Original Research Article

Publications trends in big data: A scientometric analysis

S. Baskaran[1],*

[1]*University of Madras Marina Campus Library, Chennai, Tamil Nadu, India*

ARTICLE INFO

ABSTRACT

The present paper is aimed at analysing the global publications trends on big data using Web of Science database for the period 2011-2020. The database contained 45249 publications. The average number of publications per year was 4524.9. There were only 1058 publications in 2011 and a continuous growth of publication was observed during the period. The highest publication 9389 was in 2020. Authors from USA have contributed maximum number of publications compared to the other countries and India stood 11th in terms of productivity in this period. Publications on big data are spread over 22 languages. The study analysed the broad features of literature on big data focusing on growth rate of publications, most prolific authors, country wise distribution of publications, highly productive institutes, language wise distribution of publications, most preferred source titles and high productivity subject areas.

1. Introduction

Big data is being generated by everything around us at all times. Every digital process and social media exchange produces it. Systems, sensors and mobile devices transmit it. Big data is arriving from multiple sources at an alarming velocity, volume and variety. To extract meaningful value from big data, you need optimal processing power, analytics capabilities and skills. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Big data is used to better understand customers and their behaviors and preferences. Big data is also used to optimize business processes, personal quantification and performance optimization, improving healthcare and public health, sports performance, science and research, security and law enforcement and optimizing machine and device performance. Therefore, the present study has been undertaken in order to know the growth and development of publications in the field of big data research as indexed in web of science database.

1.1. Objectives of the study

The objective of the study was to perform a scientometric analysis of all big data publications in the world. The parameters studied include:

1. Annual Growth Rate, Relative Growth Rate and Doubling Time of publications
2. Highly productive countries
3. Highly productive institutes
4. Most preferred source titles for publication
5. Language-wise distribution of big data research output
6. High productivity subject areas

2. Materials and Methods

Data was collected from the Web of Science database during 2011-2020, using search terms namely 'big data' in 'topic filed'. Web of Science database is one of the very

---

* *Corresponding author.*
*E-mail address*: baskaranphd@gmail.com (S. Baskaran).

comprehensive bibliographic databases covering all aspects of science and technology. [1,2] A total of 45249 publications and analysed by using the spread sheet application as per the objectives of the study.

2.1. Data analysis and interpretations

2.1.1. Form of publications

Table 1: Form of publications

S. No.	Form of publications	No. of publications	Percentage
1	Journal Articles	36958	81.68
2	Review Articles	3278	7.24
3	Editorial Materials	2195	4.85
4	Meeting Abstracts	971	2.15
5	Proceeding Papers	880	1.94
6	Early Access	393	0.87
7	Book Chapters	163	0.36
8	Letters	157	0.35
9	Book Reviews	136	0.30
10	News Items	90	0.20
11	Data Papers	28	0.06
Total		45249	100
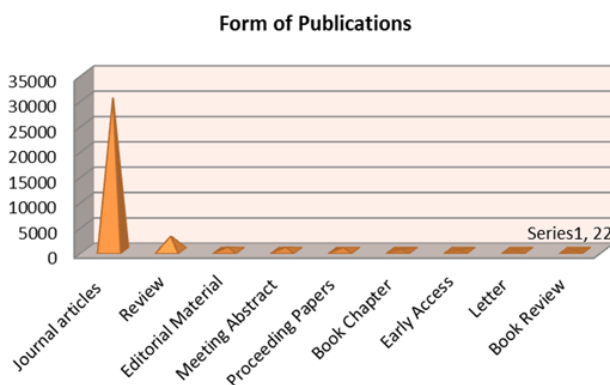
Fig. 1: Form of publications

The Table 1 reveals that the major source of publications covered by web of science databases on big data research is Journal Articles with 36,958 publications (81.68%) followed by Reviews with 3278 publications (7.24%). Editorial Materials ranks the third position with 2195 publications (4.85%), Meeting abstracts with 971 publications (2.15%), Proceeding papers with 880 publications (1.94%) and remaining forms are less than one percentage as seen in the table. The results indicate that the research outputs on the subject of the period covered by the study are mostly published in the form of journal articles. [3,4]

2.2. Annual Growth Rate (AGR) of publications

provides the AGR and CAGR of the number of documents for period 2006 to 2015.

$$AGR = \frac{End\ Value - First\ Value}{First\ Value} * 100$$

Table 2: GR of publications

Year	No. of publications	Cumulative total	Annual growth rate (AGR)
2011	1058	1058	-
2012	1187	2245	112.19
2013	1686	3931	75.10
2014	2392	6323	60.85
2015	3394	9717	53.68
2016	4545	14262	46.77
2017	5592	19854	39.21
2018	7147	27001	36.00
2019	8859	35860	32.81
2020	9389	45249	26.18

During the period of 2011 to 2020, a total of 45,249 publications were published on big data research. The highest number of publications is 9389 was published in 2020. The lowest publications of 1058 are published in 2011. The average number of publications published per year was 45289. Table 2 show that there has been a steady growth in research publications on big data during the study period.

