Research Article

Big Data Research in India: A Scientometric Review

Ravindranath Wodeyar[1,*], K R Mulla[2]

[1]Library and Information Centre, Proudhadevaraya Institute of Technology, Hosapete, Karnataka, INDIA.

[2]Dr. S G Balekundri Central Library and Information Centre, Visvesvaraya Technological University. Belagavi, Karnataka, INDIA.

ABSTRACT:
Big Data paradigm has grown rapidly in recent years, resulting in significant research from a tremendous multitude of perspectives. A literature review is an important methodology for gathering information about the ongoing activity of science. There are several methods including bibliographic or scientometric studies in reviewing the research literature. This paper studied with a scientometric analysis of research work on the emerging field 'Big Data' in India using a Python software called 'ScientoPy'.[1] A data set of 3080 papers on Big Data Research in India was collected which are indexed in two major scientific databases, i.e. Clarivate Web of Science and Scopus over a 10-year period (2012–2021). This aims to examine the main authors, country affiliation, most published authors, top research outputs, and emerging themes to provide insight into recent developments in the field.

Keywords: Big Data, Bibliometric, Scientometric, ScientoPy, Web of Science, Scopus.

INTRODUCTION

Big Data is a new paradigm that has grown momentum recently all around the world, offering intensive studies from a multidimensional perspective. "Big data" originally referred to managing, handling, and analyzing very large datasets and has been used to refer to this ever since the mid–1990s. In the era of the World Wide Web and Web 2.0 technologies, an incredible amount of structured and unstructured data is being generated continuously from various sources, including emails, social media platforms, blogs, online transactions, articles, and forums. In addition, different types of sensor data produced from different sources such as health care science, environmental organizations, metrological departments, business data, census data, company data, etc. in larger volume and enormous velocity are called Big data. Research on the emerging area of Big Data started during recent years and gained tremendous momentum in computational sciences and related disciplines.

LITERATURE REVIEW

The evolution, growth, and scientific collaboration of Indian publications in the field of big data were examined.[2] Data for the study were gathered from the Scopus database from 2001 to 2015.To demonstrate the current state, trends, and collaboration in big data research, bibliometric analysis, visualisation, and mapping software are employed.

Analyses global big data research included in the Scopus database from 2007 to 2016, Big data grew at an alarmingly rapid rate (135.2 percent), yet had a low average citation impact per piece (3.75).[3] In the last ten years, big data has accounted for a very small fraction of highly cited publications (0.86 percent). Big data is quickly growing as a key field due to its potential to impact business, governance, society, healthcare, industry, and many other industries.

Analyze Google Trends on 657 research papers to examine various variables in IT, communications, medical, health, business, and economics. A fresh source of big data was used to investigate the key nodes of the networks and to review the research directions of representative publications.[4]

Provides a comprehensive assessment of studies related to big data in social media. It covers 2000 highly cited articles from 2012 to 2019. According to thematic analysis, the subject nearly retained a major and well–developed research field.

According to the findings, research has risen at an exponential rate since 2014.[5]

Seeks to map the scientific work done in the field of big data by doing a rigorous examination of scholarly articles published between 2010 and 2016.[6]

Thematic analysis of big data articles reveals that it is a discipline with research interests ranging from medical to social sciences.

Analysed the association between influential tweeters and highly cited articles in the field of information sciences using altmetric.com twitter data The dataset contains over 10,000 tweets, with mentions, retweets, and followers utilised to construct a linked, undirected graph.[7] A feature vector holding the accumulated total of the rank scores of those significant individuals who tweet a certain item is used in this work's machine-learning model.

Exploring the growth of big data research is a great method to improve technology management and add value to R&D plans. From 2000 to 2015, this work employs a learning-enhanced bibliometric investigation to uncover interactions between big data and its basic parts.[8]

Studied The literatures on big data (bd) and data science (ds) are studied using bibliometric indicators, which aid in determining the trajectory of publications in these research areas.[9] An increase in big data publications as well as a gradual increase in data science publications. Surprisingly, a new publication course combining the big data and data science concepts arises.

Examine the trend of scientific papers in the field of 'big data and policy' research during the previous two decades.[10] Bibliometrics is used as a technique to highlight the dynamics of scientific debates that take place through articles published in international journals. According to the findings, the United States is the country of origin for the majority of productive author affiliations in publishing papers. Smart cities, open data, privacy, artificial intelligence, machine learning, and data science are among the most often debated issues.

