research in the last decade with an emphasis on data collection and data management. The purposes of this study are to (1) evaluate research performance by country, institute, journal, subject category, and keywords; and (2) briefly identify future research directions in big data and social media data research with an emphasis on data collection and data management.

4.2 Methodology, Data Collection, and Analysis

4.2.1 Applications

Bibliometric analysis was carried out in order to evaluate the characteristics and trends in big data and social media data research. Bibliometric analysis was introduced by Pritchard [16] as a mathematical and statistical approach to analyze pertinent literature and understand the global research trends in a specific area. It has been applied to environmental engineering and science, soil science, ecology, food safety, new energy utilization, and other aspects. Bibliometric indicators analyzed in this study include publication number, subject categories, source journals, countries, institutions, journals, and keywords.

4.2.1 Data Collection

The dataset was derived from the database of the Science Citation Index (SCI) and Social Science Citation Index (SSCI) publications by Web of Science. The following keywords were used, including Topic= (“social media*” OR “big data*”) AND (“data collection*” OR “data management*”) to search all archived documents. The publications that contain any of those keywords and their variants (with *) in their titles, abstracts and keyword lists were included. The following information was downloaded: title, authors, institutions, abstract, keywords and cited references. The studies period spanned 2010 to 2019. Our bibliographic search resulted in 1,436 records.

4.2.3 Analysis Tools

In order to conduct bibliometric analysis, an R package 'Bilbiometrix' [22], VOSviewer [23], Ucinet [24], or other packages can be applied (see [25]) for a review on bibliometric software). In this study, we use the R package "Bibliometrix" and VOSviewer. VOSviewer is a freely available texting software for generating bibliometric maps and analyzing trends in scientific literature. Natural language processing techniques are built in the VOSviewer package to enable one to create term co-occurrence networks based on English-language textual data. State-of-the-art techniques for network layout and network clustering are available in the software. VOSviewer software uses a circle and label to represent an element, in which the circle size represents importance, and circles with the same color belong to the same cluster. In the following section, we will focus on aspects that describe global scientific production for big data and social media data research:

(1) Growth of output during 2010-2019
(2) Distribution of output in subject categories and journals
(3) Most cited documents
(4) Geographic and institutional distribution of publications
(5) Institution collaboration network analysis
(6) Keywords analysis

4.3 Results and Discussion

4.3.1 Characteristics of Article Outputs

1,436 publications have been identified as being data collection and data management of social media and big data-related during 2010-2019. A total of 1,436 documents includes 1,233 articles, 154 reviews, 10 book chapters, 28 early accesses, 30 proceeding papers, two letters, 39 editorial materials, one software review, and one book review. After removing records that do not have completed authorship and publication year information, we were left with 1,408 records. The characteristics of the article outputs are shown in Figure 4.1 and Table 4.1. The number of annual publications rose from four in 2010 to 362 in 2019, illustrating a dramatic rise and upward growth of this are of the research in the past decade. The average annual growth rate of all SCI and SSCI publications in the field was 56.74%. In addition, Figure 4.1 shows that the annual growth rate of publication has accelerated since 2014.

The average number of authors and references increased from three and 50 in 2010 to 4.439 and 59.948 in 2019, respectively (Table 4.1). The growth of an average number of authors per article indicates that collaboration in the field has

increased. The growth of average references per article indicates that there is increased knowledge about this topic. It is interesting to see that the average number of citations per article has decreased overall since 2010.

