Mapping of Literature on Data Mining by J-Gate Database

Conference Paper · December 2019

CITATIONS
0

READS
4

3 authors, including:

Sidlingappa M Huded
Azim Premji University
7 PUBLICATIONS **1** CITATION
SEE PROFILE

Suresh Balutagi
Indian Institute of Science
6 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Mapping of Literature on Data Mining by J-Gate Database

Sidlingappa M Huded[*], Suresh Balutagi[**] & Aditya Ranjan[***]

Abstract

The present study discusses the "Data Mining" as reflected in J-gate for the period from 2010–2019. The present paper investigates the research area, productive authors, highly ranked authors, year-wise growth, and geographical distribution of the research output. The result indicates that there was a total of 6107 articles had been published on Data Mining from 2010 to 2019. The result shows that India and the United States of America have contributed the highest articles on Data Mining. This study reveals that the highest number of publications were published in the subject Information Science and Systems, and Jason H Moore from the University of Pennsylvania, USA has the highest number of articles during the selected period, i.e. 2010-2019.

Keywords: *Data mining, Bibliometric study, J-Gate*

1. INTRODUCTION

The data is an important aspect of human lives; it helps in making decisions whether it may have related to academic, business or any other areas. Data mining is also known generally as data knowledge or data discovery. Data mining is a mixture of traditional and modern data analysis techniques with the modern method of data process. Different techniques are used to analyse the data of published literature, librametric, bibliometric, scientometrics, and informetrics are different approaches of analysis in which published literature become subject. Bibliometric as measuring technique used to quantify and understanding the growth of published literature. The bibliometric method used to find authorship, geographical distribution, subject, year-wise growth, citations, publications, and their rankings.

[*] *siddumh20@gmail.com* *Azim Premji University Bengaluru*
[**] *bbsuresh@iisc.ac.in* *Indian Institute of Science Bengaluru*
[***] *adityakarn118@gmail.com* *Azim Premji University Bengaluru*

2. DATA MINING

Encyclopedia of Britannica defines "Data mining, also called knowledge discovery in databases, in computer science, the process of discovering interesting and useful patterns and relationships in large volumes of data. The field combines tools from statistics and artificial intelligence (such as neural networks and machine learning) with database management to analyse large digital collections, known as data sets. Data mining is widely used in business (insurance, banking, retail), science research (astronomy, medicine), and government security (detection of criminals and terrorists)." In the 1960s, statisticians used terms like "Data Dredging" to refer to the process of analysing data without a theoretical hypothesis. The term "Data Mining" appeared around 1990 in the database community Senagupta (2017).

The Technology Review published in 2001 by MIT called Data Mining will be one of the most important emerging technology that changing the world.

3. ABOUT J-GATE

J-Gate is an electronic gateway to global e-journal literature. Launched in 2001 by Informatics India Limited, J-Gate provides seamless access to millions of journal articles available online offered by 12,05 Publishers. It presently has a massive database of journal literature, indexed from 54,715 e-journals with links to full text at publisher sites. J-Gate also plans to support online subscription to journals, electronic document delivery, archiving and other related services.

4. OBJECTIVES OF THE STUDY

The following are the objectives of the Study:

- To know the number of full-text research articles on data mining available during 2010-2019.
- To identify the highly productive authors on data mining.
- To determine the geographical distribution of significant articles on data mining from 2010-2019.
- To know the growth of publications during 2010-2019 on data mining.
- To identify the ranking list of most productive journals preferred by researchers to write the articles on data mining
- To study the availability of articles on data mining in the J-Gate database.

5. STATEMENT OF THE PROBLEM

Data Mining is an important application that helps to extract the required data from huge datasets, due to the technology, the application of data mining has been extended to almost all the fields such as bioinformatics, digital libraries, medicine, business, information industry, etc. Hence, the bibliometric study of

Data Mining will help the researchers to know the trends, publishing patterns, source journals, growth, etc. for a better understanding of recent trends in Data Mining, therefore the researchers felt to study the particular area by using J-Gate Database.