The Table 2 also provides that the annual growth rate of the total publications calculated year wise. AGR reveals that it has decreased from 112.19 in 2012 to 26.18 in 2020. There is a downward trend in the growth rate as seen in the Figure 2.
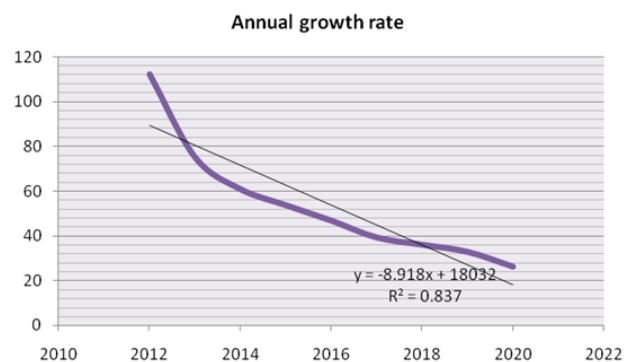
Fig. 2: Annual growth rate of publications

2.3. Relative growth rate and doubling time

The Relative Growth Rate (RGR) is the increase in a number of articles or pages per unit of time. This definition derived

from the definition of relative growth rates in the study of growth analysis in the field of big data. The mean relative growth rate (R) over the specific period of the interval can be calculated from the following equation.

1. Relative Growth Rate (RGR
2. 2. 1 - 2R$_=$Log W$_2$ – Log W$_1$/ T$_2$-T$_1$
3. Whereas
4. 1-2 R- mean relative growth rate over the specific period of interval
5. Loge W1 — log of the initial number of articles
6. 6. Log$_e$ W$_2$ - log of the final number of articles after a specific period of interval
7. T2-T1- the unit difference between the initial time and the final time
8. The year can be taken here as the unit of time
9. Doubling Time (DT = 0 693/R

Table 3: Relative growth rate (RGR) and doubling time (DT) of publications

Year	No. of Publications	Cumulative Total	W1	W2	RGR	DT
2011	1058	1058	-	6.96	-	-
2012	1187	2245	6.96	7.72	0.76	0.91
2013	1686	3931	7.72	8.28	0.56	1.24
2014	2392	6323	8.28	8.75	0.47	1.47
2015	3394	9717	8.75	9.18	0.43	1.61
2016	4545	14262	9.18	9.57	0.39	1.78
2017	5592	19854	9.57	9.90	0.33	2.1
2018	7147	27001	9.90	10.20	0.30	2.31
2019	8859	35860	10.20	10.49	0.29	2.39
2020	9389	45249	10.49	10.72	0.23	3.01

Table 3 Indicates that the RGR is decreased from 0.76 in 2012 to 0.23 in 2020. The highest value 0.76 corresponds to the year 2012, whereas the lowest value 0.23 for the year 2020. Correspondingly, the Doubling Time of the publications gradually increased from 0.91 in 2012 to 3.01 in 2020.
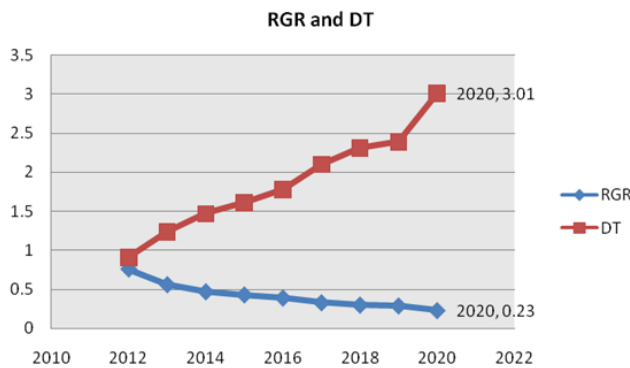

Fig. 3: Relative growth rate and doubling time

2.4. Most prolific authors

Table 4: Most prolific authors

S. No.	Author	No. of publications	Percentage
1	Zhang Y	268	0.59
2	Liu Y	232	0.51
3	Li Y	197	0.44
4	Wang Y	191	0.42
5	Wang J	160	0.35
6	Li X	156	0.34
7	Zhang J	152	0.33
8	Li J	144	0.32
8	Wang L	141	0.31
10	Liu J	122	0.27

