Winter 10-5-2020

Mapping of Data Mining Research Productivity in India: A Scientometric Analysis

SUMATHI Meyyar
Bharathidasan University, sagusumathi@gmail.com

RANGANATHAN CHANDRAKASAN
Bharathidasan University, cranganathan72@gmail.com

MAPPING OF DATA MINING RESEARCH PRODUCTIVITY IN INDIA: A SCIENTOMETRIC ANALYSIS

By

***M.Sumathi and ** Dr. C. Ranganathan**

*Research Scholar, Department of Library and Information Science, Bharathidasan University, Tiruchirappalli-24, TamilNadu, India. e-mail: sagusumathi@gmail.com.,

**Associate Professor, Department of Library and Information Science, Bharathidasan University, Tiruchirappalli-24, TamilNadu, India. e-mail: cranganathan72@gmail.com., cranganathan@bdu.ac.in

ABSTRACT

This study analyses the Indian Scientists contributions of research papers related to the topic in Data Mining was undertaken from Web of Science Databases has been used to retrieve the data for 22 years (1999-2020) by the searching the keyword "Data Mining". The study reveals that, most of the researchers preferred to publish their research results in journals; as such 88.59% of articles were published in journals. More numbers of articles were published in the year 2019. The authorship trend shows that, out of total 1096 literature published, 95.53 % of the publication published under the joint author. It is observed that author productivity is not in agreement with Lotka's law, but productivity distribution data partially fits the law when the value of Chi-square to 25212.62. Further this study also identified to analyses source wise. Degree of collaboration, Areas of research concentration, word frequency, Geographical distribution of the literature and citation analysis is also noted

Keywords: Data Mining, Scientometrics, Author Productivity, Bradford's law, Citation, India

0. INTRODUCTION

"We are living in the information age" is a popular saying; however, we are actually living in the data age. Terabytes or petabytes1 of data pour into our computer networks, the World Wide Web (WWW), and various data storage devices every day from business, society, science and engineering, medicine, and almost every other aspect of daily life. This explosive growth of available data volume is a result of the computerization of our society and the fast development of powerful data collection and storage tools. Communities and social media have become increasingly important data sources, producing digital pictures and videos, blogs, Web

communities, and various kinds of social networks. The list of sources that generate huge amounts of data is endless.

This explosively growing, widely available, and gigantic body of data makes our time truly the data age. Powerful and versatile tools are badly needed to automatically uncover valuable information from the tremendous amounts of data and to transform such data into organized knowledge. This necessity has led to the birth of data mining. The field is young, dynamic, and promising. Data mining has and will continue to make great strides in our journey from the data age toward the coming information age. "Data mining is a process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems".

The major focus of the study is to apply the Scientometric analysis with a view to analyze the evaluation and productivity of growth and development of research output in Data Mining in India. This study related to authors and their productivity; collaborative patterns and other aspects is important and useful to understand the mechanism underlying the growth of knowledge of a discipline. This study also to analyses the evaluation growth and development and of Data Mining research output interns of its content and coverage growth rates, Source wise, author productivity, authorship Pattern, Degree of collaboration, Lotka's law , Broad ford's law, geographical distributions and citation analysis is also noted.

1. OBJECTIVES OF THE STUDY

1. To identify and analysis the pattern of distribution of Data Mining research output in India.
2. To identify the year wise distribution of Publications
3. To study the Document wise distribution of Publications
4. To study the Ranking of Authors based on Publications and citations
5. To identify the nature of Authorship pattern and determine the degree of collaboration.
6. To identify the proportion of single and multi-authored papers of Data Mining research output.
7. To identify the Journal wise distribution of Publications
8. To study the Institution wise distribution of Publications
9. To identify the Country wise distribution of Publications

2 HYPOTHESES

The following hypotheses have been formulated with a view to test the above framed objectives.

1. The implication of Lotka's law related with scientific productivity of authors in Data Mining.
2. To test the Bradford's Law of Scattering in Data Mining research output in India

3. METHODOLOGY

The present study aims at analyzing the research output of Researchers in the field of Data Mining. The growth rates of output in terms of research productivity are analyzed from 1999 to 2020. The data has analyzed and classified into His cite software it is also analytical in nature in strengthening the empirical validity due to application of suitable statistical tools.