5. SCOPE AND LIMITATION

The present study is limited to search results on data mining in the J-Gate database from 2010-2019. Basic data has been collected, processed and analysed manually.

6. METHODOLOGY

The present study uses the J-Gate database for retrieving the data for the study. This study uses 10 years' data (2010-2019) for study in the field of data mining. The J-Gate was accessed on 25th September 2019 for the collection of data on data mining. The statistical data acquired from the database was put to excel worksheet to analysis and understand the objectives set for the study. The data were interpreted and analysed to get the precise picture of the research area, productive authors, highly ranked authors, year-wise growth, and geographical distribution of the research output. The data has been calculated and represented in tables. Quantitative and inferential methods have been used to analyse the data.

7. REVIEW OF LITERATURE

Wu et al. (2009) have discussed data mining for a traffic flow; the authors have analysed 3870 papers published in Science Citation Index SCI) from 1976-2005. They have evaluated the contribution of authors, annual production, document type, language distribution of publications, page count, number of REFERENCES, the geographical distribution of research output, and keywords. The authors found that the USA is the top producing country with 1359 publications of the total output; there were eight keywords with frequency over 100 times; the highest frequency keyword was traffic flow with 378 (12.52%).

Yang et al. (2014) have analysed the articles on data mining using "Data Mining or Knowledge Discovery" keywords from Science Citation Index, Social Sciences Citation Index and Arts & Humanities Citation Index during 1983-2013. The authors have analysed 16689 articles and discussed the distributions, top-ranking authors, affiliations of the top-ranked authors, and year wise growth of the literature during 1983-2013.

Le and Zeng (2017) have analysed current research hotspot and trends in data mining. The authors have extracted the data from the Web of Science (WoS) from 1993 to 2016 and covered 11071 articles from 131 countries. The researchers found that the USA topped the list followed by China and Taiwan.

Xiao et al. (2019) have discussed the future directions of data mining in nursing, researchers have extracted the data for 407 papers from English and

Chinese databases from 1990 to 2017 with data mining and nursing-related keywords. This study investigates the comparative study of trends of data mining in nursing from English and Chinese databases.

8. DATA ANALYSIS

8.1. Research Area

Table 1 presents the area of research where most articles on data mining have been published. This table indicates that the maximum number of contributions were in the area of Information Science and Systems, i.e. 1650 followed by Computer Science (Hardware & Networks) 1502, Electrical Engineering 625, Civil Engineering 488, Mechanical Engineering 443, Electronics 379, Communication Networks & Technology 364 Artificial Intelligence 259, and Neural Networks 255.

Table 1: Analysis of Data Mining Articles by Top -10 Research Area

Distribution of research area	*No. of articles*	*Rank*
Information Science and Systems	1650	1
Computer Science (Hardware & Networks)	1502	2
Software Engineering	849	3
Electrical Engineering	625	4
Civil Engineering	488	5
Mechanical Engineering	443	6
Electronics	379	7
Communication Networks & Technology	364	8
Artificial Intelligence	259	9
Neural Networks	255	10

8.2 Productive Authors

Among the distribution of authors, the most productive author is Jason H Moore from the University of Pennsylvania who topped the list with 12 publications during the period, followed by Kaori Kadoyama, Saurabh Pal, Toshiyuki Sakaeda, and Yasushi Okuno who have contributed 10 articles each during 2010-2019. Toshio Kobayashi has 9 publications, followed by A Govardhan, D Aruna Kumari, Kanwal Garg and R R Shelke who have penned 8 articles each in this period. Kyoto University, Japan has 30 publications to its credit and three authors from the University are in the top five authors.

