International Journal of Research in Library Science (IJRLS)

ISSN: 2455-104X

DOI: 10.26761/IJRLS.7.4.2021.1457

Volume 7, Issue 4 (October-December) 2021, 76-84, Paper ID: IJRLS-1457

Received: 12 October. 2021 ; Accepted: 24 October. 2021 ; Published: 27 October. 2021

Research Trend on Data Mining: A Scientometric Study during 2011-2020, based on Web of Science

Dr. S. Baskaran

Assistant Librarian, Marina Campus Library, University of Madras, Chennai - 600 005

baskaranphd@gmail.com

ABSTRACT

The present study analyzing the publication trends on data mining research output based on Web of Science database. During 2011-2020, the database contained 46446 publications were published in the field. The average number of publications per year was 46446 and the highest number of publications 7680 was published in 2020. Relative Growth Rate is decreasing throughout the study period and corresponding Doubling time is increasing. Authors from China have contributed maximum number of publications compared to the other countries and India stood 7th rank in terms of productivity in this period. The most prolific author is Zhang, Y who contributed 245 (0.53%) *publications followed by Liu, Y with 222 (0.48%) publications, Li, J with 214 (0.46%) publications. Chinese Academy of Science, China is the most productive institution with 1248 (2.69%) publications followed by University of California System, USA with 833 (1.79%) publications, China University Mining Technology, China with 797 (1.72%) publications. The most productive source title is IEEE Access the list with the highest number of publications 1188 (25.57%), followed by Expert systems with applications with a share of 794 (1.71%) publications. PLOS one occupy the third position with 656 (1.41%) publications.*

Chinese Academy of Science, China with 1248 (2.69%) publications is the most productive institutions in the field of data mining research followed by.

KEYWORDS: Data mining, Scientometric analysis, Annual growth rate, Relative growth rate and Doubling time.

INTRODUCTION

Data mining occasionally called data or knowledge discovery is the process of analyzing data from different perceptions and transitory it into useful information. Data mining software is one of a number of logical tools for analyzing data. It allows users to analyze data from many different proportions or angles, categorize it, and

summarize the relationships identified. Technically, data mining is the process of finding associations or patterns among several of fields in large relational databases. Companies have used powerful computers to sift through volumes of hypermarket scanner data and analyze market research reports for years to identify end users requirement pattern. Data mining is mainly used to companies with a strong consumer focus - retail, financial, communication, and marketing organizations. Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. The parameters of data also may not be uniform enough to analyze the data. It needs to be integrated from various heterogeneous data sources.

The Scientometric techniques are used to understand the magnitude of the growth of a particular subject/discipline. Especially the trends and growth pattern of publications, contribution of a particular author/ organizations, collaboration pattern, relative growth rate, doubling time and so on. Therefore, the present paper has been undertaken in order to know the growth and development of publications in the field of data mining research as indexed in web of science database.

2. OBJECTIVES FOR THE STUDY

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

- Annual growth Rate, compound growth rate of publications
- Highly prolific authors
- Highly productive countries
- Highly productive institutes
- Most preferred source titles for publications
- Language-wise distribution of publications
- High productive subject areas

3. METHODOLOGY

Web of science provides researchers, faculty and students with quick, powerful access to the world's leading citation databases covering all aspects of sciences, social sciences, arts and humanities with coverage dating back to 1900. The Web of Science database was used for retrieving data on data mining during 2011-2020, using search terms namely 'data mining' in topic filed. A total of 46446 publications were downloaded with no. of bibliographical data, the data were transferred to spread sheet application and analyzed the data as per objectives of the study.

4. DATA ANALYSIS AND INTERPRETATIONS

4.1 Form of publications

Table 1: Form of publications

S. No.	Form of publications	No. of publications	Percentage
1	Journal Articles	41899	90.21
2	Review Articles	2232	4.81
3	Proceeding Papers	1018	2.19

4	Meeting Abstract	472	1.02
5	Editorial Material	420	0.90
6	Early Access	210	0.45
7	Book Chapters	73	0.16
8	Data Papers	41	0.09
9	Book Reviews	31	0.07
10	Letters	23	0.05
11	News Items	16	0.03
12	Retracted Publications	11	0.02
Total		**46446**	**100.00**

The table 1 reveals that the major source of publications covered by web of science databases on animation research is Journal Articles with 41,899 publications (90.21%) followed by Review articles with 2232 publications (4.81%). Proceeding Papers ranks the third position with 1018 publications (2.19%) and Meeting abstract with 472 publications (1.02%) 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.

