Summer 4-2019

Where Does Data Science Research Stand in the 21st Century: Observation from the Standpoint of a Scientometric Analysis

Arindam Sarkar Mr.
Department of Library and Information Science, Jadavpur University, infoarindam83@gmail.com

Ashok Pal Mr.
Institute of Development Studies Kolkata, pal.sunrise.ashok@gmail.com

Where Does Data Science Research Stand in the 21st Century: Observation from the Standpoint of a Scientometric Analysis

Arindam Sarkar[1] **& Ashok Pal**[2]

Abstract: In this scientometric study through the meticulous analysis of year and language wise distribution of publications, document type wise distribution of contributions, year wise citation analysis and country wise productivity in the field of data science research, an effort has been made to delineate a vibrant image of the present condition of data science research. Data have been collected from Scopus database for the purpose of the study. The study reflects that among the total 3793 publications on data science, the highest number of publications i.e. 604, were published in 2018 and lowest number of publications i.e.42, were published in 2002. Among the published documents mostly were in English language i.e. 3654 and then it is followed by German, Chinese, Spanish, French, Japanese, Portuguese, Russian, Polish, and Italian. The predominance of English language in data science research is clearly visible. Journal articles (1509) were the highest in number among different types of publications as nascent information on a subject mainly get reflected in journal articles. Researchers from USA top the list with 1801 publications on data science in the whole world. The year 2011 has received maximum number of citations i.e. 4138. Finally there is a significant positive correlation between time and growth of citation denoting growth trend in the number of citations with the passage of time.

Keywords: Data science, Scientometric analysis, Publication types, Geographical distribution, Pearson correlation

---

[1]PhD Research Scholar of Department of Library and Information Science, Jadavpur University, Kolkata-700032

E-mail: infoarindam83@gmail.com

ORCID: 0000-0002-8728-3378

[2]Assistant Librarian, Institute of Development Studies Kolkata, Salt Lake, Kolkata-700064 & PhD Research Scholar of Department of Library and Information Science, Jadavpur University, Kolkata-700032

E-mail: pal.sunrise.ashok@gmail.com

ORCID: 0000-0002-8428-6864

Introduction

Data science is "the Sexiest Job of the 21st Century", as called by Harvard Business Review in 2012 and thus it became a buzzword. Data science is now often used interchangeably with earlier concepts like business analytics, business intelligence, predictive modeling, and statistics. Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. Data science is the same concept as *data mining* and *big data* and uses the most powerful hardware, the most powerful programming systems, and the most efficient algorithms to solve problems. Data science is a concept to unify statistics, data analysis, machine learning and their related methods in order to understand and analyze actual phenomena with data. It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science. Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that everything about science is changing because of the impact of information technology and the data deluge (Data science, n.d.).

To know the state of the art condition of a research domain scientometric evaluation of that subject is the ultimate solution. This scientometric analysis is a method for analyzing scientific production and is used as a tool for evaluating the quality of scientific production. Nalimov and Mulchenko (1969), of Russia, defined scientometrics as the quantitative methods which deal with the analysis of science viewed as an information process. According to Beck (1978), scientometrics has been defined as the quantitative evaluation and intercomparison of scientific activity, productivity and progress.

Humongous growth in the field of data science research gets its flagrant reflection in the original outputs of the journals in this field. A specific subject area nowhere gets amore nascent information than a suitable journal in that specific field. Journals are always the primary sources of information and are the torchbearers of the growth of literature in different areas of knowledge (Pal & Sarkar, 2018). Along with journal articles, reviews, book chapters, conference papers, letters etc. also carry new ideas in a specific field. In this scientometric study through the panoramic analysis of year and language wise distribution of publications, document type wise distribution of contributions, year wise citation analysis and country wise productivity in the said field of research, a vibrant picture of the present condition of data science research has been portrayed.