3.1 DATA DOWNLOADING

Data was downloading on 28th February 2020 for a period of 22 years (1999-2022) from the Web of Science. Web of Science has wide acceptance and is frequency used standard database of choice for undertaking Scientometric studies. It was necessary to search strategies because of the inherent limitations of the number of keywords which can be accommodated in a single strategy in WoS. The researcher has used the search string "Data Mining" for getting data from the Web of Science database which includes Science Citation Index (SCI). The researcher has downloaded the bibliographical data in the form of notepad files. Overall data retrieved by the researcher are 1109 records and eliminated 13 duplicate records hence, the refined data consists only 1096 records taken for analyzing the present study. The data has analyzed and classified into Histcite software. Finally, the unique data are rearranged in MS –EXCEL format to eliminate duplication from the downloaded data and to analyze the scattering of research in different dimensions

3.2 APPLICATION OF METRICS AND BASIC LAWS OF BIBLIOMETRICS

The following Metrics and bibliometrics law have been used in the analysis of data

3.2.1. Degree of Collaboration Co-Efficient

In order to identify the degree of collaboration, the research or has adopted K. Subramanyam's formula. The formula is C = Nm/ (Nm+Ns)

Where, C = Degree of collaboration in a discipline

Nm = Number of multiple authored papers

Ns = Number of the single authored papers

3.2.2 Lotka's Law of Author's Productivity

Lotka's law of author productivity explains number of authors contributed 'n' number of paper. Potter identified the Lotka's fraction $1/n^a$ – 4.65 on the basis of Euler – Maclaurin formula of summation. This model is applied in the present study.The sum was used as deviser for 1/n 4.65 to determine the proportion of the total number of authors expected to produce 'n' paper (in the case of present study n=1, 2, 3, 4… 10), the following formula was used to find the proportions.

$$S = \sum_{n=1}^{10} 1/n\ 4.65$$

For present study S is the sum of Lotka's modified rations for the value a= 4.65.

The formula

$$An = 1/n\ 4.65\ T/S\ (n = 1, 2, 3 \ldots 10)$$

Where T is total number of authors in the sampling and 'An' is the total number of expected authors producing 'n' papers.

The Lotka's law also tested with the application of scientific productivity chi-square model in relation to a number of authors who contributed 'n' number of publication.

It can be expressed by the equation an = a1/n2, n =1,2,3

In other words, for every 100 authors making one contribution each, there would be 25 others contributing two articles each ($100/2^2$= 25) about 11 contributing three articles each 100 $/3^3$ = 11.1, and so on.

Where 'an' is the numbers of authors contributing 'n' papers each; and al is the number of authors contributing each one paper.

The chi-square can be computed as (F-p) 2/p.

F = observed number of authors with 'n' publications

P = expected number of authors.

3.2.3 Bradford's Law

The law is mathematically expressed as

$$F(x) = a+b\log X$$

Where, F(x) is the cumulative number of references contained in first 'x' most productive journals.' a' and 'b' are constant

4. ANALYSIS AND DISCUSSION

4.1 Growth of Publications

To analysis the year wise publication of research on Data Mining, the data has been presented from the below table-1, we could clearly see that during the period 1999 – 2020 a total of 1096 publications were published. The highest publication is 202 in 2019 with 186 Global Citation Scores followed by 143 papers in 2018 with 603 Global Citation Score and 123 papers in 2016 with 1159 Global Citation Scores. The lowest publication is 3 in 1999 and 2000 with 102 and 19 Global Citation Scores. It shows that even minimum numbers of records were scored higher global citations. The study also reveals all these 1096 publications have 38298 cited references it shows that there is a healthy trend in citing reference is found among the global Scientists belongs to "Data Mining".