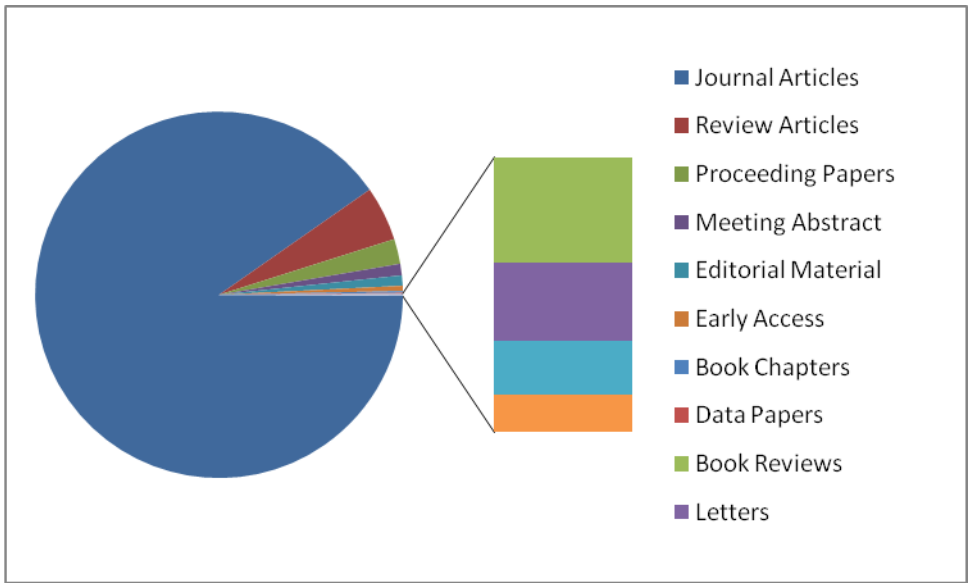

Figure 1: Form of publications

4.2 Growth of publications

Table 2: provides the AGR of the number of documents for period 2012 to 2020.

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

Table 2: AGR of Publications

Year	No. of publications (%)	Cumulative total	Annual growth rate (AGR)
2011	2730 (5.88%)	2730	-
2012	2943 (6.34%)	5673	107.80
2013	3316 (7.14%)	8989	58.45
2014	3560 (7.66%)	12549	39.60
2015	4005 (8.62%)	16554	31.91
2016	4464 (9.61%)	21018	26.97
2017	4896 (10.54%)	25914	23.29
2018	5755 (12.39%)	31669	22.21
2019	7097 (15.28%)	38766	22.41
2020	7680 (16.54%)	46446	19.81

During the period of 2011 to 2020, a total of 46,446 publications were published on data mining research. The highest number of publications is 7680 was published in 2020. The lowest publications of 2730 are published in 2012. The average number of publications published per year was 4644.6. Table 2 shows that there has been a steady growth in research publications on data mining during the study period except in the year 2018.

The table 2 also provides that the annual growth rate of the total publications calculated year wise. AGR reveals that it has decreased from 107.80 in 2012 to 19.81 in 2020. There is a downward trend in the growth rate as seen in the figure 2.

