Global Research Trends in "Big Data" during 2012-21: A Data Mining based on Scientometric Tools

Dhruba Jyoti Borgohain, Sunil Kumar Yadav and Manoj Kumar Verma

The term "big data" is becoming widespread throughout the world, as it has wide usage because it is no longer limited to the IT industry and entrepreneurship but has entered every aspect of media and communications. But the reason for using big data is only it's ability in searching, collecting and interpreting huge datasets. Now, the purpose of this paper is to scrutinize the papers related to big data and other relevant fields. Using bibliometric methods and techniques these papers are analyzed and reviewed. The author examined the citation behavior, co-authorship patterns, research hot spots, and trending topics using both traditional and state-of-the-art network visualization and text mining in bibliometric techniques. From the analysis of keyword co-occurrence, a set of 12 clusters of keywords are discovered. Each cluster with identical color and theme. The research hot spots discovered are blockchain, digital twin, artificial intelligence, and the internet of things. The emergence of these keywords indicate that these areas are having lot of scope for future research. To best of knowledge, this study is a first attempt to discover the research streams and will help the researchers to work in these areas to explore new areas and applications of big data in relevant fields.

Introduction

A large amount of data has been generated in every sphere of human life with the advent of the Internet of things (IoT), diffusion of Web 2.0 and Web 3.0 technologies, and initiation of industry 4.0 (Atzori et al., 2010; Gubbi et al., 2013; Devenport et al., 2012). This has begun the era of big data. In this era, the data is having three characteristics, more particularly 3Vs: volume, variety and velocity (McAfee and Brynjolfsson, 2012). Volume signifies the quantity of data; variety signifies the types and velocity signifies the robustness of data generated. These 3Vs develop the process of filtering data into valuable knowledge as it involves a very complex task of storage, standardization, quality, and security issues (Cai and Zhu, 2015; Tankard, 2012). Research has highlighted that the requirements for big data are increasing, which will be having a powerful impact on fields like computer science, medicine, social media, government, economic systems, etc. (Chen et al., 2014; Murdoch and Detsky, 2013; Xiang et al., 2015). So, many studies are encountered to help data analytics with efficient and effective algorithms with an aim to solve the shortcomings of data analytics (Landset et al., 2015; Meng et al., 2014; Singh and Reddy, 2014).

It is evident that most of the studies related to big data are limited to specific areas. The study that is focused solely to discover research papers on big data as a whole are very limited. To be specific, a robust review of big data research taking a large dataset has not been discovered yet. To address these gaps in research in

big data in different areas, researchers need to evaluate and review the comprehensive sources and databases regarding the papers published in the field. It has been observed that Scopus, a product of Elsevier is the most reliable, efficient, and popular abstract and citation database, which serves the data requirements to conduct any kind of bibliometric study (Borgohain et al., 2021; Sweileh, 2021). For these the following research questions are framed:

RQ1. What is the extent of growth of research in big data in recent years?

RQ2. What are the suggestions which can be offered to boost research in this area?

To make the analysis scientific, this study used the bibliometric method, a technical and efficient analysis method of publications. The manuscript is structured into 5 sections. Section 2 is divided into two sub-sections: one reviews some papers that are defining big data and its characteristics, and the second section reviewed some papers which are adopting bibliometrics on papers related to big data application in different fields. Section 3 describes the methodology used. Section 4 presents the data analysis and interpretation. Section 5 highlights further discussion on the analysis performed, implications of the study, conclusion, limitations, and future research scope of this study.

2. Literature Review

2.1 Big data: Definition and Features

With the advent of the internet in the 1980s, it was observed that many datasets are observed to be too huge to get processed with the usual software and algorithm, so in the late 1990s, the term "big data" came into existence (Cox and Ellsworth, 1997). The wide applicability and practicality of the concept bought new definitions of the term, "large growing dataset that includes heterogeneous formats: structured, unstructured and semi-structured data" (Oussous et al., 2018). Another one stated, "a large complex collection of data sets, which is difficult to process using on-hand database management tools and traditional data processing applications" (Furht and Villanustre, 2016). But a uniform idea has already been established, which distinguished big data from common traditional data.

As mentioned, big data is characterized by 3Vs: volume, velocity, and variety (Gandomi and Haider, 2015; Laney, 2001). But additionally, two Vs are amalgamated: value and veracity (Manogaran et al., 2016). These features facilitate better understanding of big data and directs the paths for exploitation.

2.2 Application of bibliometrics to evaluate the papers on big data and its application in different fields

A statistical method that is usually used in academic literature review for assessing the scientific outputs is a typical bibliometric study (Bellis, 2009). Many studies have been found to use this technique to study the big data-related literature. For instance, a study by Kalantari et al. (2017) used bibliometrics to evaluate the 6572 papers relevant to big data research, which are published from 1980 to 2015 and indexed on the Web of Science (WoS). Nobre and Tavares (2017) evaluated the papers on the application of big data and IoT in the

context of circular economy with data from Scopus from 2006 to 2015. An in-depth analysis was conducted by Liang and Liu (2018) used the bibliometric method to explore the research trends of research on big data and business intelligence by taking data from three premium databases: Social Science Citation Index (SSCI), Science Citation Index Expanded (SCIE) and Arts and Humanities Citation Index (AHCI). Liu et al. (2019) conducted a bibliometric analysis of papers related to big data published between 2013 and 2018 with data from Scopus that used SciVal metrics. Rialti et al. (2019) adopted the bibliometric method to review 170 papers on big data obtained from WoS core collection. Extracting data from WoS, Zhang et al. 2019 investigated the 5840 papers and revealed the research hotspots and core components of big data by applying text mining techniques.