Table 1: Shows Year wise Distribution of Citation Score

S.No	Publication Year	Publications	Percent	TLCS	TGCS
1	1999	3	0.3	0	102
2	2000	3	0.3	3	19
3	2001	8	0.7	1	17
4	2002	7	0.6	20	1280
5	2003	13	1.2	6	371
6	2004	22	2.0	6	2121
7	2005	19	1.7	12	292
8	2006	21	1.9	15	542
9	2007	24	2.2	20	731
10	2008	23	2.1	17	396
11	2009	26	2.3	12	819
12	2010	24	2.2	19	498
13	2011	40	3.6	12	866
14	2012	40	3.6	18	504
15	2013	71	6.4	21	1070
16	2014	66	6.0	22	1142
17	2015	81	7.3	38	1003
18	2016	123	11.1	25	1159
19	2017	113	10.2	14	711
20	2018	143	12.9	15	603
21	2019	202	18.2	4	186
22	2020	24	2.2	0	2
Total		**1096**	**99**	**300**	**14434**

4.2 Source Wise Distribution of Publications

A study of data in table-2 indicates the source wise distribution of research output in Data Mining during the period of twenty two years from 1999 to 2020. Out of various sources of publications in Data Mining, journal articles that appeared in the journals have shown a predominant contribution (88.59%) with Global citation score is 12615 and this source occupies the first position. The source of review comes second in order (5.84 %) of sharing total research output in Data Mining" during the period of analysis. The source of Proceedings Paper comes in the third position (3.56%) with respect to total output in "Data Mining" research during the study period.

Table 2: Shows Source wise distribution of Publications

S.No	Document Type	Publication	%	TLCS	TGCS
1	Article	971	88.59	268	12615
2	Review	64	5.84	16	1292
3	Article; Proceedings Paper	39	3.56	14	458
4	Article; Early Access	6	0.55	0	0
5	Editorial Material	5	0.46	0	38
6	Meeting Abstract	2	0.18	0	2
7	Article; Book Chapter	2	0.18	0	14
8	Correction	2	0.18	0	0
9	Letter	2	0.18	0	2
10	Review; Early Access	2	0.18	0	0
11	Article; Retracted Publication	1	0.09	2	13
	Total	**1096**	**100**	**300**	**14434**

4.3 Ranking of Authors Productivity Based on Publications

Table- 3 indicates ranking of authors by number of publications. Authors "Pal SK" published highest number of articles for the study period with 22 records, consecutive authors "Maji P" and Samantaray SR" are published next highest number of articles for the study period with 14 records. **"**Pal SK" having highest Global Citation Scores of 2085 with just 22 publications followed by "Mitra P" is having Global Citation Score of 1537 with 11 publications, while Balamurugan SAA having lowest Global Citation Score of 13 with just 8 publications. Thus the most-cited authors are distinguished from the most-published ones.

Table- 3 shows Ranking of Prolific Authors

S. No	Author	Articles	%	TLCS	TLCS/t	TLCSx	TGCS	TGCS/t	TLCR	TLCSb	TLCSe
1	Pal SK	22	2.0	45	3.02	22	2085	122.43	13	6	4
2	Maji P	14	1.3	21	1.64	4	438	36.75	20	3	1
3	Samantaray SR	14	1.3	19	3.06	3	312	56.00	16	6	
4	Kumar S	12	1.1	0	0.00	0	116	19.30	3	0	
5	Tiwari MK	12	1.1	4	0.33	2	432	41.74	4	1	0
6	Ghosh A	11	1.0	7	0.71	5	202	25.06	6	0	0
7	Mitra P	11	1.0	27	1.48	19	1537	82.59	2	3	2
8	Das AK	10	0.9	5	0.85	2	134	28.59	6	1	
9	Sharma A	10	0.9	0	0.00	0	20	4.78	0	0	
10	Biswas SK	9	0.8	3	0.83	0	42	10.52	5	0	
11	Dehuri S	9	0.8	13	1.29	9	245	23.65	7	1	0
12	Kumar R	9	0.8	2	0.23	1	209	33.88	2	1	
13	Mukhopadhyay A	9	0.8	6	0.86	4	378	51.55	6	2	1
14	Singh A	9	0.8	1	0.20	0	52	8.13	3	0	
15	Singh S	9	0.8	3	0.62	3	108	18.94	1	0	
16	Balamurugan SAA	8	0.7	1	0.14	0	13	2.29	0	0	
17	Bandyopadhyay S	8	0.7	6	0.86	4	426	53.47	6	2	2
18	Gupta A	8	0.7	0	0.00	0	78	18.34	3	0	
19	Gupta S	8	0.7	0	0.00	0	39	5.76	0	0	
20	Kumar M	8	0.7	1	0.14	0	56	11.01	0	1	
21	Maulik U	8	0.7	6	0.86	4	361	52.68	7	2	1
22	Mitra S	8	0.7	17	0.94	10	798	63.21	2	3	3
23	Sastry PS	8	0.7	27	2.06	2	287	21.25	22	9	4
24	Sharma S	8	0.7	1	0.20	0	39	6.97	1	0	
25	Chakraborty M	7	0.6	3	0.92	0	25	7.25	3		
26	Ghosh S	7	0.6	0	0.00	0	64	11.29	1	0	
27	Jacob SG	7	0.6	0	0.00	0	30	3.88	0	0	
28	Jena MK	7	0.6	7	1.32	0	81	16.54	12	4	
29	Kumar N	7	0.6	0	0.00	0	19	4.61	1	0	
30	Laxman S	7	0.6	19	1.41	1	254	18.84	16	6	3
	Total	**284**	**25.3**	**244**	**23.97**	**95**	**8880**	**861.3**	**168**	**51**	**21**