Literature Review

Khiste,. Maske and Deshmukh (2018) in their study has focused on big data as reflected in J-gate for the period from 2013–2017. Their result indicated that there were total 8930 articles on this subject domain during 2013 to 2017. United States of America and United Kingdom are the most attentive countries in the area of big data analytics. Liao et al. (2018) focused on the bibliometric analysis and visualization of medicalbig data research They analysed a total of 988

references which were downloaded from the Science Citation Index Expanded and the Social Science Citation Index databases from Web of Science.The GraphPad Prism 5, VOSviewer and CiteSpace software are used for data analysis. Annual trends, the top players in terms of journal and institute levels, the citations and H-index in terms of country level, the keywords distribution, the highly cited papers, the co-authorship status and the most influential journals and authors on that specific domain had been reflected in their study. Zhang et al. (2018) in their research had analysed the question, "How are data analytics involved in policy analysis to create complementary values?" from the perspective of bibliometrics.

Objectives

The objectives of this study are to:

i. To represent the year-wise distribution of publications on data science worldwide.
ii. To trace the language wise distribution of publications on data science.
iii. To show the document wise distribution of publications in this research domain.
iv. To delineate the country wise distribution of publications on the basis of author affiliations.
v. To reflect the chronological distribution of citations as well to represent the correlation between year and number of citations.

Scope and Limitation

The study is restricted within a particular database, i.e. Scopus.com (Scopus, 2019). In this study the documents published on data science from 2001 to 2018 have been collected.

Methodology

To find out the objectives of this study a general scientometric process has been used. As a registered user of Scopus database by using a search string **TITLE (data science) AND PUBYEAR > 2000 AND PUBYEAR < 2019** (Noruzi, 2017). After retrieval, data have been collected and consolidated and then analysed keeping in view the objectives of the study. GunnMap (http://gunnmap.herokuapp.com/) tool has been used to delineate the country wise distribution of publications. Pearson's correlation formula has been used to represent the relationship between year and number of citations through R statistical software. Data have been collected between April 21st to 25th, 2019.

Data Analysis and Findings

In the following few paragraphs retrieved data are presented and analysed through some tables and figures.

- ***Year-wise distribution of publication***

Year wise distribution of publications helps us to identify the research trends regarding the topic data science.