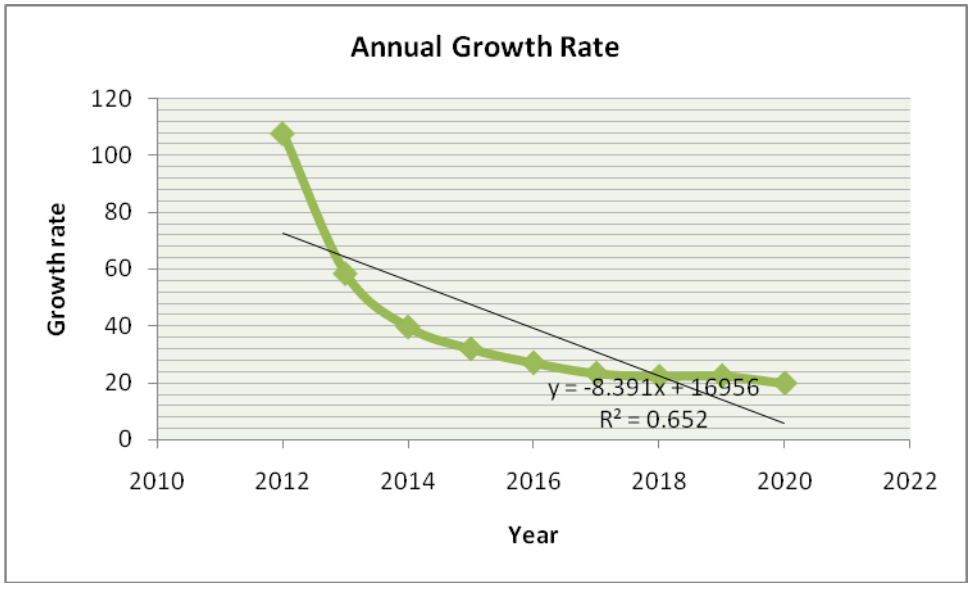

Figure 2: Annual growth rate of publications

4.3 Relative Growth Rate (RGR) and Doubling Time

The Relative Growth Rate (RGR) is the increase in 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 data mining. The mean relative growth rate (R) over the specific period of interval can be calculated from the following equation.

Relative Growth Rate (RGR)

$$1 - 2R = \text{Log } W_2 - \text{Log } W_1 / T_2 - T_1$$

Whereas

1-2 R- mean relative growth rate over the specific period of interval

$\text{Log}_e W_1$ - log of initial number of articles

$\text{Log}_e W_2$ - log of final number of articles after a specific period of interval

T_2-T_1- the unit difference between the initial time and the final time

The year can be taken here as the unit of time.

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	2730	2730	-	7.91	-	-
2012	2943	5673	7.91	8.64	0.73	0.95
2013	3316	8989	8.64	9.10	0.46	1.51
2014	3560	12549	9.10	9.44	0.34	2.04
2015	4005	16554	9.44	9.71	0.27	2.57
2016	4464	21018	9.71	9.95	0.24	2.89
2017	4896	25914	9.95	10.16	0.21	3.3
2018	5755	31669	10.16	10.36	0.20	3.47
2019	7097	38766	10.36	10.57	0.21	3.3
2020	7680	46446	10.57	10.75	0.18	3.85

The year wise RGR is found to be in the range of 0.73 to 0.18. Year wise calculation of RGR reveals that it has decreased from 2012 to 2020 (figure 3). It is found that the RGR is decreasing throughout the study period and corresponding Doubling time is increasing. Average RGR is 0.32 and corresponding Doubling time is 2.65. The highest value of RGR is corresponds to 2012, whereas the lowest value for the years 2020. The Doubling Time has shown a year wise increase from 0.95 to 3.85.

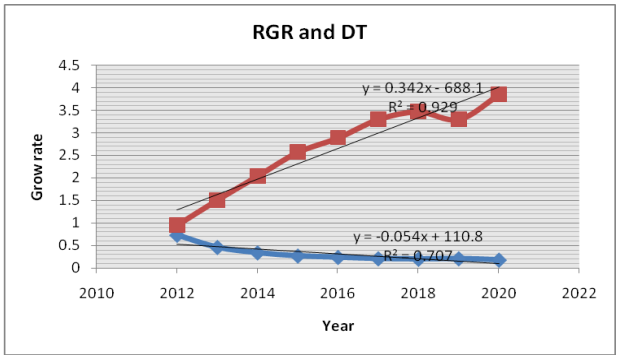

Figure 3: Relative growth rate for research output

4.4 Most prolific authors

Table 4: Most prolific authors

S. No.	Author	No. of publications	Percentage
1	Zhang, Y	245	0.53
2	Liu, Y	222	0.48
3	Li, J	214	0.46
4	Wang, J	203	0.44
5	Li, Y	194	0.42
6	Zhang, J	185	0.40
7	Wang, H	155	0.33
8	Zhang, L	155	0.33
8	Li, X	152	0.33
10	Wang, L	145	0.31
11	Li, L	144	0.31
12	Li, H	131	0.28

The authors having 130 or more publications during 2011-2020 are given in Table 4. Zhang, Y is the most productive author with 245 (0.53%) publications followed by Liu, Y with 222 (0.48%) publications, Li, J with 214 (0.46%) publications, Wang, J with 203 (0.44%) publications, Li, Y with 194 (0.42%) publications, Zhang, J with 185 (0.40%) publications, Wang, H, Zhang, L and Li, X each with 155 (0.33%) publications respectively. And a total of 83,641 authors are contributed entire research output of the period under study.