Supplementary to this, some studies used the bibliometric technique to analyze the papers related to the application of big data. Ardito et al. (2018) produced a research paper concerning big data analytics for business and management gathering data from WoS and analyzed 478 highly relevant and filtered articles. Mishra et al. (2018) conceived a literature review on big data and supply chain management assessing the 286 articles published in the last decade. Czarnecka and Olczyk (2020) investigated the papers related to cyber ethics and big data with papers indexed in WoS published between 1900 and July 2020. Another study by Nobanee (2020) used visualization tools to evaluate the 2408 relevant documents from Scopus on big data in business. For analyzing the papers on big data in finance, Nobanee (2021) applied the bibliometric method on 1059 relevant papers with data from Scopus. Sahoo (2021) analyzed the application of big data in manufacturing in the field of business management to under the research trends in the field and the scope of investigations in the future. Using data from Scopus and factorial analysis in Biblioshiny, the study presents three macro research clusters in this field: strategic change management using big data, automation, and smart manufacturing and big data analytics for energy conservation and sustainable production.

An extensive quantitative analysis of the papers related to big data will complement this analysis. The paper will visualize the research collaboration network using VOSviewer, analyze these maps and interpret them to reveal the research trends.

3. Methodology

An "iterative cycle of defining appropriate search keywords, searching the literature, and completing the analysis" is needed for designing a literature review paper (Fahimnia et al., 2015; Tranfield et al., 2003). A study by Fahimnia et al., (2015) devised a five-step methodology for conducting a systematic review. It starts with the definition of the database to extract data, designing the search strategy and query (using relevant search terms), screening of initial search results, refinement of the search results, development of descriptive statistics, and finally the detailed analysis using bibliometrics.

As our target is to scrutinize the papers on pure big data research. It may include its application-related papers too. For this purpose, Scopus is the most reliable and efficient database to extract data for this purpose (Borgohain et al., 2021a). As it covers articles from almost all top journals in a subject discipline. The restriction has not been taken on quality or impact factor of the journals maintaining an inclusive nature

for this. But conference proceedings, book chapters and all non-peer reviewed has been left out to ensure an examination of quality articles only so as to make our results accurate (Borgohain et al., 2022; Meier, 2011). A screening routine is deployed to include articles which is identified to be the best search strategy (Apio et al., 2014).

The search strategy developed is put into the 'Title-Abstract-Keyword' field of the Scopus: "Data Analytics" OR "Machine Learning" OR "Large dataset" OR "Big data" OR "Hadoop". This resulted a total of 85476 documents. Further, the subject area excluded are "Social Sciences", "Materials Science", "Earth and Planetary Science", "Chemical Engineering", "Arts and Humanities", "Agricultural and Biological Sciences", "Pharmacology", "Dentistry", "Veterinary", "Nursing", "Immunology and Microbiology", "Psychology", "Pharmacology, Toxicology and Pharmaceutics", "Multidisciplinary". This resulted a total of 1998 papers. These papers are considered to be final for analysis. The Boolean operator "OR" is used as it is more inclusive than "AND", so that it incorporates all the related and relevant papers in the topic as such the analysis of which will generate some reliable results (Borgohain et al., 2021b).

After careful scrutiny of each paper, these are mined into software VOSviewer for visualization of co-operative networks (country & author co-authorship) to visualize the clusters. This reveals the cooperation network. After this, the file is imported to Biblioshiny (an R platform) for analysis of prolific authors, journals, author and journal impact, analysis of keywords, Bradford's law, Lotka's law. This reveals the trending topics of research, hot spots, whether the authors follow the productivity law (Lotka's law) and the journal scattering law (Bradford's law). The study specifically is determined to discover the research trends, collaboration behaviour, trending areas of research, author productivity behaviour and journal impact. To specify, the core objectives of this study are: (i) to reveal the chronological growth of papers in big data research from 2012 to 2021 (ii) to identify the prolific author, journal in big data research based on number of papers (NP), total citations (TC), h-index, citations per paper (CPP) (iii) to visualize the collaboration networks with analysis of co-authorship of countries and authors (iv) to discover the research hot spots and trending topics of research in this subject (v) to test the fitness of the Lotka's law.

4. Data Analysis and Interpretation

4.1 Chronological growth

Based on the general information of the 1998 articles, the chronological growth of the papers is examined with the Price Law (Price, 1963). As per the law, if the correlation coefficient (R2) of the exponential trendline is greater than the correlation coefficient (R2) of the linear trendline, then the Price law is fulfilled and the growth of papers per annum is exponential.

Here, from Figure 1 it is evident that value of R2 for the exponential trendline (red colour) is 0.0097 which is smaller than that of the linear trendline (green) 0.0375. This is clear that the publication trend is not exponential. As, highest number of papers is in the year 2018 (408) and lowest in 2012 (25), the curve in Figure 1 is not symmetrical i.e., data is not normally distributed.

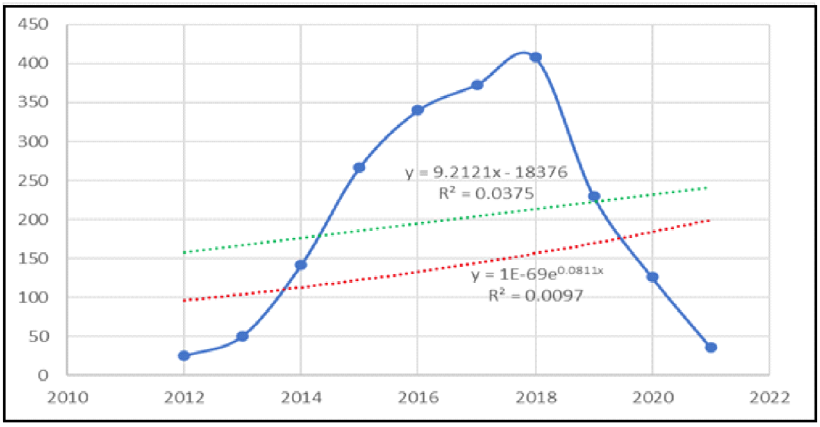

Figure 1: Chronological Growth of publications

4.2 Prolific Authors

Table 1 provides the top 10 authors in big data research. Authors with at least 10 papers and maximum of 25 papers are listed in the table. Analysis reveals that the distribution of the number of papers is highly skewed. For instance, these top 10 authors contributed 164 papers with 1012 with CPP of 6.17 papers. The top 10 authors together contributed 8.2% (164) global publication share and 9.47% (1012) global citation share. The group average of the productive top 10 authors is 16.4. On further analysis it was observed that, the 4 of the top 10 authors have contribution above the group average (21.5): X Li, Y Wang, L Afraites, G Andrienko. Moreover, the highly impactful author as per the normalized parameter, CPP is G Andrienko (11) with 198 citations in 18 papers, affiliated to City University London with h-index of 64.