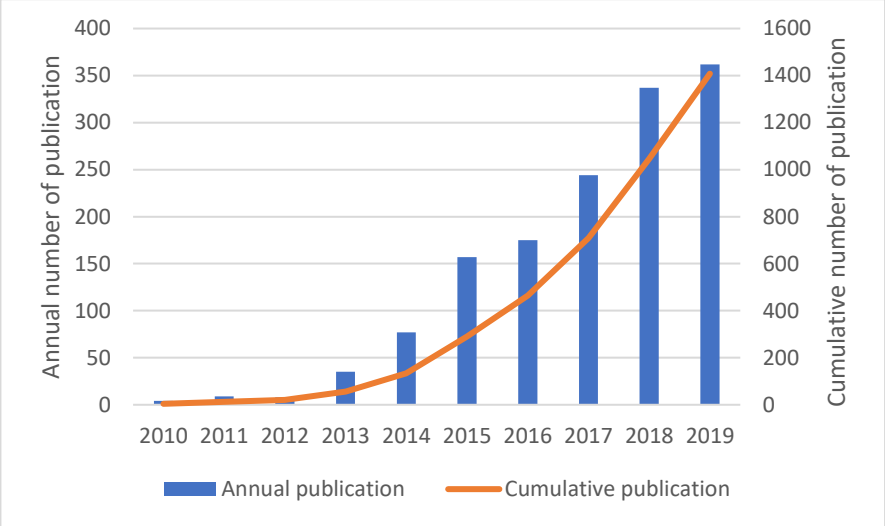

Figure 0.1. Growth of publication outputs. (Horizontal axis: year; Vertical axis: number of publications)

Table 0.1. Scientific outputs descriptors during 2010-2019

PY	TP	AU	AU/TP	NR	NR/TP	TC	TC/TP
2010	4	12	3.000	200	50.000	453	113.250
2011	9	35	3.889	301	33.444	555	61.667
2012	8	25	3.125	329	41.125	792	99.000
2013	35	102	2.914	1476	42.171	1793	51.229
2014	77	326	4.234	3128	40.623	3123	40.558
2015	157	600	3.822	7616	48.510	3123	19.892
2016	175	813	4.646	9053	51.731	3119	17.823
2017	244	1017	4.168	13739	56.307	2626	10.762
2018	337	1622	4.813	19529	57.950	1769	5.249
2019	362	1607	4.439	21701	59.948	442	1.221

PY: year; TP: number of publications; AU: number of authors; TC: total citation count; NR: number of cited references; AU/TP, NR/TP, and TC/TP: an average of authors, references, and citation per paper.

4.3.2 Subject Categories and Major Journals

Based on the classification of the Web of Science categories, the sample documents covered 194 subject categories. The research domain covered a wide variety of themes and disciplines. The top 20 subject categories were presented in Table 4.2. The results show that data collection and management for big data and social media research spanned a wider range of disciplines, but such studies mainly stemmed from computer science information systems, engineering electrical electronics, telecommunications, computer science theory methods, computer science interdisciplinary applications, and information science library science.

Table 0.2. Distribution of the top 20 subject categories.

Subject Category	**TP(%)**	
Computer science information systems	233	16.19
Engineering electrical electronic	158	10.98
Telecommunications	126	8.76
Computer science theory methods	108	7.51
Computer science interdisciplinary applications	90	6.25
Information science library science	90	6.25
Computer science software engineering	76	5.28
Health care sciences services	68	4.73
Medical informatics	60	4.17
Environmental sciences	51	3.54
Public environmental occupational health	51	3.54
Computer science artificial intelligence	50	3.47
Communication	49	3.41
Management	49	3.41
Computer science hardware architecture	44	3.06
Business	43	2.99
Medicine general internal	40	2.78
Multidisciplinary sciences	34	2.36
Environmental studies	33	2.29
Green sustainable science technology	29	2.02
Engineering civil	28	1.95
Operations research management science	28	1.95

The top 20 active journals are summarized in Table 4.3. This table shows that *IEEE (The Institute of Electrical and Electronics Engineers) Access* is the most productive journal, followed by *Future Generation Computer Systems - The International*

Journal of eScience, *Journal of Medical Internet Research*, *ISPRS (International Society for Photogrammetry and Remote Sensing) International Journal of Geo-information*, *Sensors*, and *Sustainability*. Regarding average citation number per article, *IEEE Transactions on Knowledge and Data Engineering*, *Computers in Human Behavior*, and *Big Data* are the three most highly cited journals, with a magnitude of 112.4, 70, and 44.429, respectively.