4.4 Single Vs Multiple Authored Research Output and Degree of Collaboration

It is observed that the single version multi author research output during the period 1999 to 2020. At the overall level, the single author contributed papers constitute 4 percent of the total publications; whereas the remaining majority (96%) of the papers is contributed by multi-authorship. In order to determine the collaboration in quantitative terms, the formula suggested by K. Subramanyam was tested.

Table-4: Shows Single Vs Multiple Authored Research Output Degree of Collaboration

Year	Single Author		Multiple Authors		Total	%	Degree of Collaboration	Mean in Degree of Collaboration
	No of Output	%	No of Output	%				
1999	1	2.04	2	0.19	3	0.27	0.67	0.90
2000	0	0.00	3	0.29	3	0.27	1.00	
2001	2	4.08	6	0.57	8	0.73	0.75	
2002	1	2.04	6	0.57	7	0.64	0.86	
2003	0	0.00	13	1.24	13	1.19	1.00	
2004	1	2.04	21	2.01	22	2.01	0.95	
2005	2	4.08	17	1.62	19	1.73	0.89	
2006	1	2.04	20	1.91	21	1.92	0.95	
2007	1	2.04	23	2.20	24	2.19	0.96	
2008	3	6.12	20	1.91	23	2.10	0.87	
2009	1	2.04	25	2.39	26	2.37	0.96	
2010	2	4.08	22	2.10	24	2.19	0.92	0.95
2011	3	6.12	37	3.53	40	3.65	0.93	
2012	0	0.00	40	3.82	40	3.65	1.00	
2013	5	10.20	66	6.30	71	6.48	0.93	
2014	3	6.12	63	6.02	66	6.02	0.95	
2015	5	10.20	76	7.26	81	7.39	0.94	
2016	2	4.08	121	11.56	123	11.22	0.98	

2017	4	8.16	109	10.41	113	10.31	0.96	
2018	7	14.29	136	12.99	143	13.05	0.95	
2019	4	8.16	198	18.91	202	18.43	0.98	
2020	1	2.04	23	2.20	24	2.19	0.96	
Total	**49**	**4.47**	**1047**	**95.53**	**1096**	**100**	**0.93**	**0.93**

It is inferred from the above table -4 that at the aggregate level, the degree of collaboration is of 0.74 during the study period 1999 to 2020 i.e., that is out of total 1096 literature published, 95.53% of them or published under the joint author of publications in "data mining" research output. The period wise analysis indicates that its level is somewhat less in the first period [1999-2009: 0.90] and it has shown. An increasing trend during the period [2010-2020: 0.95]. This brings out clearly the high level of prevalence of collaborative research in "Data Mining". Based on this study, the result of the degree of collaboration **C=0.93** i.e., 93 percent of collaboration authors articles published during the study periods.

4.5 Lotka's Law of Author Productivity

The Lotka's law of author productivity is tested with the applications of scientific productivity Chi-square model, and it is applied in relation to number of authors contributing to the number of publications. It is relevant to analyze the implications of Lotka's law in relation to author productivity on Data Mining. It explains that number of authors making 'n' contribution is about $1/n^2$ of those making a single contribution and the proportion of contribution that makes a single contribution is about 60 percent. In this study, Data Mining Scientists productivity is examined. At the first observation, the analyzed data invalidate the Lotka's findings that the proportion of all contributions that make a single contribution is less than 60 percent.