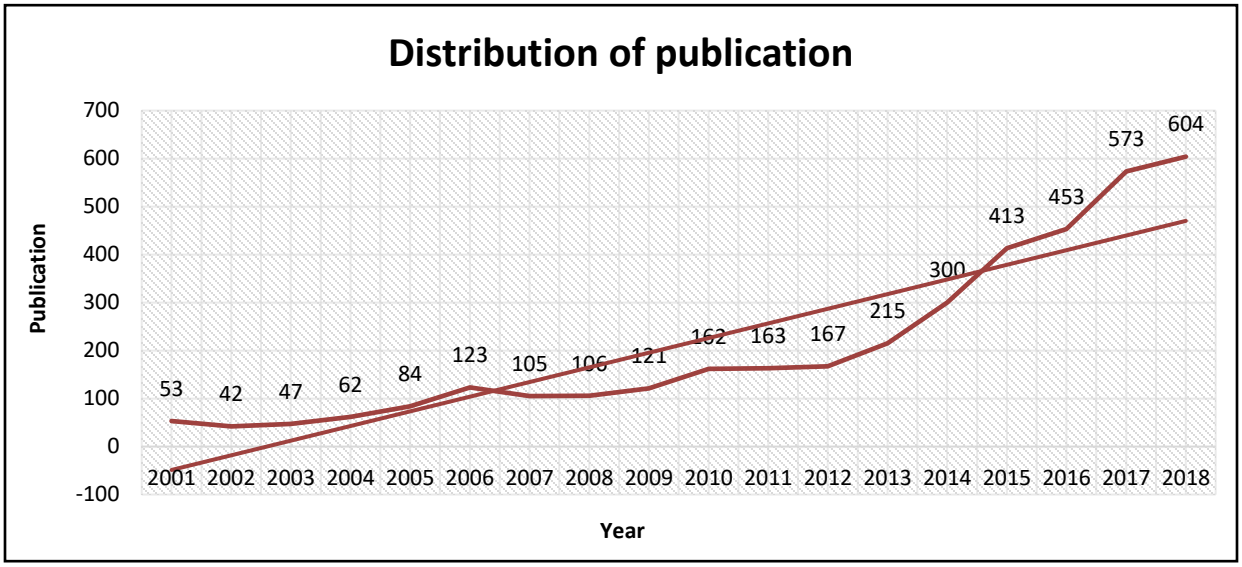

Figure 1: Year-wise distribution of publication

As per this figure, total 3793 number of publications were identified in Scopus within the studied time period. Among the total 3793, the highest number of publications i.e. 604 were published in 2018 and lowest number of publications i.e.42 were published in 2002. The trend line clearly hints at the upward growth of this research domain with every passing day.

- ***Language wise distribution of publication***

Following Table-1 reveals the number of top ten languages of published documents on data science.

Table-1: Language (top ten) wise distribution of published documents

Language	Number	Language	Number
English	3654	Japanese	14
German	31	Portuguese	11
Chinese	21	Russian	7
Spanish	21	Polish	4
French	17	Italian	3

From the above table it can be concluded that, among the published documents mostly were in English (3654) and then it is followed by German, Chinese, Spanish, French, Japanese, Portuguese, Russian, Polish, and Italian. The predominance of English language in data science research is clearly visible.

- ***Document-type wise distribution of publication***

Following table (Table-2) shows the differences in document types on data science.

Table 2: Document type wise distribution of publications

Type of Document	Number	Type of Document	Number
Article	1509	Conference review	111
Conference paper	1162	Book	58
Review	241	Letter	42
Editorial	172	Short survey	33
Book chapter	166	Article in Press	23
Erratum	161	Retracted	1
Note	114	**TOTAL**	**3793**

The table reflects that articles are maximum in number **(**1509). It is followed by conference papers (1162) and reviews (241). It again establishes the supremacy of journal articles in research domain as journal articles always carry the nascent information in any specific field of research.

➢ ***Country wise distribution of publication***

Following figure-2 portrays the distribution of 41 countries (on the basis of the authors' affiliations) those have more than 10 plus number of publications. Researchers from USA top the list with 1801 publications on data science in their works while UK holds the 2nd position with 456 and Germany holds the 3rd rank with 366 publications. Among the African countries only South Africa has some publications on this subject domain.