Table 1: Scientometric Profile of top 10 most productive authors in research on 'big data'

Name	Affiliation	NP	TC	CPP	h-index
X Li	University of California at Davis	25	74	2.96	14
Y Wang	Nanchang University	23	218	9.48	12
L Afraites	University Sultan Moulay Slimane	20	76	3.8	8
G Andrienko	City University London	18	198	11	64
S Anter	Université Hassan II de Casablanca	16	91	5.69	5
C Boldrini	IIT-CNR Italy	15	88	5.87	23
B Bouikhalene	Sultan Moulay Slimane University	14	76	5.43	18
N Falih	University of Basrah	13	52	4	7
S Garg	Amazon	10	42	4.2	7
Giannotti	Universit di Pisa	10	97	9.7	54

4.3 Prolific Journals

The top 10 prolific journals are listed in Table 2. Journals with at least 48 or more papers are in the list. The maximum number of papers are for the journal published by Springer, Journal of Big Data with impact factor 10.835 and these articles were cited 198 times. These journals together produced 629 articles and received 1559 citations with CPP of 2.48 and 31.43% share in global share of number of publications (1998). The distribution of number of papers and citations for these top journals is also skewed. For example, the top 4 journals received total citations of 754 and have 282 papers with CPP of 2.67 which is more than the group average CPP of 2.48. Top 5 most productive journals are Journal of Big Data (75 papers), Electronics (71 papers), Energies (69 papers), International Journal of Electrical and Computer Engineering (67) and Big Data and Cognitive Computing (65 papers). Top 5 highly impactful journals in terms of CPP are: Journal of Business Research (4.48), Big Data Research (3.69), Electronics (3.21), Energies (3.04), Journal of Big Data (2.64) (Table 2).

Table 2: Scientometric profile of top 10 journals in "big data" research

Journal (IF)	Publisher	NP	TC	CPP
Journal of Big Data (10.835)	Springer	75	198	2.64
Electronics (2.690)	MDPI	71	228	3.21
Energies (3.252)	MDPI	69	210	3.04
International Journal of Electrical and Computer Engineering	Institute of Advanced Engineering and Science	67	118	1.76
Big Data and Cognitive Computing	MDPI	65	126	1.94
Mathematics (2.592)	MDPI	64	102	1.59
Big Data Research (3.739)	Elsevier	61	225	3.69
Frontiers in Big Data	Frontiers	58	95	1.64
Indian Journal of Computer Science and Engineering	Engg Journals Publications	51	42	0.82
Journal of Business Research (10.969)	Elsevier	48	215	4.48

4.4 Co-authorship of countries

In total, 90 countries participated in global research on "big data" but here also the distribution of productivity is highly skewed. Analysis of co-authorship of countries reveals the cooperation network of the nations that are participating. A network map of the top productive nations is generated using the VOSviewer. This is shown in Figure 2 which gives a visual presentation of their productivity, cooperation in work and the inter networks of collaborative research in the subject. The size of the circle in the figure representing a country is proportional to the productivity in terms of number of papers. The similarity in color of the circles representing a country indicated the close cooperation between the nations. The thickness of the lines

between the nations will represent the strength of collaborations. The network of the nations is divided into 7 clusters. Cluster 1 (red) represents 17 countries, some are: United Kingdom, Denmark, Ireland, Spain, Italy, Poland, Norway, France. Cluster 2 (green) represents 11 countries, some are: India, Iran, Malaysia, Taiwan. Cluster 3 (blue) has 10 countries, like South Korea, Saudi Arabia, Finland, Pakistan, Qatar, Saudi Arabia, Turkey, Jordan. Cluster 4 (yellow) has 8 papers, some are United States, Sweden, New Zealand, Japan, Hong Kong, China, Macau. The prominent U.S.A and China are included in this group. The 5th cluster (violet) has 4 countries Brazil, Portugal, Mexico, Russia. Cluster 6 and Cluster 7 (shallow blue & orange) have Canada and South Africa respectively. This figure is developed taking minimum number of documents for a country to be 5 it was found that out of the 90 participating nations, 52 countries meet the threshold.