Table 0.3. The 20 most active journals

Journals	TP	(%)	TC	(%)	TC/TP
IEEE Access	43	3.054	234	1.315	5.442
Future Generation Computer Systems - The International Journal of eScience	29	2.060	399	2.242	13.759
Journal of Medical Internet Research	29	2.060	384	2.158	13.241
ISPRS International Journal of Geo-information	14	0.994	46	0.258	3.286
Sensors	14	0.994	124	0.697	8.857
Sustainability	14	0.994	51	0.287	3.643
BMJ Open	13	0.923	42	0.236	3.231
PLOS ONE	12	0.852	235	1.321	19.583
Cluster Computing - The Journal of Networks Software Tools and Applications	10	0.710	24	0.135	2.400
Concurrency and Computation: Practice and Experience	10	0.710	63	0.354	6.300
IEEE Transactions on Knowledge and Data Engineering	10	0.710	1124	6.316	112.400
International Journal of Information Management	10	0.710	118	0.663	11.800
Transportation Research Record	10	0.710	39	0.219	3.900
IEEE Internet of Things Journal	9	0.639	222	1.248	24.667
Computers in Human Behavior	8	0.568	560	3.147	70.000
IEEE Transactions on Industrial Informatics	8	0.568	274	1.540	34.250
SIGMOD Record	8	0.568	56	0.315	7.000
Big Data	7	0.497	311	1.748	44.429
Computer Law & Security Review	7	0.497	54	0.303	7.714
Frontiers in Psychology	7	0.497	17	0.096	2.429
International Journal of Distributed Sensor Networks	7	0.497	20	0.112	2.857
Online Information Review	7	0.497	38	0.214	5.429

TP: number of publications; TC: total citation count; TC/TP: an average of citations per paper

4.3.3 Most Cited Documents

Based on the number of citations, we list the 10 most cited documents relating to data collection and data management in Table 4.4. The citation number has updated base on Google Scholar figures. If audiences wish to heavily research big data collection and data management, it might prove worthwhile to read those highly cited documents first to gain an overview of the topic. For example, one could read Chen & Zhang [26] to learn more about challenges, techniques, and technologies as they relate to big data. In order to learn how to process and generate large datasets, it could prove beneficial to learn more about a programming model- MapReduce- by reading Dean & Ghemawat [27].

Table 0.4. The 10 most cited documents

Title	Year	Publication	Author	Citation Count[2]
Users of the world, unite! The challenges and opportunities of Social Media	2010	*Business horizons*	AM Kaplan, M Haenlein	17,762
MapReduce: Simplified data processing on large clusters	2004	*usenix.org*	J Dean, S Ghemawat	12,758
Big data: A revolution that will transform how we live, work, and think	2013	*Houghton Mifflin Harcourt*	V Mayer-Schönberger, K Cukier	5,340
Big data: The next frontier for innovation, competition, and productivity	2011	*McKinsey*	J Manyika	5,319
Business intelligence and analytics: From big data to big impact.	2012	*MIS quarterly*	H Chen, RHL Chiang, VC Storey	4,620
Big data: the management revolution	2012	*Harvard Business Review*	A McAfee, E Brynjolfsson, TH Davenport	4,130
Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon	2012	*Information, communication & society*	D Boyd, K Crawford	3,254
Big data: A survey	2014	*Mobile networks and applications*	M Chen, S Mao, Y Liu	2,403
Data-intensive applications, challenges, techniques and technologies: A survey on Big Data	2014	*Information sciences*	CLP Chen, CY Zhang	2,255
Beyond the hype: Big data concepts, methods, and analytics	2015	*International journal of information management*	A Gandomi, M Haider	2,212

[2]The citation count is updated based on Google Scholar on January 9, 2020.

4.3.4 Geographic and Institutional Distribution of Publications

The geographic and institutional distributions of publications were generated based on author affiliation information. We summarized the 10 most productive countries in Figure 4.2, in terms of the number of total publications, single country articles, and international collaborations, respectively. Out of these 10 countries, five were from Europe, two from North America, one from Australia, and two from Asia. The U.S. was the most productive country with a total of 506 articles. China ranked second with 234 articles, followed by the UK with 195 articles. Figure 4.2 also reveals that some countries achieved higher rates of international collaboration. These countries include France (collaboration rate: 65.00%), Spain (collaboration rate: 63.16%), Canada (collaboration rate: 62.65%), Australia (collaboration rate: 60.31%), and Germany (collaboration rate: 60.32%).