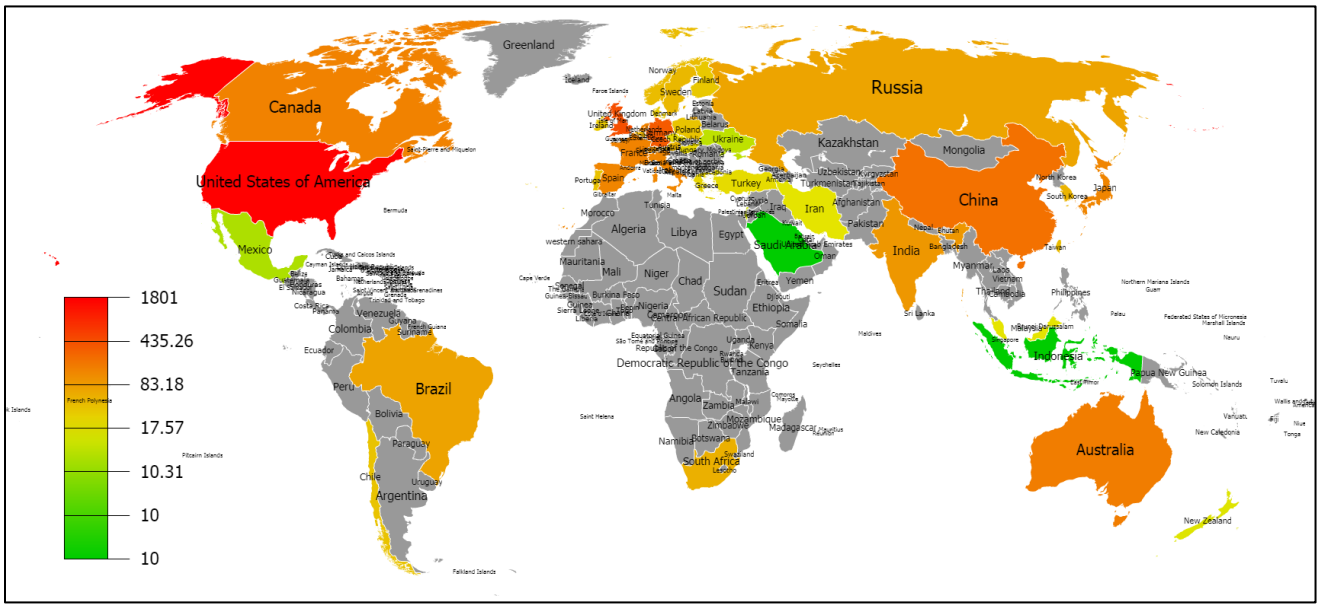

Figure-2: Heat map of countries with active research

(Grey: countries with less than 10 publications and/or no publications; Red: upper limit of publication and Green: lower limit of publication on data science). *Tool*: GunnMap

➢ ***Year wise distribution of citations***

Table-3: Chronological distribution of citations

Year	Number of Citation	Year	Number of Citation
2018	540	2009	1567
2017	1910	2008	1517
2016	3652	2007	2254
2015	3019	2006	2793
2014	2854	2005	2151
2013	3319	2004	1007
2012	2839	2003	1884
2011	4138	2002	414
2010	2201	2001	1409

The table 3 demonstrates that year 2011 has received maximum number of citations i.e. 4138. Using Pearson correlation formula the relationship between year and number of citations can be tested. Pearson correlation is a test used to know the correlation (degree of association) between two variables. In this study correlation has been observed between time (year) and growth of citation (number of citation).

Pearson correlation as per R-Statistical Software:

```
> Year<- c (2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018)
> Citation<- c (1409, 414, 1884, 1007, 2151, 2793, 2254, 1517, 1567, 2201, 4138, 2839, 3319, 2854, 3019, 3652, 1910, 540)
>cor (Year, Citation, method="pearson")
[1] 0.3970787
```

There is a significant positive relationship between time and growth of citation. In this case Pearson's **r =0.3970787**. Following figure also shows the positive correlation between two variables.