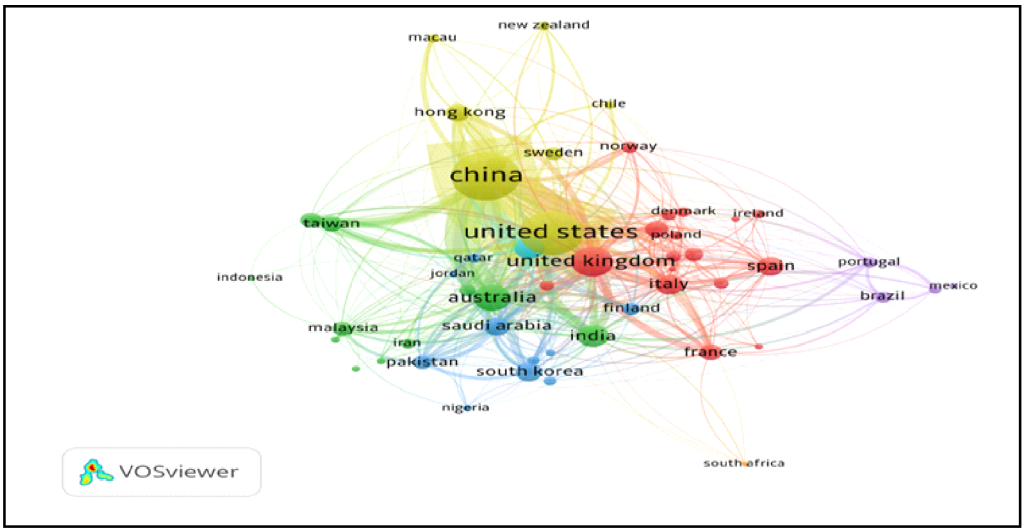

Figure 2: Network map of the collaboration countries

4.5 Co-authorship of authors

Analysis of co-authorship of authors reveals the network of authors that worked published papers in common affiliation. The network visualization, Figure 3 of the top authors contributing papers is developed using the VOSviewer using normalization method "Association Strength". Taking the minimum number of documents of an author to be 5, out of the 4964 authors that are participating in this field 232 have a minimum of 5 documents meeting the threshold. For each 232 authors, the total strength of the co-authorship links with other authors will be calculated. The authors with the greatest link strength are selected and the largest set of connected authors is found to be 217. Figure 2 depicts the collaborative network of these connected authors.

This map depicts the authors in 13 clusters. Of this, 7 clusters have the maximum papers. Cluster 1 (red) has the maximum number of 27 authors. Such as Y Chen, X Zhou, J Wu, B Wang, X Cheng, X Jiang. Cluster 2 (green) has 24 authors such as Y Wang, Q Qi, Q Chen, H Song, J Yang. Cluster 3 (blue) has 21 authors in all

such as H Liu, H Zhou, X Wu, W Zhang, M Wang. Cluster 4 (yellow) has 20 authors such as H Chen, D Zhang, Q Sun, J Zhu. 5th cluster (violet) has 19 authors in total such as M Song, KE Psannis, H Wang, Q Wang, X Huang. Cluster 6 (shallow blue) has 19 authors in all. Such as R Ranjan, J Shen, L Wang, N Kumar, D Chen. The 7th cluster (orange) has 19 authors in all. Such as W Wang, J Zhou, Z Zhou, W Wang, C Liu. Cluster 8 (brown) has 18 authors such as C Zhang, S Tang, D Li, S Wang, M Imran. The 9th cluster (purple) has total of 18 authors such as D Wang, J Wang, J Sun, J Zhao, J Lee. The 10th cluster marked in shallow red has 15 authors in all such as S Lan, J Lee, P Li, Z Chen, MJ Deen. The 11th cluster marked in light green has 6 authors such as A Alamri, S Singh, M Qiu, MS Hossain. 12th cluster (shallow blue) has 6 items like G Jeon, M Khan, A Paul, A Ahmad, MM Rathore. 13th cluster (shallow yellow) has 5 authors like S Gracia, F Herrera, S Del Rio. The top 4 authors with highest collaborative linkages are Y Zhang, Y Liu, Y Li, J Chen. YK Dwivedi, SP Singh, J Lee and K Yang are authors with least collaborative linkage.

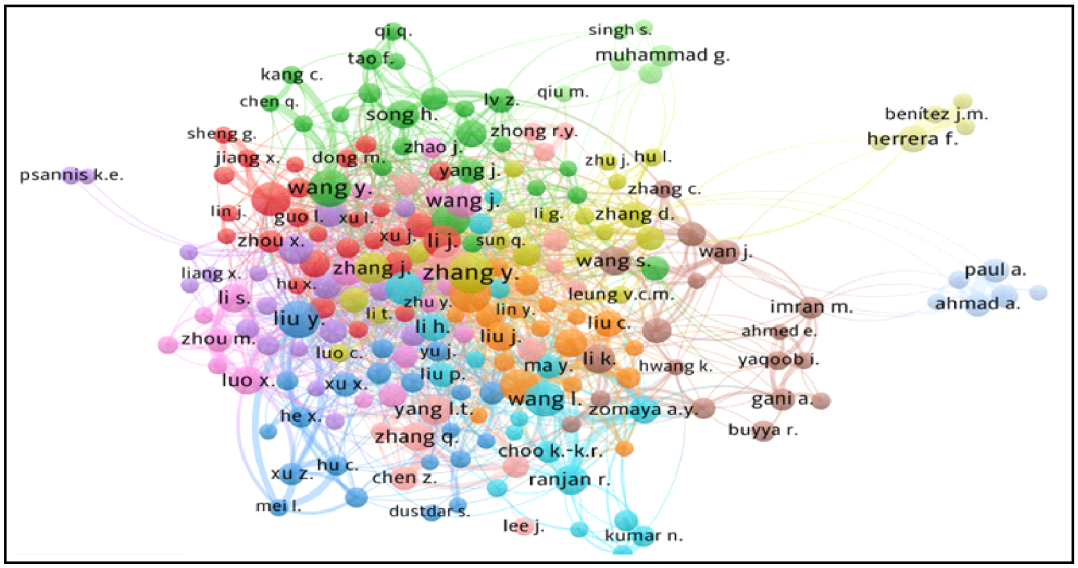

Figure 3: Network visualization of authors working in collaboration

4.6 Appropriateness of the Lotka's law

Lotka was the first to device the explicit idea called Lotka's law, one of the three basic bibliometric laws (Lotka, 1926) which was developed in order to examine the author productivity pattern in data sample of Chemistry and Physics papers. His formulation was challenged by Bookstein who came with a formulation (Bookstein, 1976) as,

$$g(x) = kx^{-n}, \text{ where } x = 1,2,3,4.....x_{max}, \quad k>0$$

here, the constants k and n are the parameters to be estimated from data while g(x) represents the fraction of authors publishing x articles.

Many studies have deployed this law into the author productivity data in different fields like Sudhier (2013) applied this law in Physics literature, Borgohain et al. 2021a applied this law in nanotechnology research. Let us observe how we applied this law to the author productivity data of big data research.