Further, Lotka's chi-square model confirms the source trend. It explains that the calculated χ^2 value is 25212.62 which is less than its tabulated value at 5 percent level of significance. Thus, the present analysis clearly invalidates the Lotka's findings. (Hence, the first hypothesis is not proved (the implication of Lotka's law related with author productivity in Data Mining)

Table – 5: Lotka's Law of Author Productivity- Chi- Square Model

No. of authors	Observed Number of authors with 'n' or (an) or (f)	Observed percentage of authors 100 x an/a1	Expected number of authors ($an=an/n^2$)or (p)	Expected percentage of authors	(F-P)^2/P
1	49	100.00	49	100.00	0
2	421	859.18	105.25	25.00	947.25
3	292	595.92	32.44444	11.11	2076.444
4	158	322.45	9.875	6.25	2221.875
5	75	153.06	3	4.00	1728
6	41	83.67	1.138889	2.78	1395.139
7	19	38.78	0.387755	2.04	893.3878
8	13	26.53	0.203125	1.56	806.2031
9	12	24.49	0.148148	1.23	948.1481
10	5	10.20	0.05	1.00	490.05
11	8	16.33	0.066116	0.83	952.0661
12	2	4.08	0.013889	0.69	284.0139
13	2	4.08	0.011834	0.59	334.0118
14	1	2.04	0.005102	0.51	194.0051
15	1	2.04	0.004444	0.44	223.0044
16	2	4.08	0.007813	0.39	508.0078
17	2	4.08	0.00692	0.35	574.0069
18	2	4.08	0.006173	0.31	644.0062
24	1	2.04	0.001736	0.17	574.0017
29	1	2.04	0.001189	0.12	839.0012
50	1	2.04	0.0004	0.04	2498
78	1	2.04	0.000164	0.02	6082
				χ^2	**25212.62**

4.6 Analysis the Ranking List of Journals and Their Published Articles

The study found that the total research output of the Data Mining for the study period (1999 – 2020) published in 470 journals. Table- 6 indicates the major portion of the research productivity (34.8%) covered by 30 journals that is coinciding with the theory of Bradford's Law of scattering of journals in research productivity. Top thirty Journals produced mostly 34.8 % of the research output. The journal "Expert Systems with Applications" topped with 33 publications

with the Global Citation Score of 700, next "Cluster Computing-The Journal of Networks Software Tools and Applications" has 28 publications with the Global Citation Score of 31 and "Sadhana-Academy Proceedings in Engineering Sciences" with 23 publications with the Global Citation Score of 194 respectively. The "Expert Systems with Applications" has scored the highest Global Citation Score of 700 with 33 publications out of top thirty journals while "Arabian Journal for Science and Engineering" has scored a lowest Global Citation Score of 4 with just 6 records.

Table-6 Distribution of Ranking list of Journals and their Published Articles

S. No	Journal	Articles	%	TLCS	TLCS/t	TGCS	TGCS/t	TLCR
1	Expert Systems With Applications	33	3.0	27	4.35	700	102.21	8
2	Cluster Computing-The Journal Of Networks Software Tools And Applications	28	2.5	4	0.82	31	11.22	3
3	Sadhana-Academy Proceedings in Engineering Sciences	23	2.1	9	1.17	194	16.85	9
4	Applied Soft Computing	21	1.9	15	1.49	521	58.57	11
5	International Arab Journal of Information Technology	19	1.7	4	0.55	45	7.86	8
6	Journal of Medical Systems	17	1.5	0	0.00	134	25.63	6
7	Journal of Medical Imaging and Health Informatics	16	1.4	3	0.60	51	9.09	8
8	Journal Of Intelligent & Fuzzy Systems	14	1.3	1	0.25	30	8.10	4
9	Neural Computing & Applications	14	1.3	6	0.81	226	38.85	3
10	Information Sciences	13	1.2	2	0.24	186	24.60	11
11	Knowledge and Information Systems	13	1.2	6	0.64	83	9.77	16
12	IEEE Transactions on Knowledge And Data Engineering	12	1.1	26	2.08	406	33.42	8
13	Knowledge-Based Systems	12	1.1	5	0.66	183	29.85	5
14	Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery	12	1.1	2	0.29	107	17.57	4
15	Biomedical Research-India	11	1.0	1	0.25	12	2.80	2
16	International Journal of Data Mining and Bioinformatics	11	1.0	6	1.73	20	4.34	9
17	Pattern Recognition Letters	11	1.0	5	0.68	281	23.09	0