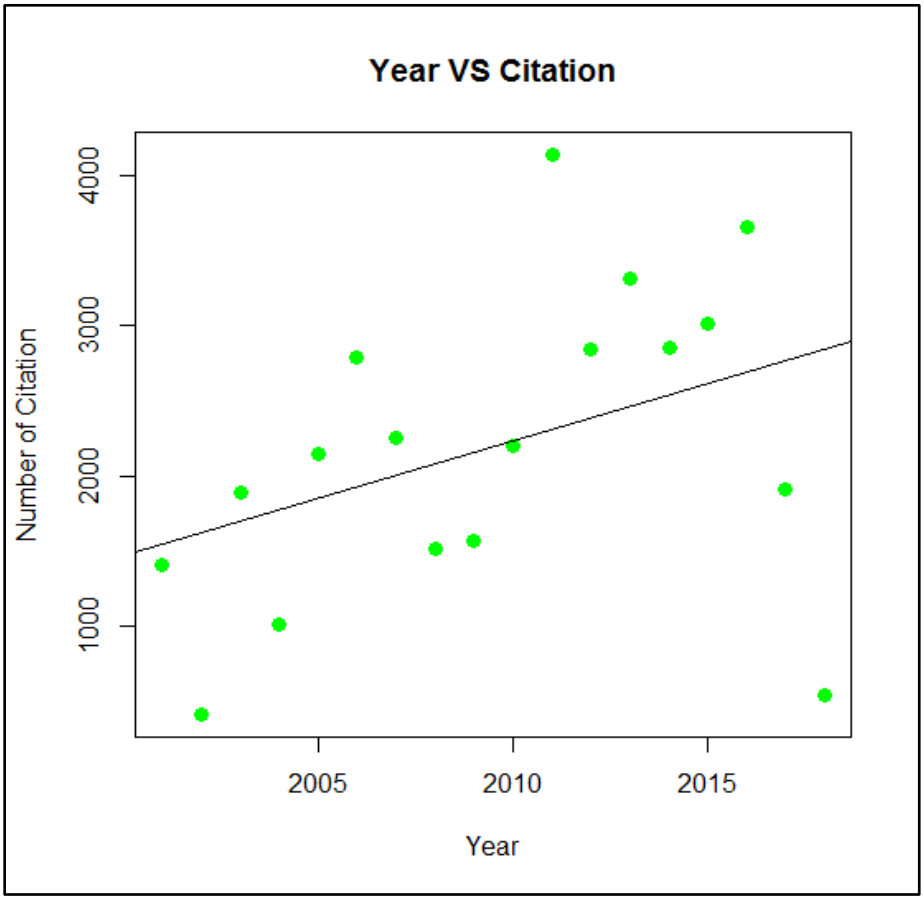

Figure- 3: Relationship between year vs citation (Using R statistical software)

Conclusion

The above study reflects that among the total 3793 publications on data science, the highest number of publications i.e. 604, were published in 2018 and lowest number of publications i.e.42, were published in 2002. Among the published documents mostly were in English language i.e. 3654 and then it is followed by German, Chinese, Spanish, French, Japanese, Portuguese, Russian, Polish, and Italian. The predominance of English language in data science research is clearly visible. Journal articles (1509) were the highest in number among different types of publications as nascent information on a subject mainly get reflected in journal articles. Researchers from USA top the list with 1801 publications on data science in the whole world. The year 2011 has received maximum number of citations i.e. 4138. Finally there is a significant positive correlation between time and growth of citation denoting growth trend in the number of citations with the passage of time.

References

- Beck, M., Dobrov, G., Garfield, E., & De Solla Price, D. (1978). Scientometrics editorial statements. *Scientometrics*, *1*(1), 3-8.
- Data science. (n.d.) In *Wikipedia.* Retrieved April 12, 2019 from https://en.wikipedia.org/wiki/Data_science.
- GunnMap. (n.d.). Retrieved from http://gunnmap.herokuapp.com/
- Khiste, G. P., Maske Dnyaneshwar, B., & Deshmukh, R. K. (2018). Big data output in J-gate during 2013 to 2017: A bibliometrics analysis. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *3*(1), 1252-1257.
- Liao, H., Tang, M., Luo, L., Li, C., Chiclana, F., & Zeng, X. J. (2018). A bibliometric analysis and visualization of medical big data research. *Sustainability*, *10*(1), 166.
- Nalimov, V. V., & Mulchenko, Z. M. (1969). *Scientometrics: The study of science as an information process*. Macoow: Nauka.

- Noruzi, A. (2017). YouTube in scientific research: A bibliometric analysis. *Webology, 14*(1). Retrieved from http://www.webology.org/2017/v14n1/editorial23.pdf

- Pal, A. & Sarkar, A. (2018). Information Systems Research in the 21st century: a bibliometric study. *Library Philosophy and Practice (e-journal). 2155.* Retrieved April 3, 2019 from https://digitalcommons.unl.edu/libphilprac/2155/.

- *Scopus*. (2019). Retrieved April 5, 2019, from https://www.scopus.com

- Zhang, Y., Porter, A., Cunningham, S., Chiavetta, D. & Newman, N. (2018), How is data science involved in policy analysis? A bibliometric perspective, *Portland International Conference on Management of Engineering and Technology*, Hwaii, US.