MATERIALS AND METHODS

The scientometric technique is a widely recognized quantitative tool for identifying and measuring the publication growth in any subject. Scientometrics is a quantitative discipline in which a large number of studies are conducted on numerical analysis of many aspects of the literature on a particular topic. It uses characteristics of bibliographic data to do statistical analysis of published content. Scientometric studies have gotten a lot of attention in recent decades and are frequently used to evaluate scientists' research and the evolution of numerous science areas (Verma and Shukla, 2020).

A scientometrics literature review tool named 'ScientoPy' (Ruiz-Rosero et al., 123 C.E.) was utilised for the analysis and display of publications linked to Big Data. This ScientoPy script can read the Clarivate Web of Science and Scopus databases, find and remove duplicate documents, generate a graph of the top topics (keywords, authors, countries), visualise the history of items within a topic, identify trending topics based on average growth rate (AGR), and calculate the h-index of authors and countries

Data Set

For this scientometrics analysis, Clarivate Web of Science (WoS) and Scopus were examined between 1 January 2012 to 31 December 2021, conference papers, articles, reviews, and proceedings papers were examined. This analysis was conducted using the search string "Big Data" on 26th January 2022 within a topic searched in both WoS and Scopus databases.

There were a total of 3080 documents in the two databases, as shown in Table 1.

Remove Duplicate Samples

In order to avoid duplicate samples across the databases, WoS publications were kept while Scopus publications were eliminated, leaving 3080 documents.

Growth of Big Data Publications

Table 2 shows the growth of Big Data before eliminating the duplicates and Figure 1 shows after deleting the duplicates of publications.

Country collaboration Analysis

Here the authors and country collaboration were analysed. Figure 2 shows the top 10 countries associated with Big Data with the highest occurrence in the data set, sorted by year.

A list of the collaborative countries with the most associated publications was compiled in Table 3. United States of America was leading collaborated country with 144 publications (66 Articles, 56 Conference papers and 22 Reviews) followed by United Kingdom with 77 publications (48 Articles,

Table 1: Documents by type and eliminate duplicates.

Source	Conference Paper	Article	Review	Total	%	Duplicated Removed
WoS	0	34	5	39	1.2	0
Scopus	1514	1495	81,	3090	98.8	49
Statics after duplication removal filter						
WoS	0	34	5	39	1.3	
Scopus	1507	1457	77	3041	98.7	
Final data for the study			3080			

Table 2: Annual Growth of Publications.

DataBase	Total	AGR	ADY	h-index
Scopus	3090	20	437	67
WoS	38	0	8	14

Year	Scopus	WoS
2012	6	0
2013	23	0
2014	66	1
2015	206	1
2016	360	4
2017	347	5
2018	535	5
2019	662	6
2020	408	8
2021	428	8

AGR: Average Growth Rate; ADY: Average documents per year;
Wos: Web of Science

17 Conference papers and 12 Reviews) and China in 3rd highest collaborated country with 61 publications (41 Articles, 12 Conference papers and 8 Reviews).

Author Analysis

The Authors of the 3080 publications connected to Big Data are represented by 9238 authors. The top ten authors with the most published documents per year are shown in Table 4. Kumar A stood in top, with 44 publications and 338 citations, with an *h*-index of 10. Figure 3. Top 10 Authors with most documents published per year.

Sharma S is positioned second with 30 documents and 177 citations, of which 14 list him as the primary author. Kumar N with 12 *h*-index is the fourth in this list with 27 publications, and received highest citations with 706 among others in the list.

Figure 1: Growth of Publications after removing the duplicated documents.
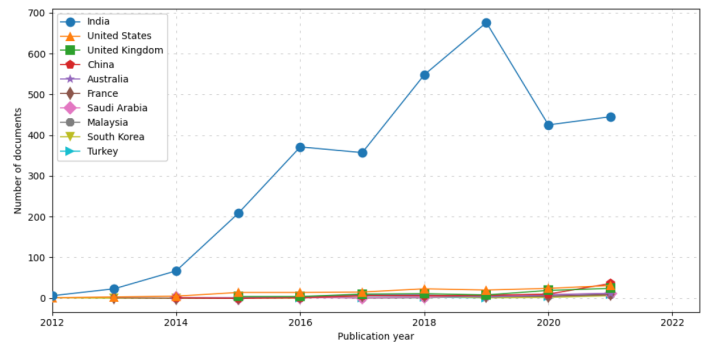
Figure 2: Top 10 Year wise Country Collaboration growth of Publications.