18	Applied Intelligence	10	0.9	1	0.06	76	9.12	4
19	Intelligent Data Analysis	10	0.9	3	0.52	63	6.48	4
20	Soft Computing	10	0.9	2	0.45	17	3.76	7
21	Computers & Electrical Engineering	9	0.8	2	0.50	53	12.28	6
22	IEEE Access	9	0.8	0	0.00	6	2.00	2
23	Journal of Scientific & Industrial Research	9	0.8	1	0.08	16	2.58	1
24	Current Science	8	0.7	1	0.07	31	3.36	2
25	Journal of Ambient Intelligence And Humanized Computing	8	0.7	0	0.00	44	14.67	1
26	Wireless Personal Communications	8	0.7	0	0.00	26	5.83	3
27	International Journal of Uncertainty Fuzziness and Knowledge-Based Systems	7	0.6	2	0.23	16	2.06	1
28	Neuro Computing	7	0.6	1	0.17	150	27.43	4
29	Arabian Journal for Science And Engineering	6	0.5	0	0.00	4	0.98	2
30	Gene	6	0.5	0	0.00	50	5.90	1
	Total	**387**	**34.8**	**135**	**18.69**	**3762**	**520.27**	**153**

4.7 Bradford's Law Distribution

The Bradford law was formulated in the year 1948. It examines essentially that a group of journals are arranged in an order of decreasing productivity. It means the journals that yield that most relevant article coming first and the most unproductive in the last. Table-7 shows clearly that the ranking list of journals contributed by Data Mining scientists in an order of decreasing productivity.

Table No.-7 indicates that the first twenty eight journals covered more than one third of the total articles published. The next hundred and forty seven journals covered another one third of the articles. The remaining 291 journals covered the last one third of the published articles. According to Bradford's distribution the relationship between the zone is 1: a: a^2, while the relationship in each zone of the present study is 28:147:291 which does not fit into Bradford's distribution. This shows that core contributions are given by a very few journals, i.e., less than Bradford formulated and the final zone contains a very large number of journals, i.e, much more than the Bradford formula.

Table-7 Showing Ranking Journals according to Bradford's Law

S.No	No of Journals	No of Contribution	Total Number of Contribution	Cumulative Total
1	1	33	33	33
2	1	28	28	61
3	1	23	23	84
4	1	21	21	105
5	1	19	19	124
6	1	17	17	141
7	1	16	16	157
8	2	14	28	185
9	2	13	26	211
10	3	12	36	247
11	3	11	33	280
12	3	10	30	310
13	3	9	27	337
14	3	8	24	361
15	**2**	**7**	**14**	**375**
16	6	6	36	411
17	12	5	60	471
18	17	4	68	539
19	42	3	126	665
20	**70**	**2**	**140**	**805**
21	**291**	**1**	**291**	**1096**

4.8 Institution Wise Distribution of Publications

In general, institutions which are specifically meant for research activities would contribute a greater level of research publications and it is not up to the mark of desired level of expectations in other institutions. The table- 8 analysis indicates Institution-wise research productivity. It is noted that 1203 institutions were contributed 1096 of the total research productivity. It indicates that the major portion of the research productivity (42.9%) contributed by top 25 institutions. It is noted that Indian Inst Technology contributed the highest number of research publications (86) with Global Citation Score 1422. Indian Statistical Institute terms second in order 23 publication of the total Global Citation score 3014.