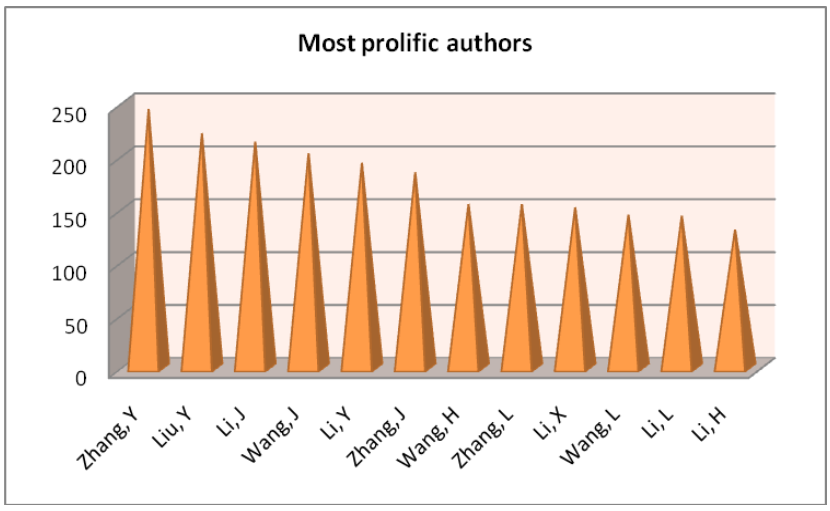

Figure 4: High prolific authors

4.5 Highly productive institutes

Table 5: Highly productive institutes

S. No.	Institutions	Country	No. of Publications
1	Chinese Academy of Science	China	1248 (2.69%)
2	University of California System	USA	833 (1.79%)
3	China University Mining Technology	China	797 (1.72%)
4	Centre National de La Researche Scientifique CNRS	France	760 (1.64%)
5	State University System of Florida	USA	451 (0.97%)
6	Helmholtz Association	Germany	447 (0.96%)
7	University Texas System	USA	402 (0.87%)
8	Indian Institute of Technology System	India	401 (0.86%)
9	University of London	UK	371 (0.80%)
10	Pennsylvania Commonwealth System of Higher Education (PCSHE)	USA	370 (0.80%)

A total of 16,371 institutions 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 1248 (2.69%) publications is the most productive institutions in the field of data mining research followed by University of California System, USA with 833 (1.79%) publications, China University Mining Technology, China with 797 (1.72%) publications, Centre National de La Researche Scientifique CNRS, France with 760 (1.64%) publications, State University System of Florida, USA with 451 (0.97%) publications, Helmholtz Association with 447 (0.96%) publications, University Texas System, USA with 402 (0.86%) publications and Indian Institute of Technology System, India with 401 (0.86%) publications.

4.6 Highly productive countries

Table 6: Highly productive countries

S. No.	Country	Total Publications	S. No.	Country	Total Publications
1	China	12516 (26.95%)	8	Spain	2048 (4.41%)
2	USA	11001 (23.69%)	9	Italy	1795 (3.86%)
3	Australia	2983 (6.42%)	10	France	1744 (3.75%)
4	England	2773 (5.97%)	11	South Korea	1552 (3.34%)
5	Canada	2563 (5.52%)	12	Taiwan	1521 (3.27%)
6	Germany	2471 (5.32%)	13	Iran	1347 (2.90%)
7	India	2176 (4.69%)	14	Brazil	1164 (2.51%)

In all, there were 168 countries involved in the research in data mining; however, China topped the list with highest share (26.95%) of publications. USA ranked second with 23.69% share of publications followed by Australia 6.42% share of publications, England with 5.97% share of publications, Canada with 5.52% share of publications, Germany with 5.32% share of publications, India with 4.69% share of publications, Spain with 4.41% share of publications

and the remaining countries are publishing less than 4% of the research output in this study period. The publication share of highly productive countries (≥1000 publications) on data mining is given in Table 6.