4.6.1 Evaluation of parameter 'n'

This is calculated with the formula of Linear Least Square Method as follows:

$$n = \frac{[N \sum(\ln(x) \cdot \ln g(x)) - \sum \ln g(x) \sum \ln(x)]}{[N \sum(\ln x)^2 - (\sum \ln x)^2]} \quad \ldots\ldots (1)$$

To estimate n, we take x as the number of articles and g(x) is the fraction of authors publishing x articles. The calculations done are shown as follows.

Table 3: Calculation of 'n'

x	g(x)	ln (x)	ln g(x)	ln (x) g (x)	ln (x) * ln (x)
1	3943	0.000	8.280	0.000	0.0000
2	592	0.693	6.384	4.424	0.4804
3	215	1.099	5.371	5.903	1.2069
4	94	1.386	4.543	6.297	1.9218
5	58	1.609	4.060	6.533	2.5902
6	39	1.792	3.663	6.564	3.2114
7	29	1.946	3.367	6.552	3.7866
8	13	2.079	2.565	5.333	4.3239
9	12	2.197	2.485	5.460	4.8287
10	11	2.303	2.398	5.523	5.3015
Total	5006	15.104	43.116	52.589	27.651

$$n = \frac{10x\ 52.589 - 43.116x15.104}{10x27.651 - 228.13} = -2.59$$

Since 'n' is taken for calculation of number of authors, so the positive value of it is taken i.e., 2.59. Now this value is utilized for calculating the number of authors anticipated using the equation (1) as shown in Table 4.

Table 4: No. of authors observed and expected

Articles (x)	Observed authors {g(x)}	% Observed	Expected authors	% Of exp. authors
1	3943	78.77	3943	78.77
2	592	11.83	98	1.98
3	215	4.3	13	0.26
4	94	1.88	3	0.06
5	58	1.16	1	0.02
6	39	0.78	0	0.00
7	29	0.58	0	0.00
8	13	0.26	0	0.00
9	12	0.24	0	0.00
10	11	0.22	0	0.00

Table 4 gives the number of authors observed and expected. These results indicate that at least one article is published by 3943 authors taking 78.77% share in cumulative author count (5006), which is both observed and anticipated. It is observed that 2 articles are contributed by 592 authors (11.83%) while 2 articles are expected to be contributed by 98 authors (1.98%) as per the calculation. Hence, a huge variation is discovered in number of authors observed and expected as per calculations performed. Hence, the dataset fails to obey the Lotka's law of author productivity.

4.7 Analysis of keywords

This is a very important analysis in a typical bibliometric analysis. Here, the analysis is performed in three ways: (1) analysis of keyword cooccurrence (2) analysis of research hot spots (3) analysis of keywords on the basis years.

The keyword cooccurrence is visualized with VOSviewer to identify the common themes of research. Taking the minimum number of cooccurrence of a keyword as 6, it was found that of the 5428 keywords that appeared in total 152 are found to meet the threshold. These 152 items are divided into 12 clusters. Cluster 1 (red) has 31 keywords in all such as neural networks, industrial big data, parallel computing, support vector machine, big data processing. The 2nd cluster (green) has 25 keywords like blockchain, cloud computing, artificial intelligence, biometrics, access control. Cluster 3 (blue) has 14 keywords in all such as bioinformatics, big data, energy efficiency, climate change. Cluster 4 (yellow) has 14 items in all which are very trending areas of research in the field like sentiment analysis, opinion mining, social media analytics, text mining, twitter, computational intelligence. Cluster 5 (violet) has 13 keywords in all like internet of

things, cybersecurity, logistics, cloud manufacturing, logistics etc. Cluster 6 (shallow blue) has 12 keywords in total like ontology, business intelligence, topic modelling, analytics, business support, decision support. The 7th cluster (orange) has 12 items in total like big data analytics, data fusion, big data analysis, energy consumption. Cluster 8 (brown) has 11 keywords in all like data analysis, visualization, predictive analytics, online social networks. The 9th cluster (purple) has 10 items in all such as higher education, collaborative filtering, learning analytics, optimization, cognitive computing. 10th cluster (shallow red) has 5 items like Hadoop, map reduce, scalability, smart city, spark. Cluster 11 (shallow green) has 4 items like big data applications, data mining, mobile computing, performance evaluation. 12th cluster (navy blue) has 1 keyword, 5G (Figure-4).

Figure 5 created using Biblioshiny represents the keywords that are prominent. The font size of the keyword is proportional to the frequency of occurrence of each keyword. Table 5 provides the detail of year wise occurrence of prominent author keywords in big data from 2012 to 2021. It reveals that highest frequency of occurrence of keyword is for the core word "big data" (995) followed by "cloud computing" (167) and "big-data analytics" (134), which are top three highly occurring keywords. The Figure 6 depicts the growth of top 10 highly occurring keyword with different colors.

The figure 7 generated with Biblioshiny to show the trending areas of research and hot spots in big data research. This is generated with: Field – Author keyword, the time span from 2012 to 2021, Word Minimum Frequency- 5, number of words per year- 5, and word label size 9. It is evident that the areas like "blockchain", "artificial intelligence", "prediction", "intelligent system", "business analytics", "internet of things". These trending areas of research indicated that big data analytics has the highest potential to be limitless which shows agility of network, integrated AI with industrial internet has the highest potential to automate diverse use of big data at hyper scale.