Table 3: Year wise publications with country collaboration.

| No.	Country	Total	AGR	h-index	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
1	India	3079	20	67	6	23	67	207	364	352	540	668	416	436
2	USA	144	5	36	1	3	5	13	14	15	22	18	24	29
3	United Kingdom	77	6	23	0	0	0	4	4	9	11	7	18	24
4	China	61	23	19	0	0	1	0	1	5	6	7	9	32
5	Australia	41	6	17	0	0	1	1	2	8	8	3	6	12
6	France	33	4	15	0	1	0	0	3	6	6	3	5	9
7	Saudi Arabia	32	1	6	0	0	1	0	1	1	3	7	9	10
8	Malaysia	21	4	6	0	0	0	0	2	0	1	6	4	8
9	South Korea	21	4	10	1	0	0	0	0	5	6	1	2	6
10	Turkey	21	6	10	0	0	0	0	1	2	3	1	4	10

AGR: Average Growth Rate
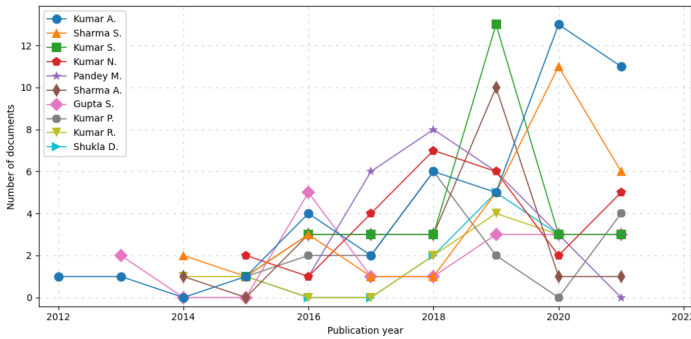
Figure 3: Top 10 Authors with most documents published per year.

Table 4: Top 10 authors with most publications, citations and *h*-index.

Author	Total	AGR	ADY	Citations	*h*-index
Kumar A.	44	-2	11	338	10
Sharma S.	30	-5	6	177	6
Kumar S.	29	0	3	207	6
Kumar N.	27	3	5	706	12
Pandey M.	24	-3	0	63	4
Sharma A.	22	0	1	187	6
Gupta S.	18	0	3	233	7
Kumar P.	17	4	4	86	5
Kumar R.	14	0	3	194	5
Shukla D.	14	0	3	21	2

AGR: Average Growth Rate; ADY: Average documents per year;
Wos: Web of Science

Table 5 displays the top five cited papers for each of the three document types viz. articles, conference proceedings, and reviews.

Wamba S.F. *et al*. obtained the most citations (614) for their article Big data analytics and firm performance published in the year 2017. For conferences and proceedings, Katal A.; Wazid M.; Goudar R.H. received the 507 citations for their study Big data: Issues, challenges, tools and Good practices.

Finally, on the reviews side, Mehta N.; and Pandit A obtained the citations of 138 for their review Concurrence of big data analytics and healthcare: A systematic review.

Research Topics

This section examines author keywords to identify trends in the various Big Data research fields. The majority of the writers in this data set indicate their research topic in the document keywords. Figure 4 depicts the trend of study subjects in several domains based on the most often used keywords in publications by top authors.

Figure 4 Trend of the research topics based on author key words per year

Next to Big data Data Analytics (487) Data mining (434) Hadoop are the top three trending topics in Big Data Research.

A word cloud can be used to visually represent keywords, highlighting those that occur frequently. Figure 5 shows the 10 most frequent author's keywords in the Big Data field.