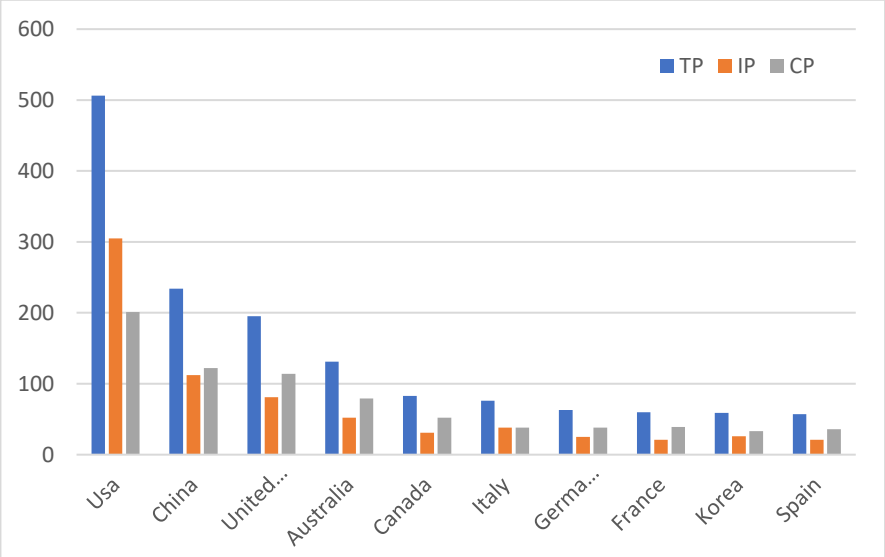

Figure 0.2. Most productive countries during 2010-2019 (TP, total publications; IP, the number of independent publications by country; CP, the number of internationally collaborative publications)

Co-authorship analysis enabled the study of the most influential countries' cooperation network, as plotted in Figure 4.3. The size of nodes reveals the productivity rate of each country, while the thickness of curved lines between countries demonstrates the strength of collaboration. The U.S., China, and the U.K. had the largest number of papers with co-authorships. As we can see from Figure 4.3, there are clusters that demonstrate inter-country collaboration. European Countries cluster together, whereas Asian countries and Australia cluster together. This suggests that countries from the same continent tend to collaborate more with one another than

with countries from distant continents. Note that the U.S. is at the center of the collaboration network, which illustrates that U.S. researchers conduct an enormous amount of collaborative research with researchers from both Asian and European countries.

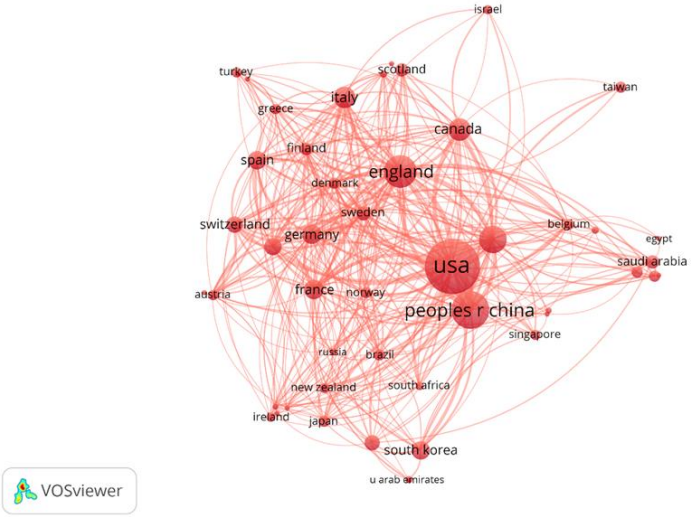

Figure 0.3. Co-authorship cooperation between productive countries

4.3.5 Institution Collaboration Network

A collaboration network of the 80 most productive institutions is visualized using the VOSviewer (Figure 4.4). The most productive institution proved to be the University of Michigan with 21 papers, followed by both the Chinese Academy of Science and the University of Sydney, each of which produced 17 papers (Table 4.5). Each node in Figure 4.4 indicates an institution. The size of each node indicates the institution's productivity. The bigger the node is, the more productive the organization is. The distance between two organizations in the visualization approximately indicates the relatedness of the organizations in terms of co-authorship links. The closer two organizations are located to each other, the stronger their relatedness. The co-authorship links between organizations are also represented by curve lines. The institutions are clustered into two groups: each group has a unique color. Asian institutions are represented by the green group, while North America institutions, European institutions are represented by the red. Institutions that are working on

social media data collection and management within the same continent are clustered closer together and have more connections than institutions scattered across different continents. This means that institutions from the same continents are more likely to collaborate with one another than institutions from different continents.