The data on big data research publication during 10 years between 2011 and 2020 reveals that in total, 87,541 authors contributed to the publishing of the 45249 publications. The authors having 120 or more publications during 2011-2020 are shown in Table 3. Zhang, Y is the most productive author with 268 (0.59%) publications followed by Liu Y with 232 (0.51%) publications, Li Y with 197 (0.44%) publications, Wang Y with 191 (0.42%) publications and Wang J with 160 (0.35%) publications respectively.
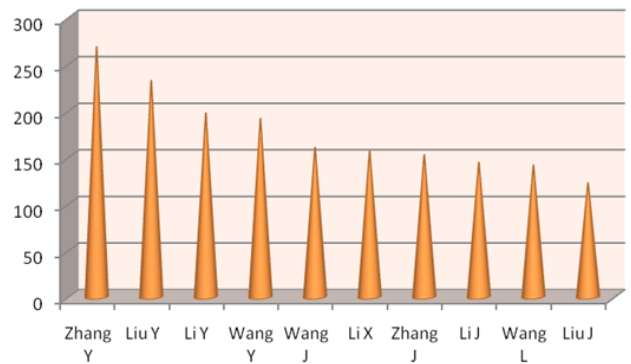
Fig. 4: Highly prolific authors

2.5. Highly productive institutions

A total of 16968 organizations are contributed entire research output of the study. The scientometric profile of top 10 institutions is presented in Table 5 . Findings revealed that Chinese Academy of Science, China with 1423 (3.14%) publications is the most productive institutions in the field of big data research followed by University of California System, USA with 1264 (2.79%) publications, University of London, UK with 812 (1.79%) publications, Harvard University, USA with 702 (1.55%) publications Centre National De La Recherche Scientifique CNRS, France with 613 (1.35%) publications, and University of Texas System USA with 539 (1.24%) publications.

Table 5: Highly productive institutions

| S. No.	Institutions	Country	No. of Publications
1	Chinese Academy of Science	China	1423 (3.14%)
2	University of California System	USA	1264 (2.79%)
3	University of London	UK	812 (1.79%)
4	Harvard University	USA	702 (1.55%)
5	Centre National De La Recherche Scientifique CNRS	France	613 (1.35%)
6	University of Texas System	USA	559 (1.24%)
7	State University System of Florida	USA	495 (1.09%)
8	Tsinghua University	China	454 (1.00%)
9	University of Chinese Academy of Science	China	441 (0.97%)
10	Pennsylvania Commonwealth System of Higher Education	USA	430 (0.95%)

2.6. Highly productive countries

Table 6: Highly productive countries

| S. No.	Country	Total Publications (%)
1	USA	13444 (29.71%)
2	China	11862 (26.21%)
3	England	4128 (9.12%)
4	Germany	2950 (6.52%)
5	Australia	2607 (5.76%)
6	Italy	2195 (4.85%)
7	Spain	2120 (4.69%)
8	Canada	2043 (4.52%)
9	South Korea	1827 (4.04%)
10	France	1743 (3.85%)
11	India	1602 (3.54%)

The publication share of highly productive countries (≥1600 publications) on big data is given in Table 6. In all, there were 172 countries involved in the research in big data; however, USA topped the list with highest share (29.71%) of publications. China ranked second with 26.21% share of publications followed by England 9.12% share of publications, Germany with 6.52% share of publications, Australia with 5.76% share of publications, Italy with 4.85% share of publications and Spain with 4.69% share of publications.

2.7. Language wise distribution

Publications on big data are spread over 22 languages. The study reveals that the maximum number of publications
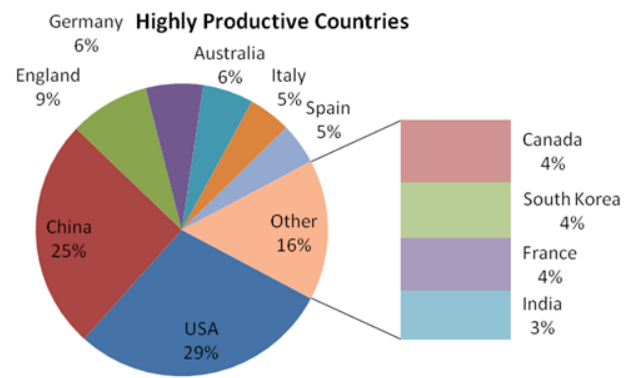
Fig. 5: Highly productive countries