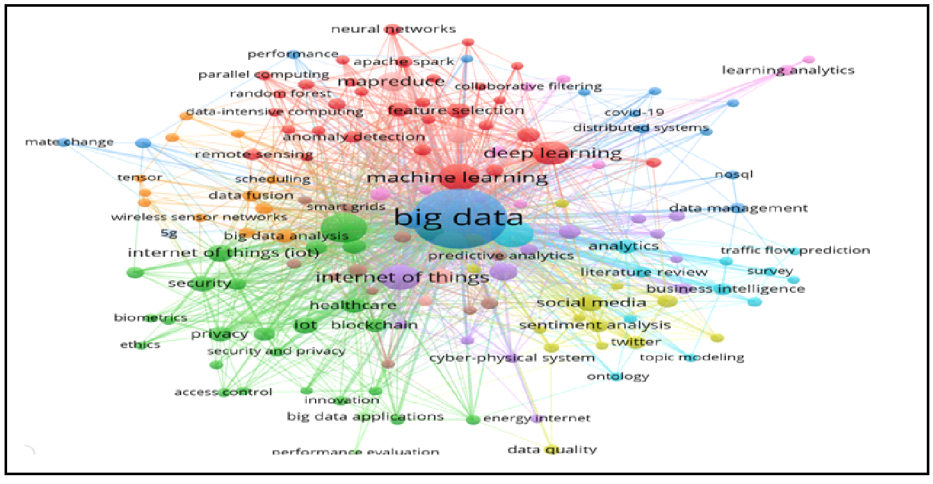

Figure 4: Analysis of keyword cooccurrence of highly occurring keywords

Figure 5: Word cloud of some prominent keywords

Table 5: Year wise occurrence of top prominent keywords

Year	Big Data	Cloud Computing	Big Data Analytics	Machine Learning	Internet of Things	Deep Learning	Data Mining	Map reduce	Industry 4.0	Data Analytics
2012	5	0	1	0	0	0	1	1	0	0
2013	40	9	2	2	1	0	5	2	0	2
2014	84	12	5	5	1	1	7	7	0	3
2015	157	37	13	13	7	5	9	18	2	4
2016	180	39	22	13	23	4	10	14	4	9
2017	171	24	16	12	20	13	10	5	7	10
2018	180	26	31	26	24	27	13	7	10	8
2019	106	15	24	25	20	27	11	2	14	5
2020	54	4	18	20	11	14	4	0	7	2
2021	18	1	2	3	1	4	0	0	2	1

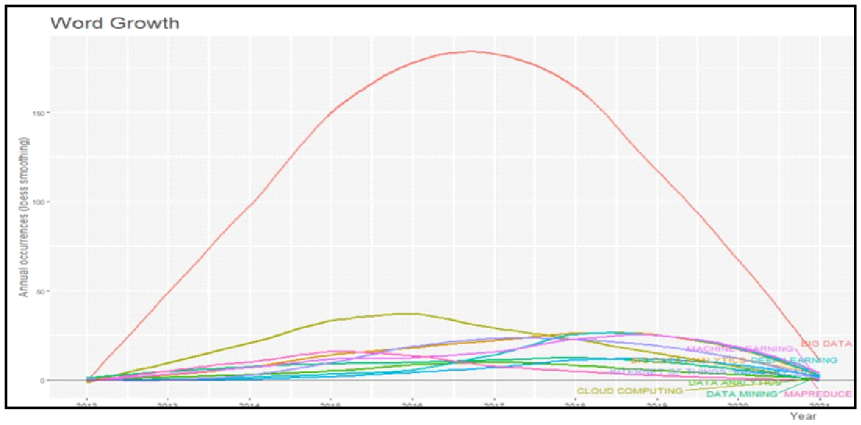

Figure 6: Year wise growth of top prolific keywords

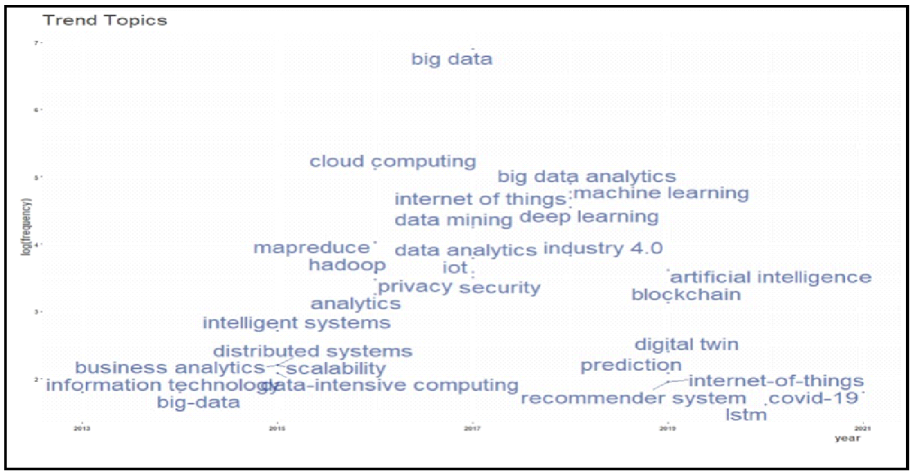

Figure 7: Visualization of research hot spots in big data research

5. Discussion

Bibliometrics is different from systematic review for its efficiency to summarize the current status the predict the future progress in a subject domain (Moller and Myles, 2016). Besides predicting development trends in a specific field, analysis can also be used to compare research performance across countries, authors, journals, and institutions. Moreover, the visualization tools like VOSviewer (van ECK and Waltman, 2010), HistCite (Gu et al., 2019), and CiteSpace (Synnestvedt et al., 2005) have been used to create maps, investigate cutting-edge research progress, visualization of trends in scientific papers (Modak et al., 2020). As per bibliometric analysis is concerned it can be classified into three types of analysis. First, analysis of performance with indicators like h-index (Hirsch, 2005), g-index (Egghe, 2006), total citations, number of papers, and collaborative coefficient. These are traditional metrics. Second, science mapping techniques include bibliographic coupling, co-word analysis, and keyword co-occurrence analysis. This is followed by visualization technique or network analysis or enriched bibliometric technique which uses visualization technique software tools like VOSviewer, Biblioshiny, CiteSpace, HistCite etc. (Donthu et al., 2021). This study incorporated all these techniques making it different from other studies reviewed here. Thereby fulfilling the research gap.