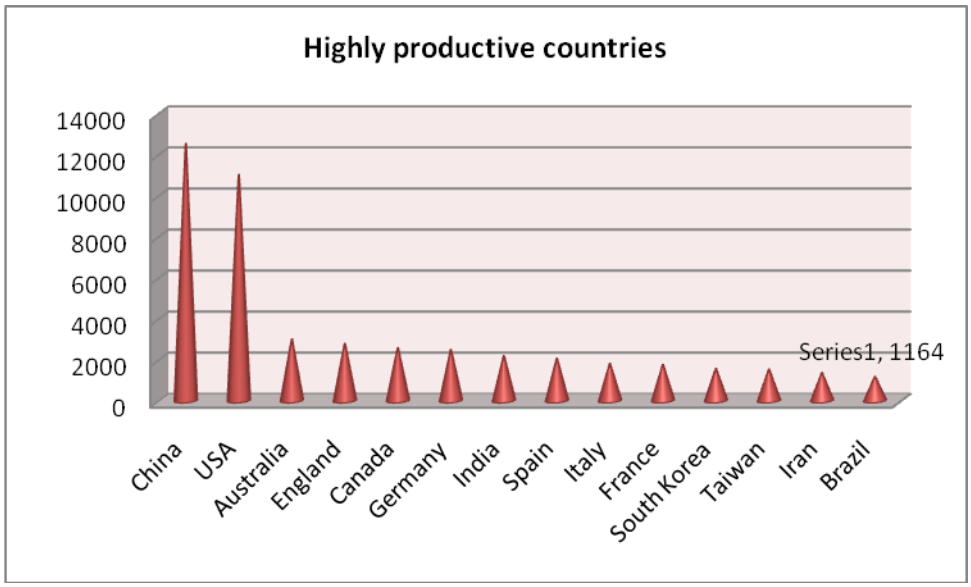

Figure 5: Highly productive countries

4.7 Most preferred source titles

Table 7: Source Title of Publications

S. No.	Source Title	No. of Publications	Percentage
1	IEEE Access	1188	25.57
2	Expert systems with applications	794	1.71
3	PLOS one	656	1.41
4	Knowledge based systems	388	0.84
5	Information sciences	372	0.80
6	IEEE transaction on knowledge and data engineering	359	0.77
7	Knowledge and information systems	310	0.67
8	Science of the total environment	300	0.65
9	Sustainability	291	0.63
10	BMC bioinformatics	271	0.58

The publication share of most productive source titles (≥270 publications) on data mining is given in Table 7. The scientific literature on data mining is spread over 5972 different source journals and conference publications. It reveals that IEEE Access the list with the highest number of publications 1188 (25.57%), followed by Expert systems with applications with a share of 794 (1.71%) publications. PLOS one occupy the third position with 656 (1.41%) publications. The fourth highest source title is Knowledge based systems with 388 (0.84%) publications, Information sciences with 372 (0.80%) publications and IEEE transaction on knowledge and data engineering with 310 (0.67%) publications.

4.8 High productivity subject areas

Table 8: High productivity subject areas

S. No.	Subject	No. of Articles	Percentage
1	Computer science	14659	31.56
2	Engineering	10397	22.39
3	Environmental sciences ecology	4716	10.15
4	Geology	3020	6.50
5	Science and technology	2448	5.27
6	Telecommunications	2282	4.91
7	Biochemistry molecular biology	2196	4.73
8	Chemistry	1881	4.05
9	Operations research management science	1684	3.63
10	Geochemistry geophysics	1668	3.59

The scientific literature on data mining is spread over 193 different subjects. Table 8 shows high productivity subjects which are contributing more than 1600 articles. It is found that Computer science has highest number of articles with 14659 (31.56%) followed by Engineering contributing 10397 (22.39%) articles. Environmental sciences ecology occupies the third position with 4716 (10.15%) articles. The fourth highest articles belonged to the subject Geology with 3020 (6.50%), Science and technology with 2448 (5.27%) and Telecommunications with 2282 (4.91%) articles respectively.

CONCLUSIONS

The present paper attempted to highlight the growth and development of research production on data mining. A total of 46446 publications were published during 2011-2020 and the average number of publications per year was 4644.6. AGR reveals that it has decreased from 107.80 in 2012 to 19.81 in 2020. China topped the list with highest share (26.95%) of publications. USA ranked second with 23.69% share of publications followed by Australia 6.42% share of publications. There were 168 countries involved in the research in data mining. Chinese Academy of Science, China with 1248 (2.69%) publications is the most productive institutions followed by University of California System, USA with 833 (1.79%) The scientific literature on data mining is spread over 5972 different source journals and conference publications. Among subjects, Computer science has highest number of articles with 14659 (31.56%) followed by Engineering contributing 10397 (22.39%) articles. Environmental sciences ecology occupies the third position with 4716 (10.15%) articles.

REFERENCES

[1] Santha kumar R. Publications Trends in Nuclear Physics: A Global Perspective (2016). *Library Philosophy and Practice (e-journal).* Paper 1361.http://digitalcommons.unl.edu/libphilprac/1361

[2] Bala, A. and Gupta, B.M. Mapping of Indian neuroscience research: A Scientometric analysis of India's research output, 1990-2008. *Scientometrics*, 2010, 58(1), 35-41

[3] Kademani, B.S et al. World literature on thorium research: A scientometric study based on Science Citation Index, *Scientometrics*, 2006, 29(2): 347-64.

[4] Santha kumar R. Research Trends in Medical Physics: A Global Perspective" (2016). *Library Philosophy and Practice (e-journal).* Paper 1362. http://digitalcommons.unl.edu/libphilprac/1362

[5] www.science.purdue.edu