Table 7: Language-wise distribution

S. No.	Languages	Total Publications (%)
1	English	44201 (97.68%)
2	Chinese	284 (0.63%)
3	German	217 (0.48%)
4	Spanish	172 (0.38%)
5	French	96 (0.21%)
6	Portuguese	72 (0.16%)
7	Polish	32 (0.07%)
8	Russian	32 (0.07%)
9	Hungarian	26 (0.06%)
10	Turkish	23 (0.05%)
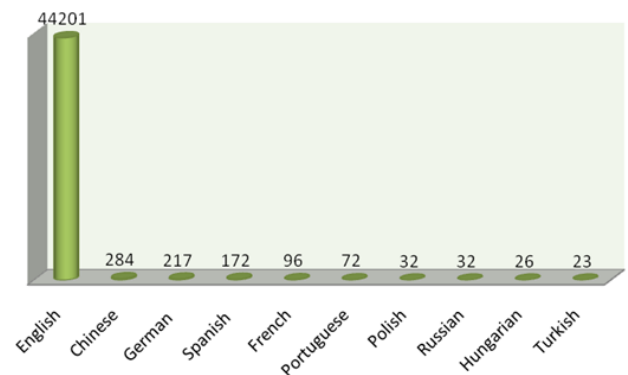
Fig. 6: Language wise distributions

have been published in the English language with 44201 (97.68%) publications, followed by Chinese language with 284 (0.63%) publications, German language ranks third position with 217 (0.48%) publications, Spanish language with 172 (0.38%) publications, French language with 96 (0.21%) publications. And the remaining languages such as Portuguese, Polish, Russian, Hungarian and other languages are constituted in negligible percentage. The English language superiority was found in every year in total productivity on the subject during the study period.

Table 8: Source title of publications

S. No.	Source Title	No. of Publications	Percentage
1	IEEE Access	1258	2.78
2	Future generation system the international journal of e-science	491	1.09
3	PLOS one	489	1.08
4	Sustainability	471	1.04
5	Sensors	304	0.67
6	Cluster computing the journal of networks software tools and applications	239	0.53
7	Concurrency and computation practice experience	231	0.51
8	Applied sciences basel	223	0.49
9	Journal of supercomputing	217	0.48
10	Multimedia tools and applications	215	0.47
11	Journal of cleaner production	213	0.47

2.8. Major source title of publications

The publication share of most productive source titles (≥200 publications) on big data is given in Table 8. The scientific literature on big data is spread over 3951 different source journals and conference publications. It reveals that IEEE Access the list with the highest number of publications 1258 (2.78%) followed by Future generation system the international journal of e-science with a share of 491 (1.09%) publications. The PLOS one occupies third position with 489 (1.08%) publications. The fourth highest source title is Sustainability with 471 (1.04%) publications and Sensors with 304 (0.67%) publications.

2.9. High productivity subject areas

Table 9: High productivity subject areas

S. No.	Subject	No. of Articles	Percentage
1	Computer science	11671	25.79
2	Engineering	8743	19.32
3	Environmental sciences ecology	3359	7.42
4	Telecommunications	3177	7.02
5	Business economics	3170	7.01
6	Science technology	2669	5.90
7	Physics	1647	3.64
8	Psychology	1619	3.58
9	Chemistry	1575	3.48
10	Mathematics	1484	3.28

Table 9 Shows high productivity subjects which are contributing more than1400 articles. It is found that Computer science has the highest number of articles with 11671 (25.79%) followed by Engineering contributing 8743 (19.32%) articles. Environmental sciences ecology occupies the third position with 3359 (7.42%) articles. The fourth highest articles belonged to the subject Telecommunications with 3177 (7.02%), Business economics with 3170 (7.01%) and Science technology with 2669 (5.90%) articles respectively.

3. Conclusions

A number of research works are being carried out all over the world in this field. A total of 45249 publications were published on big data research during a 2011-2020 and average number of publications per year was 4524.9. The density of the research output during the year 2020 with 9389 publications.

Among 172 countries, USA topped the list with highest share (29.71%) of publications. China ranked second with 26.21% share of publications followed by England 9.12% share of publications. Chinese Academy of Science, China with 1423 (3.14%) publications is the most productive institutions followed by University of California System, USA with 1264 (2.79%) publications. Among source titles, IEEE Access the list with the highest number of publications 1258 (2.78%) followed by Future generation system the international journal of e-science with a share of 491 (1.09%) publications. The maximum number of publications has been published in the English language with 44201 (97.68%) publications, followed by Chinese language with 284 (0.63%) publications.

4. Source of Funding

None.

5. Conflict of Interest

None.

References

1. Kumar RS. Big Data Literature output of Web of Science: A Scientometric Profile. *KIIT J Libr Inf Manag*. 2017;4(2):93–9.
2. Gupta BM. World cataract research: A scientometric analysis of publications output during 2002-11. *L Philos Pract*. 2013;1:18.

3. Kumar RS. Publications trends in nuclear physics: A global perspective. . *Libr Philos Pract*. 2016;p. 1361. Available from: http: //digitalcommons.unl.edu/libphilprac/1361.

4. Kumar RS. Research trends in medical physics: A global perspective. *Libr Philos Pract*. 2016;p. 1362. Available from: http: //digitalcommons.unl.edu/libphilprac/1362.

Author biography

S. Baskaran, Assistant Librarian