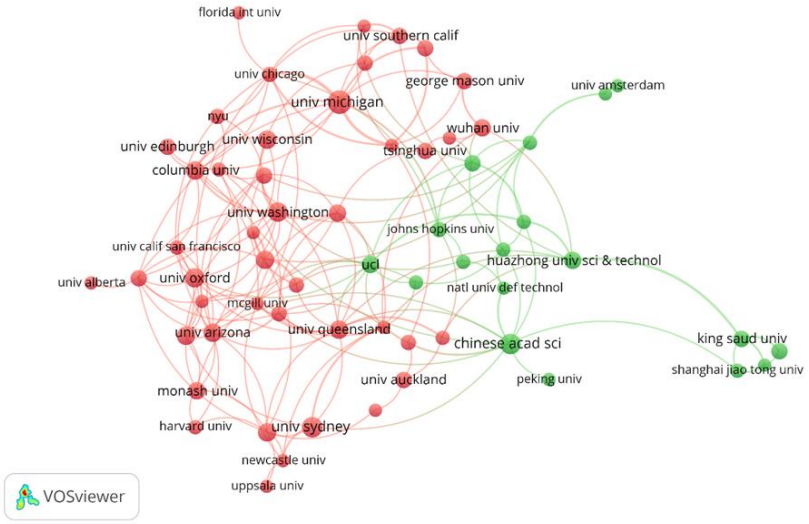

Figure 0.4. Institution collaboration network of most 66 central institutions

Table 0.5. Top 17 institutions based on the total publications

Rank	Organization	Country	Number of publications	Total citations
1	University of Michigan	USA	21	226
2	Chinese Academy of Sciences	China	17	370
3	University of Sydney	Australia	17	266
4	University of Washington	USA	15	350
5	University of Oxford	UK	15	177
6	University of Arizona	USA	14	230
7	University of Cambridge	UK	14	312
8	Columbia University	USA	14	178
9	University of Melbourne	Australia	14	502
10	University of Toronto	Canada	14	84
11	University of Queensland	Australia	13	105
12	University of Wisconsin	USA	13	243

13	University of College London	UK	12	443
14	Huazhong University Science & Technology	China	12	54
15	University of Maryland	USA	12	379
16	Monash University	Australia	12	56
17	Wuhan University	China	12	72

4.3.6 Keywords Analysis – Network Analysis

Keywords supplied by the authors provide a very basic idea of the topics covered within the article. The 30 most frequently used keywords in the study period are calculated and ranked in Column (1) of Table 4.6. Co-occurrence links indicate the frequency whereby keywords occur simultaneously in a study. The co-occurrence relationships between keywords can be shown by the co-occurrence links in the co-occurrence word network. In this study, the co-occurrence relationships between the top 86 high-frequency keywords were examined, and the co-occurrence word networks were visualized by VOSviewer software (Figure 4.5). The nodes represent high-frequency keywords, and the size of each represents the degree of frequency the keyword is used. The high-frequency keywords are selected based on how many times keywords have. A higher value for a keyword represents a higher frequency at which the word was referenced as a keyword in the last 10 years. The distance between two keywords in the visualization approximately indicates the relatedness of the keywords in terms of co-occurrence links. The closer two keywords are located to each other, the stronger their relatedness. The strongest co-occurrence links between keywords are also represented by curved lines.

As shown in Figure 4.5, the 86 most frequent author keywords are grouped into four clusters. The red cluster mainly focuses on social media; the green and yellow clusters focus mainly on big data dimensions, and the blue cluster is mainly about data analytics. The keywords that are referenced with highest frequency are “big data”, “social media”, “data management”, “data collection”, “cloud computing”, “Twitter”, “machine learning”, “Facebook”, and “data mining”. Twitter and Facebook are the two most popular social data platforms referenced in human dynamics research. It is not surprising to see these two words occurring in high frequency. We also see that machine learning is frequently referenced, which coincides with the fact that the machine learning approach has been popular and been applied heavily in social media data analysis and human dynamics. Other significant topics over the past 10 years include: “Internet of things”, “privacy”, “big data analytics”, “crowdsourcing”, “data analytics”, “Internet”, “ethics”, “data Science”, “big data management”, and “citizen science”. The high frequency of the aforementioned topics suggests that these are critical areas of social media data collection and data management throughout human dynamics research. For example, the keyword “ethics” and “privacy” both occur frequently in the literature, which reflects that

privacy and ethics are among the critical concerns relating to social media data collection and management.