Table-8 Institution wise Distribution of Publications

S. No	Institution	Publication	%	TLCS	TGCS
1	Indian Institute of Technology	86	7.8	35	1422
2	Indian Statistical Institute	53	4.8	58	3014
3	Anna University	49	4.4	7	208
4	Indian Institute of Science	34	3.1	37	723
5	National Institute of Technology	30	2.7	3	108
6	Jadavpur University	19	1.7	10	511
7	Thapar University	19	1.7	1	124
8	VIT University	18	1.6	2	37
9	Thiagarajar College of Engineering	14	1.3	3	125
10	Sathyabama University	12	1.1	1	10
11	University of Hyderabad	12	1.1	1	83
12	Birla Institute of Technology	11	1.0	4	51
13	Indian Inst Technology Kharagpur	11	1.0	1	17
14	Visvesvaraya National Institute of Technology	11	1.0	6	11
15	Kongu Engineering College	10	0.9	1	19
16	SASTRA University	10	0.9	1	61
17	Bharathiar University	9	0.8	0	1
18	CSIR	9	0.8	2	73
19	Indian Inst Technology Bhubaneswar	9	0.8	12	182
20	PSG College of Technology	9	0.8	0	38
21	University of Delhi	9	0.8	1	81
22	Jamia Millia Islamia	8	0.7	1	63
23	Jawaharlal Nehru University	8	0.7	0	98
24	Sri Krishna College of Engineering& Technology	8	0.7	1	13
25	University of Kalyani	8	0.7	6	347
		476	**42.9**	**194**	**7420**

4.9 Country – Wise Collaborative Distribution of Publications

The study of Country wise distribution of a number of research output is an important factor in highlighting the research and development in any discipline of science. In this context, the analysis of performance of Indian Data Mining scientists is quite obvious with a view to

reflect their achievements in attracting the attention of foreigners in terms of published research articles in the journals of various countries.

Table -9: Country – Wise Collaborative Distribution of Publications

S.No	Country	Publication	%	TLCS	TGCS
1	India	1104	99.5	300	14307
2	USA	90	8.1	42	3166
3	Peoples R China	31	2.8	5	670
4	UK	28	2.5	5	723
5	Australia	21	1.9	2	264
6	South Korea	21	1.9	10	391
7	Canada	16	1.4	9	526
8	Japan	14	1.3	3	330
9	Iran	13	1.2	2	105
10	Italy	12	1.1	2	218
11	Singapore	12	1.1	2	200
12	Germany	11	1.0	1	210
13	Malaysia	11	1.0	4	265
14	Vietnam	9	0.8	2	211
15	France	8	0.7	1	161
16	Saudi Arabia	8	0.7	1	27
17	Brazil	6	0.5	0	24
18	Egypt	6	0.5	1	53
19	Netherlands	6	0.5	0	303
20	Spain	6	0.5	0	17
21	Mexico	4	0.4	6	257
22	Norway	4	0.4	1	174
23	Taiwan	4	0.4	0	33
24	Unknown	4	0.4	0	62
25	Finland	3	0.3	0	48

The above table-9 shows that among the country wise distribution of “Data Mining” covered by the study tops India has published 1104 (99.5 %) publications with global citation score 14307 followed by USA has published 90 (8.1%), Peoples R China with 31 (2.8 %), research publications respectively. First place goes to India having total Global Citation Score of 14307 with 1104 publications. USA has secured second rank in terms of GCS with 3166 but with only 90 publications.

4.10 Documentation of Word Frequency in the Publications

Publications convey precisely the thought contents of the papers. The potency of information concentrated on the titles of the papers is more than the rest of the section of the papers. Therefore, if a word occurs more frequently than expected it to occur, then it reflects the emphasis given by the authors about the research field of their interest. The important words called 'Key Word' are one of the best indicators to understand and grasp instantaneously the thought content of the papers, methodologies used and areas of research addressed to the high frequency keywords were **"Data"** is topped with 373 publications with the Global Citation Score of 6144, next "**Mining"** has scored the highest Global Citation Score of 5703 with 346 and followed by word "**Using**" has scored the Global Citation Score of 3546 with 281 publications respectively.