Table 2: Distribution of Top Authors Written on Data Mining

Name of the author	*Affiliation of the author*	*No. of publications*	*Rank*
Jason H Moore	University of Pennsylvania, USA	12	1
Kaori Kadoyama	Kyoto University, Japan	10	2
Saurabh Pal	VBS Purvanchal University, India	10	2
Toshiyuki Sakaeda	Kyoto University, Japan	10	2
Yasushi Okuno	Kyoto University, Japan	10	2
Toshio Kobayashi	Miyagi University, Japan	9	3
A Govardhan	JNTUH College of Engineering, India	8	4
D Aruna Kumari	K L University, India	8	4
Kanwal Garg	Kurukshetra University, India	8	4
R R Shelke	Dr. Panjabrao Deshmukh Krishi Vidyapeeth, India	8	4

8.3 Highest Ranked Journals

Table 3 presents the top 10 most productive journals which have published the highest number of research articles on data mining. Out of the 6107 articles, the first 10 highest ranked journals have published 1061 articles which share 17.37% of the total output during 2010-2019. Among the top 10 journals, *International Journal for Scientific Research and Development* has published the highest articles i.e. 209, followed by, *International Journal of Advanced Research in Computer and Communication Engineering* 165, *International Journal of Advanced Research in Computer Science* 165, *International Journal of Advanced Research in Computer Science* 139, *International Journal of Computer Science and Information Technologies* 138, *International Journal of Computer Science and Mobile Computing* 88, *Procedia Computer Science* published by Elsevier 88, *International Journal of Innovative Research in Computer and Communication Engineering* 68, *Indian Journal of Science and Technology* occupied 7th place by publishing 66 articles, *International Journal of Scientific and Engineering Research* 64, *and International Journal of Engineering Research and Applications* 56 during 2010-2019.

Table 3: Highest-ranked Journals

Highest ranked journals	*No. of articles*	*Rank*
International Journal for Scientific Research and Development	209	1
International Journal of Advanced Research in Computer and Communication Engineering	165	2
International Journal of Advanced Research in Computer Science	139	3
International Journal of Computer Science and Information Technologies	118	4

Highest ranked journals	*No. of articles*	*Rank*
International Journal of Computer Science and Mobile Computing	88	5
Procedia Computer Science	88	5
International Journal of Innovative Research in Computer and Communication Engineering	68	6
Indian Journal of Science and Technology	66	7
International Journal of Scientific And Engineering Research	64	8
International Journal of Engineering Research and Applications	56	9

8.4 Year-wise Growth

Table 4 depicts the year wise publications output on data mining for the period of 2010 to 2019. In the year, i.e. 2010, there were only 197 articles published which shares only 3.23% of the total output. The highest number of articles, i.e. 926 were published in the year 2014 which shares 15.16%, followed by, 883 in 2015, 804 in 2016, 724 in 2017, 706 in 2013, 622 in 2018, 552 in 2012, 374 in 2011, and 319 in 2019. However, the year 2019 may get some more articles until December 2019 as the data was taken in September 2019. More specifically, an average of 611 articles means 10% were published during this period.

The research in data mining from 2013 to 2017 has seen drastic growth, these five years alone shares 4043 (66.20%) of the total output. Filippov and Hofheinz (2016) have identified the reason that during these years, Asia has emerged the highest contributor to the world outputs that replaced the European union in publishing academic research on data mining. During 2011-2016, the outputs rose to 32.4% from 31.1% in 2000.

Table 4: Year-wise Growth

Year of publication	*No. of articles*	*Percentage*
2019	319	5.22
2018	622	10.18
2017	724	11.87
2016	804	13.17
2015	883	14.46
2014	926	15.16
2013	706	11.56
2012	552	9.03
2011	374	6.12
2010	197	3.23
TOTAL	**6107**	**100**

8.5 Geographical Distribution

The table-5 depicts the country-wise distribution of the research articles on data mining from 2010-2019, here articles are analysed according to their country of origin. Out of 6107 articles, the 10 countries have contributed 5260 articles which share 86.13% of the total output. The table 5 shows that India is the most productive country with 2709 articles, followed by USA 1152, UK 458, Netherlands 256, Pakistan 153, Iran 138, Switzerland 129, Germany 105, Singapore 89, and Romania 71 during the period.