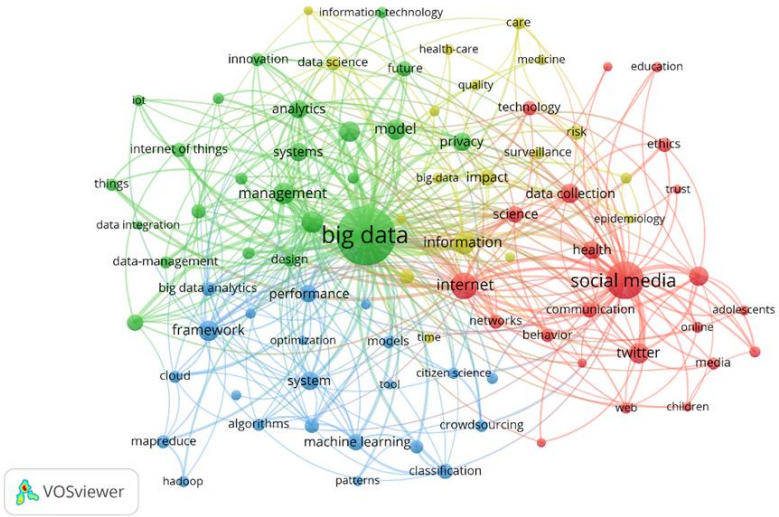

Figure 0.5. Co-occurrence network of top 86 high-frequency keywords.

4.3.7 Keywords Analysis – Temporal Evolution

By examining the temporal evolution of these keywords would give us insights on trending areas of research. To closely examine the temporal evaluation, we divide the 10 studied years into three consecutive periods (2010-2012, 2013-2015, and 2016-2019). The 30 most frequently used keywords throughout the entire studies period (2010-2019) are calculated and ranked in Columns (2), (3), and (4) of Table 4.6. We identify keyword trends on the basis of whether or not the rank of a keyword rises upward across the three consecutive periods. Eleven keywords ("data management", "cloud computing", "machine learning", "data mining", "Internet of things", "big data analytics", "crowdsourcing", "data analytics", "data Science", "big data management", "Hadoop", "MapReduce", "sentiment analysis", "surveillance", "business intelligence", and "IoT") are becoming increasingly popular over the past 10 years. For example, the keyword "machine learning" does not occur in articles in the 2010-2012 period, but usage of this keyword increased from 18th in 2013 to 2015 to fifth from 2016 to 2019, suggesting that machine learning rose as a trending research topic in the past years. This dramatic increase coincided with the popularity of AI and machine learning. It is not surprising that the keyword "cloud computing" has also been frequently referenced frequently in recent years. This can be attributed

to the recent advance of cloud computing technology and applications in social data analysis. In addition, “crowdsourcing” and “surveillance” rose in popularity. Human crowdsourcing and surveillance cameras are the two important approaches for sensing and data acquisition. Surveillance cameras generate a huge volume of images and videos while human crowdsourcing generates data via mobile devices. These data acquisition approaches provide information for traffic analysis, human mobility, and urban structure research needs. Another example for treading research topics is the keyword “sentiment analysis”. With social media data, researchers are capable of analyzing the change of human sentiments and their mobility pattern as a result of an event (e.g., disaster, pandemic, or presidential candidate campaign).