Table- 10 showing Word Frequency in the Publications

S.No	Word	Publication	Percent	TLCS	TGCS
1	Data	373	33.6	103	6144
2	Mining	346	31.2	124	5703
3	Using	281	25.3	57	3546
4	Based	234	21.1	58	2500
5	Approach	128	11.5	32	771
6	Classification	114	10.3	21	724
7	Algorithm	109	9.8	27	1004
8	Clustering	99	8.9	17	954
9	Analysis	85	7.7	16	703
10	Feature	76	6.9	22	1534
11	Selection	70	6.3	27	1303
12	Detection	67	6.0	15	745
13	Fuzzy	67	6.0	36	1102
14	Prediction	63	5.7	17	480
15	System	63	5.7	5	583
16	Network	61	5.5	14	588
17	Association	58	5.2	26	389
18	Learning	58	5.2	11	548
19	Novel	57	5.1	8	367
20	Neural	55	5.0	10	577
21	Hybrid	52	4.7	21	641
22	Techniques	50	4.5	5	521

23	Model	49	4.4	18	404
24	Multi	49	4.4	14	557
25	Efficient	46	4.1	13	298
26	Rule	45	4.1	20	369
27	Algorithms	44	4.0	19	810
28	Rough	44	4.0	31	1119
29	Optimization	43	3.9	12	445
30	Decision	42	3.8	6	434

5. MAJOR FINDINGS

Based on the analysis undertaken the present study, the following findings are drawn.

1. The findings of Indian research productivity in Data Mining has the highest publication as 202 in the year 2019 with 186 Global Citation Scores followed by 143 papers in 2018 with 603 Global Citation Score and 123 papers in 2016 with 1159 Global Citation Scores. The lowest publication is 3 in 1999 and 2000 with 102 and 19 Global Citation Scores.
2. The authorship pattern of Indian research productivity on Data Mining has identified that majority of papers are multi-authored.
3. The study found that the total research output of the Data Mining for the study period (1999 – 2020) published in 470 journals. As the major portion of the research productivity (34.8%) covered by 30 journals that is coincide with the theory of Bradford's Law of scattering of journals in research productivity.
4. Top 25 institutions were contributed 476 (42.9%) articles of the total research productivity.
5. The findings of distribution of Indian Data Mining scientists published articles in the journals of various countries reveal the fact that Indian Data Mining scientists have contributed their research focus mainly in Indian journals. The countries such as USA, People R China United Kingdom (UK) and Australia have considerably recognized the research articles of Indian Data Mining scientists and published the same in their journals. It is not up to the mark in the case of other countries.
6. The formulated of the applicability of Bradford's law of scattering in various journals is identified as invalidated.

7. The formulated of the implication of Lotka's law related with author productivity in Data Mining identified as invalidated

REFERENCES

1. Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering,* 137, 85–86.
2. Dhawan, S M., Gupta, B M and Ritu Gupta (2017). Mobile computing: A Scientometric Assessment of global publications output, Annals of Library and Information Studies, 64, 172-180.
3. Gupta, B M., Dhawan, S M ., Mand Ritu Gupta,S (2018). Mobile Research in India : A Scientometric Assessment of Publications Output during 2007-16, Journal of Library and Information Technology, 38(1),41-48
4. Gupta, B.M. HarKaur and AvinashKshitig (2012). Dementia research in India: A scientometric analysis of research output during 2002-11, *Annals of Library and Information Studies (ALIS),* 59(4), -288.
5. Lotka's AJ, (1926). The Frequency distributing scientific products,*Journal of washing for Academic of Science,* 6, 317-323.
6. Potter WG, (1981).Lotka's Law Revised, *Library Trends*, 30(1),39.
7. Rahul,L.R and Nishy,P (2016). Mycobacterial tuberculosis and leprosy in India: a scientometric Study, Annals of Library and Information Studies, 63,140-153.
8. Ranganathan, C (2014).Indian Scientists Contribution of Data Mining Research: A Scientometric Profile. Library Progress (International). 34 (2),113-128**.**
9. Ranganathan, C (2014).Mapping of Oceanography Research Productivity in India: A Scientometric Analysis. Library Philosophy and Practice (e-Journal).Winter 12-23-2014. Paper 1205. Link: http://digitalcommons.unl.edu/libphilprac/1205
10. Tripathi,H.K and Garg,K.C (2016). Scientometrics of Cereal crop science research in India as seen through SCOPUS database during 1965-2010, Annals of Library and Information Studies, 63, 222-231.
11. Varaprasad, S.J.D., Ramesh, D.B and Mitali, M. (2011) Scientometrics of India's chemistry during 1987 to 2007, *A Journal of Library and Information Science*, 5(3),67-74.