Table 5: Most cited papers for three document types:Articles, Proceedings and Reviews.

| No.	First Author	Times Cited	Publication Year	Country	Title
				Article Document	
1	**Wamba S.F.**	**614**	**2017**	**France**	**Big data analytics and firm performance**
2	Akter S.	419	2016	Australia	How to improve firm performance using big data analytics capability and business strategy alignment?
3	Gunasekaran A.	382	2017	USA	Big data and predictive analytics for supply chain and organizational performance
4	Papadopoulos T.	265	2017	UK	The role of Big Data in explaining disaster resilience in supply chains for sustainability
5	Plageras A.P.	262	2018	Greece	Efficient IoT-based sensor BIG Data collection processing and analysis in smart buildings
				Conference Paper	
1	Katal A.	507	2013	India	Big data: Issues, challenges, tools and Good practices
2	Ishwarappa	163	2015	India	A brief introduction on big data 5Vs characteristics and hadoop technology
3	Archenaa J.	149	2015	India	A survey of big data analytics in healthcare and government
4	Kiran M.	130	2015	UK	Lambda architecture for cost-effective batch and speed big data processing
5	Singh S.	96	2012	India	Big Data analytics
				Reviews	
1	Mehta N.	138	2018	India	Concurrence of big data analytics and healthcare: A systematic review
2	Mishra D.	115	2018	India	Big Data and supply chain management: a review and bibliometric analysis
3	Kumari A.	93	2018	India	Multimedia big data computing and Internet of Things applications: A taxonomy and process model
4	Palanisamy V.	69	2019	India	Implications of big data analytics in developing healthcare frameworks â€" A review
5	Gupta S	68	2020	France	Big data in lean six sigma: a review and further research directions
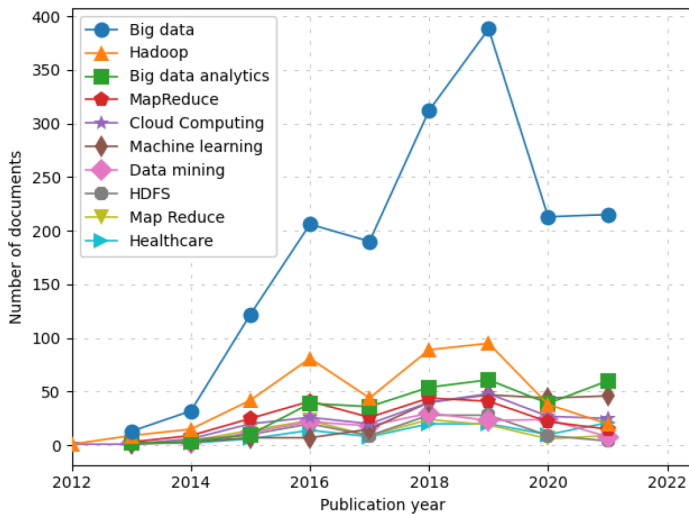
Figure 4: Trend of the research topics based on author key words per year.
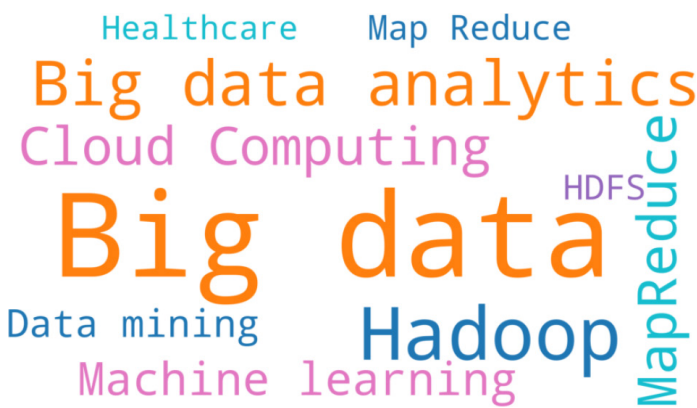

Figure 5: Word Cloud map of Top authors' keywords documents.

Most Productive Journals

Scholars of Big data research in India preferred the 10 most productive journals for their publishing research papers are shown in the Table 6. *Advances in Intelligent Systems and Computing* (199) was the top journal by publication output, followed by Journal of Advanced Research in Dynamical and Control Systems (91). The Procedia Computer Science (39) received the *h*-index of 15 among the top 10 journals.

Most Productive Institutions

A total of 1248 affiliations were identified where the universities were the most participative. Table 7 shows the top ten institutions with the highest number of documents identified. It is observed that Vellore Institute of Technology has the leadership, followed by the Amity University.

CONCLUSION

A scientometrics review of Big Data was conducted using a data collection of 3080 papers produced over a 10-year period

Table 6: Most Productive Journals in Big Data Research, 2012-2021.