Table 0.6. Temporal evolution of the 30 most frequently used keywords

Keywords	Periods									Rising Trend
	2010-2019			2010-2012		2013-2015		2016-2019		
	N	R	(%)	N	R	N	R	N	R	
Big data	383	1	5.75	---	---	69	1	314	1	
Social media	155	2	2.33	7	1	42	2	106	2	
Data management	61	3	0.92	1	22	9	7	51	3	X
Data collection	58	4	0.87	1	19	14	3	43	4	
Cloud computing	43	5	0.65	---	---	8	9	35	6	X
Twitter	42	6	0.63	2	3	11	5	29	8	
Machine learning	40	7	0.60	---	---	4	18	36	5	X
Facebook	35	8	0.53	2	2	13	4	20	13	
Data mining	34	9	0.51	---	---	7	10	27	9	X
Internet of things	34	10	0.51	---	---	3	32	31	7	X
Privacy	32	11	0.48	---	---	9	8	23	11	
Big data analytics	28	12	0.42	---	---	4	15	24	10	X
Crowdsourcing	24	13	0.36	---	---	2	45	22	12	X
Data analytics	21	14	0.32	---	---	3	26	18	14	X
Data analysis	20	15	0.30	---	---	5	12	15	16	
Internet	20	16	0.30	---	---	10	6	10	29	
Ethics	19	17	0.29	---	---	4	16	15	18	
Data science	17	18	0.26	---	---	2	53	15	17	X
Big data management	16	19	0.24	---	---	1	162	15	15	X
Citizen science	16	20	0.24	1	10	3	22	12	20	
Hadoop	15	21	0.23	---	---	1	420	14	19	X
Data	14	22	0.21	---	---	6	11	8	36	
MapReduce	13	23	0.20	---	---	2	67	11	24	X

Sentiment analysis	13	24	0.20	---	---	1	737	12	21	X
Surveillance	13	25	0.20	---	---	2	95	11	26	X
Business intelligence	12	26	0.18	---	---	2	39	10	28	X
IoT	12	27	0.18	---	---	1	491	11	23	X
Recruitment	12	28	0.18	1	70	3	34	8	40	
Security	12	29	0.18	---	---	3	35	9	34	
Technology	12	30	0.18	---	---	4	20	8	41	

N: the number of articles in the study period; R: the absolute rank of author keywords; ---: no such author keyword in a specific period.

4.4 Conclusions

In conclusion, this study social data collection and data management as they are referenced in human dynamics research through a bibliometric approach. We presented an overview and brief picture of existing studies in this area. Audiences looking to understand social media data collection and management as it is referenced in human dynamics can use this study to help determine which articles they should read, which journals they should consider for publication submissions, and which research trends are most significant.

In summary, the number of annual publications about big data and social media research with an emphasis on data collection and data management increased from just four in 2010 to 362 in 2019. The annual growth rate for such publications has accelerated since 2014. The studies covered wide variety of subjects, such as computer science information systems, engineering electrical electronic, telecommunications, computer science theory methods, computer science interdisciplinary applications, and information science library science. The three most productive journals in these areas were *IEEE Access*, *Future Generation Computer Systems - The International Journal of eScience*, and *Journal of Medical Internet Research*.

This study suggests that “big data”, “social media”, “data collection”, “Twitter”, “Facebook”, and “privacy” have been the most popular topics in this area over the past 10 years. Some keywords, such as “data management”, “cloud computing”, “machine learning”, “data mining”, “Internet of things”, “big data analytics”, “crowdsourcing”, “data analytics”, “data Science”, “big data management”, “Hadoop”, “MapReduce”, “sentiment analysis”, “surveillance”, “business intelligence”, and “IoT”, attracted increasing attention, reflecting research trends. Furthermore, most of those 30 most frequently referenced keywords in this area were not listed as keywords during the 2010-2012 period; rather, they grew in popularity in the most recent periods.