5.1 Strength and Limitations of the study

First strength is that the study can provide a panoramic knowledge in this area to the researchers and some keywords may predict the research hot spots and future directions. Second the study deployed two state-of-the-art network visualization tools to do the present study which provides comprehensive, reliable and efficient results. Third, use of premier database, Scopus for data extraction for this study is itself a strength.

The study has two major limitations. First, use of a single database (Scopus) for analysis may limit the results to a particular direction only. The findings in this case may not be diverse and may not predict the

future directions accurately. Use of two databases (like Scopus or WoS) would visualize the research trends in a more perfect way (Shi et al., 2019; Wu et al., 2021). Second, only papers in written in English language are considered and some important research papers might have been neglected. These limitations open up new directions for future research like if one more database like WoS is included then it would be a comparative study of papers from two major database which in turn will make the findings more inclusive and accurate.

6. Conclusion

This study has elaborated the present research status and emerging trends in big data research in global context. The number of papers has been increasing and USA & China has been ahead in terms of number of papers as well as citation frequency. Emergence of trending areas like blockchain, artificial intelligence, prediction, social media analytics and sentiment analysis depicts that this area of research is going to expand rapidly, expand and more studies are going to get published in the ensuing years as evident from the analysis of chronological growth of papers. To be specific, the shifting of areas of research from "big data" to "blockchain", "artificial intelligence", studies on "digital twin", "prediction", "internet of things" will be the next potential areas of research. In future, the scientist and funders working and funding in this field may get attention to this topic for research and open up new techniques of application of big data in different fields.

References

1. Appio , F., Cesaroni , F., & Di , M. A. (2014). Visualizing the structure and bridges of the intellectual property management and strategy literature: a document co-citation analysis. Scientometrics. 101(1), 623-661.

2. Ardito , L., Scuotto , V., Giudice, M. D., & Petruzzelli, A. M. (2019). A bibliometric analysis of research on Big Data analytics for business and management. Management Decision, 57(8), 1993-2009. doi:10.1108/MD-07-2018-0754

3. Atzori, L., Iera, A., & Morabito, G. (2010). The Internet of Things: a survey. Computer Networks, 54(15), 2787-2805.

4. Bellis, N. D. (2009). Bibliometrics and Citation Analysis: from the Science Citation Index to Cybermetrics. Lanham, MD: Scarecrow Press.

5. Bookstein , A. (1976). The Bibliometric distributions. Library Quarterly, 46(4), 416-423.

6. Borgohain , D. J., Mohammad , N., & Verma , M. K. (2022). Cluster analysis and network visualization of research in mucormycosis: a scientometric mapping of the global publications from 2011 to 2020. Library Hi Tech, ahead-of-print. doi:10.1108/LHT-04-2022-0171

7. Borgohain, D. J., Sohaimi , Z., & Verma , M. K. (2021a). Cluster Analysis and Network Visualization of Global Research on Digital Libraries during 2016–2020: A Bibliometric Mapping. Science & Technology Libraries, 40. doi:10.1080/0194262X.2021.1993422

8. Borgohain, D. J., Verma , M. K., & Daud, S. C. (2021b). Scientometric Profile of Fisheries Research in SAARC Countries. DESIDOC Journal of Library and Information Technology, 41(6), 429-437. doi:10.14429/djlit.41.6.16986

9. Cai , L., & Zhu , Y. (2015). The challenges of data quality and data quality assessment in the Big Data era. Data Science Journal, 14(2), 1-10.

10. Chen , M., Mao , S., & Liu , Y. (2014). Big data: a survey. Mobile Networks and Applications, 19(2), 171-209.

11. Cox , M., & Ellsworth , D. (1997). Managing big data for scientific visualization. ACM SIGGRAPH, 97, 21-38.

12. Davenport, T. H., Barth, P., & Bean, R. (2012). How 'Big Data' is different. MIT Sloan Management Review, 54(1), 43-46.

13. Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. (2021). How to conduct a bibliometric analysis: An overview and guidelines. Journal of Business Research, 133, 285-296. doi:https://doi.org/10.1016/j.jbusres.2021.04.070

14. Egghe , L. (2006). Theory and practise of the g-index. Scientometrics, 69(1), 131-152.

15. Fahimnia, B., Sarkis, J., & Davarzani, H. (2015). Green supply chain management: a review and bibliometric analysis. International Journal of Production Economics, 162, 101-114.

16. Furht, B., & Villanustre, F. (2016). Introduction to big data. (B. Furht, & F. Villanustre, Eds.) Springer International Publishing.

17. Gandomi , A., & Haider, M. (2015). Beyond the hype: big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137-144.

18. Gu , D., Li, T., Wang , X., Yang , X., & Yu , Z. (2019). Visualizing the intellectual structure and evolution of electronic health and telemedicine research. International Journal of Medical Informatics, 130, 103947. doi:doi:10.1016/j.ijmedinf.2019.08.007

19. Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of Things (IoT): a vision, architectural elements, and future directions. Future Generation Computer Systems, 29(7), 1645-1660.

20. Hirsch, J. (2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America, 102 (46), pp. 16569-16572.

21. Kalantari, A., Kamsin, A., Kamaruddin, H. S., Ale Ebrahim, N., Gani, A., Ebrahimi, A., & Shamshirband, S. (2017). A bibliometric approach to tracking big data research trends. Journal of Big Data, 4(1), 4-30.

22. Kuc-Czarnecka , M., & Olczyk, M. (2020). How ethics combine with big data: a bibliometric analysis. Humanities & Social Sciences Communication, 7(137), 1-9. doi:10.1057/s41599-020-00638-0

23. Landset , S., Khoshgoftaar, T. M., Richter, A. N., & Hasanin, T. (2015). A survey of open source tools for machine learning with Big Data in the Hadoop ecosystem. Journal of Big Data, 2(1), 24-49.