No	Source Title	Total	*h_*index
1	Advances in Intelligent Systems and Computing	199	10
2	Journal of Advanced Research in Dynamical and Control Systems	91	3
3	International Journal of Innovative Technology and Exploring Engineering	81	4
4	International Journal of Recent Technology and Engineering	78	4
5	International Journal of Applied Engineering Research	59	6
6	International Journal of Engineering and Technology (UAE)	47	7
7	Communications in Computer and Information Science	43	3
8	Procedia Computer Science	39	15
9	International Journal of Engineering and Advanced Technology	37	3
10	Smart Innovation, Systems and Technologies	36	5

Table 7: Most Productive Institutions.

No	Affiliation	Documents
1	Vellore Institute of Technology	172
2	Amity University	94
3	K L Deemed to be University	80
4	Sathyabama Institute of Science and Technology	50
5	Anna University	47
6	Vels Institute of Science, Technology and Advanced Studies	47
7	Thapar Institute of Engineering and Technology	46
8	Kalinga Institute of Industrial Technology, Bhubaneswar	46
9	Bharathiar University	45
10	Jamia Millia Islamia	33

between 2012–2021 from two databases Web of Science and Scopus.

This study provides an insight into recent Big Data research trends. 3080 publications were examined to know the top authors, country collaboration, document types, publication sources, keywords, and affiliations.

The findings revealed the number of publications in the Big Data research in India increased significantly from 2015 to 2021, and the citation rate increased overall. There was a lot of cooperation between India and the United States. With 47 documents. Kumar A, ranks first among the top ten most relevant authors. Half of the publications 1507were conference papers. "Advances in Intelligent Systems and Computing, Journal of Advanced Research in Dynamical and Control Systems and International Journal of Innovative Technology and Exploring Engineering" were the top three journals.

The keywords analysis showed that Big Data Analysis is the research hotspot. Vellore Institute of Technology has produced the highest publications in Big Data research publications in India. From the present research it is identified that in India within a short period of time, Big Data research has got accelerated. It is presently regarded as one of the most important developing study fields in computer sciences and associated disciplines.

CONFLICT OF INTEREST

The authors declare that no conflict of interest.

ABBREVIATIONS

AGR: Average Growth Rate; **ADY:** Average Documents per year; **Wos:** Web of Science.

REFERENCES

1. Ruiz-Rosero J, Ramirez-Gonzalez G, Viveros-Delgado J. Software survey: ScientoPy, a scientometric tool for topics trend analysis in scientific publications. Scientometrics. 2019;121(2):1165-88. doi: 10.1007/s11192-019-03213-w.

2. Singh PK Singh AP, Kumar P. Diffusion of big data in Indian scientific literature: Study of research productivity and scientific collaboration [internet]; 2017. [cited 25/5/2022].

3. Dhawan SM, Gupta BM, Gupta R. Mapping global research output in big data during 2007-16. Emerg Sci J. 2018 Jun 1;2(3):117-29. doi: 10.28991/esj-2018-01135.

4. Jun SP, Yoo HS, Choi S. Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. Technological Forecasting and Social Change. 2018 May 1;130:69-87. doi: 10.1016/j.techfore.2017.11.009.

5. Esfahani HJ, Tavasoli K, Jabbarzadeh A. Big data and social media: A scientometrics analysis. Int J Data Netw Sci. 2019;3(3):145-64.

6. Gupta V, Singh VK, Ghose U, Mukhija P. A quantitative and text-based characterization of big data research. J Intell Fuzzy Syst. 2019;36(5)(5, SI): 4659-75. doi: 10.3233/JIFS-179016.

7. Hassan SU, Bowman TD, Shabbir M, Akhtar A, Imran M, Aljohani NR. Influential tweeters in relation to highly cited articles in altmetricbig data. Scientometrics. 2019 Apr;119(1):481-93. doi: 10.1007/s11192-019-03044-9.

8. Zhang Y, Huang Y, Porter AL, Zhang G, Lu J. Discovering and forecasting interactions in big data research: A learning-enhanced bibliometric study. Technological Forecasting and Social Change. 2019;146(Jun):795-807. doi: 10.1016/j.techfore.2018.06.007.

9. Raban DR, Gordon A. The evolution of data science and big data research: A bibliometric analysis. Scientometrics. 2020 Mar;122(3):1563-81. doi: 10.1007/s11192-020-03371-2.

10. Bhairawa Putera P, Pasciana R. Big Data for Public Domain: A bibliometric and visualized study of the scientific discourse during 2000-2020. pgr. 2021 Jul 23;5(3):220. doi: 10.30589/pgr.v5i3.440.