References

1. Shaw S-L, Tsou M-H, Ye X (2016) Editorial: human dynamics in the mobile and big data era. Int J Geogr Inf Sci 30:1687–1693
2. Wang Z, Lam NSN, Obradovich N, Ye X (2019) Are vulnerable communities digitally left behind in social responses to natural disasters? An evidence from Hurricane Sandy with Twitter data. Appl Geogr 108:1–8
3. Wang Z, Ye X (2018) Social media analytics for natural disaster management. Int J Geogr Inf Sci 32:49–72
4. Depoux A, Martin S, Karafillakis E, Preet R, Wilder-Smith A, Larson H (2020) The pandemic of social media panic travels faster than the COVID-19 outbreak. J Travel Med. https://doi.org/10.1093/jtm/taaa031
5. Ye X, Li S, Peng Q (2021) Measuring interaction among cities in China: A geographical awareness approach with social media data. Cities 109:103041
6. Gong J, Li S, Ye X, Peng Q (2020) Measuring the Dynamic Impact of High-Speed Railways on Urban Interactions in China. ArXiv201008182 Cs
7. Ye X, She B, Li W, Kudva S, Benya S (2020) What and Where Are We Tweeting About Black Friday? In: Thakur RR, Dutt AK, Thakur SK, Pomeroy GM (eds) Urban Reg. Plan. Dev. 20th Century Forms 21st Century Transform. Springer International Publishing, Cham, pp 173–186
8. Bao J, Zheng Y, Mokbel MF (2012) Location-based and Preference-aware Recommendation Using Sparse Geo-social Networking Data. In: Proc. 20th Int. Conf. Adv. Geogr. Inf. Syst. ACM, New York, NY, USA, pp 199–208
9. Thorson K, Driscoll K, Ekdale B, Edgerly S, Thompson LG, Schrock A, Swartz L, Vraga EK, Wells C (2013) Youtube, Twitter and the Occupy Movement. Inf Commun Soc 16:421–451
10. Karamshuk D, Noulas A, Scellato S, Nicosia V, Mascolo C (2013) Geo-Spotting: Mining Online Location-based Services for Optimal Retail Store Placement. Proc 19th ACM SIGKDD Int Conf Knowl Discov Data Min 793–801
11. Ye X, Zhao B, Nguyen TH, Wang S (2020) Social Media and Social Awareness. In: Man. Digit. Earth. Springer, Singapore, pp 425–440
12. Mayr P, Weller K (2017) Think before you collect: Setting up a data collection approach for social media studies. SAGE Handb Soc Media Res Methods 679
13. Hussain A, Vatrapu R (2014) Social Data Analytics Tool: Design, Development, and Demonstrative Case Studies. In: 2014 IEEE 18th Int. Enterp. Distrib. Object Comput. Conf. Workshop Demonstr. pp 414–417
14. Zheng Y, Capra L, Wolfson O, Yang H (2014) Urban Computing: Concepts, Methodologies, and Applications. ACM Trans Intell Syst Technol 5:38:1-38:55
15. Garfield E (1970) Citation indexing for studying science. Nature 227:669–671
16. Pritchard A (1969) Statistical bibliography or bibliometrics. J Doc 25:348–349
17. Li Q, Wei W, Xiong N, Feng D, Ye X, Jiang Y (2017) Social Media Research, Human Behavior, and Sustainable Society. Sustainability 9:384

18. Ding Y, Chowdhury GG, Foo S (2001) Bibliometric cartography of information retrieval research by using co-word analysis. Inf Process Manag 37:817–842
19. He Y, Cheung Hui S (2002) Mining a Web Citation Database for author co-citation analysis. Inf Process Manag 38:491–508
20. Glänzel W, Schubert A (2005) Analysing Scientific Networks Through Co-Authorship. In: Moed HF, Glänzel W, Schmoch U (eds) Handb. Quant. Sci. Technol. Res. Use Publ. Pat. Stat. Stud. ST Syst. Springer Netherlands, Dordrecht, pp 257–276
21. Schmoch U, Schubert T (2007) Are international co-publications an indicator for quality of scientific research? Scientometrics 74:361–377
22. Aria M, Cuccurullo C (2017) bibliometrix: An R-tool for comprehensive science mapping analysis. J Informetr 11:959–975
23. Van Eck NJ, Waltman L (2010) Software survey: VOSviewer, a computer program for bibliometric mapping. scientometrics 84:523–538
24. Borgatti SP, Everett MG, Freeman LC (2002) Ucinet for Windows: Software for social network analysis. Harv. MA Anal. Technol. 6:
25. Cobo MJ, López-Herrera AG, Herrera-Viedma E, Herrera F (2011) Science mapping software tools: Review, analysis, and cooperative study among tools. J Am Soc Inf Sci Technol 62:1382–1402
26. Philip Chen CL, Zhang C-Y (2014) Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Inf Sci 275:314–347
27. Dean J, Ghemawat S (2008) MapReduce: simplified data processing on large clusters. Commun ACM 51:107–113

View publication stats