24. Laney, D. (2001, June 30). 3-D data management: controlling data volume, velocity and variety. META Group Research Note. Stamford. Retrieved July 2, 2022, from META Group Research Note, Stamford: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

25. Liang , T. P., & Liu , Y. H. (2018). Research landscape of business intelligence and big data analytics: a bibliometrics study. Expert Systems with Applications, 111, 2-10.

26. Liu , X., Sun , R., Wang , S., & Wu, Y. J. (2019). The research landscape of big data: a bibliometric analysis. Library Hi Tech, 38(2), 367-384. doi:10.1108/LHT-01-2019-0024

27. Lotka , A. (1926). The Frequency Distribution of Scientific Productivity. Journal of the Washington Academy of Sciences, 16(12), 317-323.

28. Manogaran, G., Thota, C., & Kumar, M. V. (2016). MetaCloudDataStorage architecture for big data security in cloud computing. Procedia Computer Science, 128-133.

29. McAfee, A., & Brynjolfsson, E. (2012). Big Data: the management revolution. Harvard Business Review, 90(10), 60-68.

30. Meier, M. (2011). Knowledge management in strategic alliances: a review of emperical evidence. International Journal of Management Reviews, 13(1), 1-23.

31. Meng , S., Dou , W., Zhang , B., & Chen , C. (2014). KASR: a keyword-aware service recommendation method on MapReduce for Big Data applications. IEEE Transactions on Parallel and Distributed Systems, 25(12), 3221-3231.

32. Mishra, D., Gunasekaran, A., Papadopoulos, T., & Childe, S. (2018). Big data and supply chain management: a review and bibliometric analysis. Annals of Operations Research, 270(1-2), 313-336.

33. Modak , N., Sinha , S., Raj , A., Panda , S., Merigo , J., & Lopes De Sousa Jabbour, A. (2020). Corporate social responsibility and supply chain management: framing and pushing forward the debate. Journal of Cleaner Production, 273 (122981). doi:10.1016/j.jclepro.2020.122981

34. Moller , A. M., & Myles , P. S. (2016). What makes a good systematic review and meta-analysis? British Journal of Anaesthesia, 117(4), 428-430. doi:doi:10.1093/bja/aew264

35. Murdoch , T. B., & Detsky , A. S. (2013). The inevitable application of big data to health care. JAMA, 309(13), 1351-1352.

36. Nobanee , H. (2020). Big data in business: a bibliometric analysis of relevant literature. Big Data, 8(6), 459-463. doi:0.1089/big.2020.29042.edi

37. Nobanee , H. (2021). A bibliometric review of big data in finance. Big Data, 9(2), 73-78. doi:10.1089/big.2021.29044.edi

38. Nobre , G. C., & Tavare , E. (2017). Scientific literature analysis on big data and Internet of Things applications on circular economy: a bibliometric study. Scientometrics, 111(1), 463-492.

39. Oussous , A., Benjelloun, F. Z., Lahcen, A. A., & Belfkih, S. (2018). Big data technologies: a survey. Journal of King Saud University – Computer and Information Sciences, 30(4), 431-448.

40. Price , D. (1963). Little science, big science. New York: Columbia University Press.

41. Rialti, R., Marzi, G., Ciappei, C., & Busso, D. (2019). Big data and dynamic capabilities: a bibliometric analysis and systematic literature review. Management Decision, 57(8), 2052-2068.

42. Sahoo , S. (2021). Big data analytics in manufacturing: a bibliometric analysis of research in the field of business management. International Journal of Production Research, 86, 122-134. doi:10.1080/00207543.2021.1919333

43. Shi , J. G., Miao , W., & Si , H. Y. (2019). Visualization and analysis of mapping knowledge domain of urban vitality research. Sustainability, 11(4), 988. doi:https://doi.org/10.3390/su11040988

44. Singh , D., & Reddy , C. K. (2014). A survey on platforms for Big Data analytics. Journal of Big Data, 2(1), 8-27.

45. Sudhier , K. P. (2013). Lotka's law and pattern of author productivity in the Area of Physics Research. DESIDOC Journal of Library and Information Technology, 33(6), 457-464.

46. Sweileh, W. M. (2021). Global research publications on irrational use of antimicbrobials: call for more research to contain antimicrobial resistance. Globalization and Health, 17(94). doi:10.1186/s12992-021-00754-9

47. Synnestvedt, M. B., Chen , C., & Holmes , J. H. (2005). CiteSpace II: visualization and knowledge disocovery in bibliographic databases . AMIA Annual Symposium Proceedings , 724-728.

48. Tankard , C. (2012). Big data security. Network Security, 2012(7), 5-8.

49. Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. British Journal of Management, 14(3), 207-222.

50. Wu, H., Tong, L., Wang, Y., Yan , H., & Sun, Z. (2021). Bibliometric Analysis of Global Research Trends on Ultrasound Microbubble: A Quickly Developing Field. Frontiers in Pharmacology, 12 (646626). doi:10.3389/fphar.2021.646626

51. Xiang , Z., Schwartz, Z., Gerdes , J. J., & Uysal , M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? International Journal of Hospitality Management, 44, 120-130.

52. Zhang , Y., Huang , Y., Porter , A. L., Zhang , G., & Lu , J. (2019). Discovering and forecasting interactions in big data research: a learning-enhanced bibliometric study. Technological Forecasting and Social Change, 146, 795-807.

Keywords: Data Visualization; Text Mining; Bibliometric Analysis; Bibliometrix R; Big Data; Network Visualization; Cluster Analysis

About Authors

Mr. Dhruba Jyoti Borgohain
PhD Reseach Scholar
Department of Library and Information Science, Mizoram University, Aizawl
Email: dhrubadlismzugu@gmail.com

Dr. Sunil Kumar Yadav
Ex-Research Scholar
Department of Library and Information Science, Mizoram University, Aizawl
Email: sunillerha@gmail.com

Dr. Manoj Kumar Verma
Professor
Department of Library and Information Science, Mizoram University, Aizawl
Email-: manojdlis@mzu.edu.in