Table 5: Country-wise Distribution

Name of the country	*No. of articles*	*Rank*
India	2709	1
USA	1152	2
UK	458	3
Netherlands	256	4
Pakistan	153	5
Iran	138	6
Switzerland	129	7
Germany	105	8
Singapore	89	9
Romania	71	10

9. INDINGS

The authors analysed the research publications on the Data Mining indexed in the J-Gate database for the period from 2010 to 2019. As per the data available on the J-Gate database, a total of 6107 articles has been published during the period with an average of 611 articles per year and witnessed an exponential growth in publications. *The International Journal for Scientific Research and Development* was the most prolific journal on Data Mining which has 209 articles in its account. Jason H Moore from the University of Pennsylvania, USA, was the most prolific author with the maximum number of publications during the period. India tops the list with 2709 (44%) articles followed by the USA with 1122 respectively, these two countries share 63.22% of the total publications. The highest publications were seen in 2014 with 926 articles (15.16%), and fewer articles were published in 2010 with 197 (3.23%).

10. CONCLUSION

The main aim of the current study was to examine the trend of 'Data Mining' in J-Gate. In the study, the authors analysed and presented the data of published literature on Data Mining indexed in J-Gate through the bibliometric study by

research area, top authors, year wise and country wise. The study found that the highest number of publications are contributed by Indian authors. Also, the study revealed that all contributors are from reputed institutes. This study is limited to search results from the J-Gate database only and not to cover the articles published in other journals and databases; therefore, additional investigation in the enormous area is required. This study can be concluded that '*The International Journal for Scientific Research and Development*' has published the highest number of articles and can be considered as one of the top sources of articles in the area of 'Data Mining'.

REFERENCES

1. Clifton, Christopher. "Data Mining." *Encyclopædia Britannica*. Encyclopædia Britannica, inc. Accessed October 15, 2019. https://www.britannica.com/technology/data-mining
2. Filippov, Sergey, and Paul Hofheinz. "Text and Data Mining for Research and Innovation ." *Lisbon Council Policy Brief*, no. 20 (2016): 1–16. https://lisboncouncil.net
3. "J-Gate." Accessed October 17, 2019. https://jgateplus.com/search/footer-html/AboutUs.jsp.
4. Li, Ray. "The History of Data Mining." Dataconomy, June 10, 2016. https://dataconomy.com/2016/06/history-data-mining/.
5. Li Z., Zeng L. (2017) Research Hotspots and Trends in Data Mining: From 1993 to 2016. In: Tan Y., Takagi H., Shi Y. (eds) Data Mining and Big Data. DMBD 2017. *Lecture Notes in Computer Science*, vol 10387. Springer, Cham
6. MIT Technology Review. "10 Breakthrough Technologies 2001." *MIT Technology Review*. Accessed October 18, 2019. http://www2.technologyreview.com/tr10/?year=2001.
7. Sengupta, Shantashree. "Applications of data mining in Library and Information Science Centres: An Overview." *International Journal of Current Research* 9, no. 1 (January 31, 2017): 45246–49. http://www.journalcra.com/sites/default/files/issue-pdf/20191.pdf.
8. Wu, Chaozhong, Hu Lei, Xinping Yan, and Feng Zhou. "Data Mining for Bibliometric Analysis of Traffic Flow." *2009 Second Pacific-Asia Conference on Web Mining and Web-Based Application*, 2009. https://doi.org/10.1109/wmwa.2009.53.
9. Xiao, Qian, Jiani Wang, Yanling Wang, and Ying Wu. "Data Mining in Nursing: A Bibliometric Analysis (1990–2017)." In *The 17th World Congress on Medical and Health Informatics*. Lyon: IOS, 2019. http://ebooks.iospress.nl/volume/medinfo-2019-health-and-wellbeing-e-networks-for-all-proceedings-of-the-17th-world-congress-on-medical-and-health-informatics.
10. Yang JM., Tseng SF., Won YL. (2016) A Bibliometric Analysis on Data Mining Using Bradford's Law. In: Juang J. (eds) Proceedings of the 3rd International Conference on Intelligent Technologies and Engineering Systems (ICITES2014). *Lecture Notes in Electrical Engineering*, vol 345. Springer, Cham.