DLINE JOURNALS

International Journal of Information Studies

Print ISSN: 1911 8408
Online ISSN: 1911 8414

IJIS 2024; 16 (2)
https://doi.org/10.6025/ijis/2024/16/2/37-48

Most Cited Articles in the Field of Data Mining: A Bibliometric Study

Varsha Singh
INFLIBNET
India
varsha6904560@gmail.com

Avinash Kumar Singh
Babasaheb Bhimrao Ambedkar University (A Central University)
Lucknow 226025, U.P. (INDIA)
India
avinashsingh4423@gmail.com

Received: 11 November 2023
Revised: 8 January 2024
Accepted: 17 January 2024

ABSTRACT

Background: *The simplest technique to discover the most recent and challenging study material across all subject areas is through bibliometrics.*

Aim: *Researchers use bibliometrics to investigate users' requirements across all fields. The results assist researchers in making sense of the numerous critical issues. The researcher employed a bibliometrics methodology in the current investigation.*

Methodology: *The current study made use of the Scopus Index core collection. Only those papers that were published in the area of data mining were focused on this research. Between 1995 and 2021, 34,011 works were discovered that were published in too many different languages (27 years). However, the researcher set a boundary and only chose those publications that were written in the English language and received the highest citations.*

Findings: *The investigation results revealed that the 2008 publication of "Top 10 methods in data mining" in the journal "Knowledge and Information Systems" received 3400 citations. The results of this study also showed that articles with several writers received the most citations. The top countries for data mining productivity are also mentioned in the study. The study's findings also look at the most popular journals and keywords utilised in the published articles. The current study's findings are very helpful for researchers who plan to conduct research in the field of data mining.*

Keywords: Data Mining, Bibliometric, Scopus Index Database, Top cited Articles, Most productive year

1. Introduction

Data mining gets its name from how similar it is to looking for important information in a big database to utilise later(VSSUT, Burla). Because this domain is so important, high-quality research communities are increasingly focussed on knowledge discovery in datasets, but they tend to operate in silos. The high-quality, technical research communities build powerful general-purpose solutions tailored to individual users or user groups and collect, represent and model user data. The higher-quality, education-focussed communities offer rich expertise in cognitive, psychological or learning science perspectives, as the intelligent techniques they adopt or suggest are focused on different kinds of data. It is a computer procedure for identifying patterns in massive data sets that combine techniques from artificial intelligence, machine learning, statistics, and database systems. Data mining aids in determining user wants and demands. Finding the elements that might draw in new users is also helpful for forecasting. Three methods are typically used in data mining analysis: traditional statistics, artificial intelligence, and machine learning (Girija & Srivatsa, 2006).

Data mining is the process of extracting knowledge from huge data sets. The information retrieved can be used for various purposes, including market analysis, fraud detection, customer retention, production control, and more(Tutorials Point (I) Pvt. Ltd, 2014).

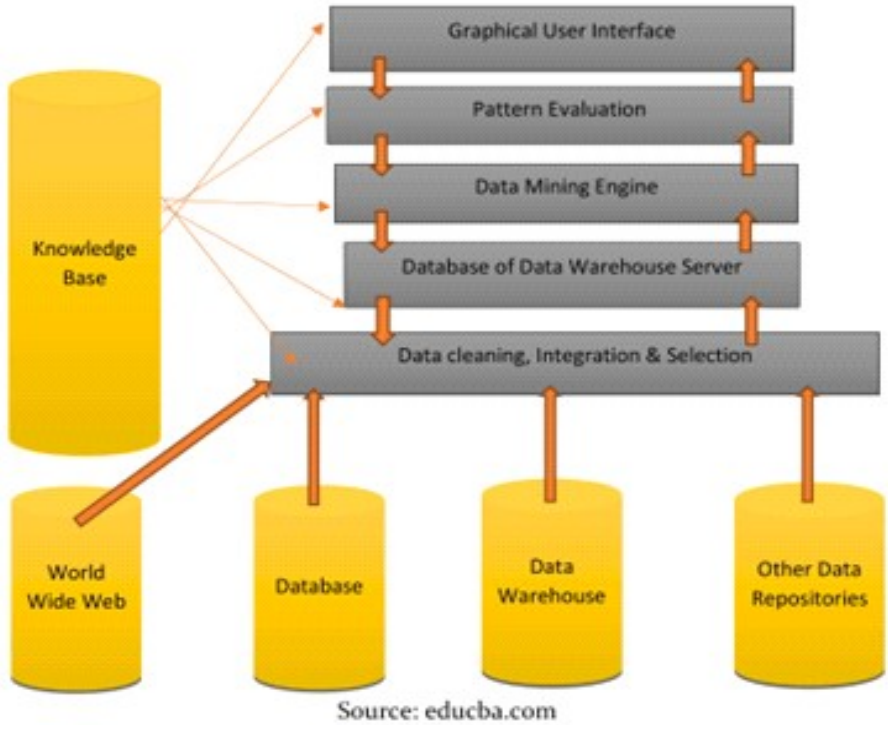

Figure 1. Data Mining Architecture

Given that data mining is the process of uncovering novel, fascinating, and potentially profitable patterns in large data sets and extracting hidden information using algorithms. Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archaeology, data dredging, and information harvesting are all terms used to describe data mining. Any data mining approach aims to provide an efficient predictive or descriptive model of a large amount of data that not only best fits or explains it but can also generalise to new data. According to a broad view of data mining skills, data mining is extracting meaningful knowledge from massive amounts of data stored in databases, warehouses, or other information repositories. Its detailed analysis is essential before going onward towards data applicability in decision-making. The goal of the study is to examine the most popular data mining articles that have been mentioned.

Theoretical Background

Aksnes introduced a conceptual distinction between quality dynamics and visibility dynamics. Quality refers to the content, and the number of citations is not an indicator that reflects the value of scientific papers. Several citations may reveal better visibility. The value and number of citations may not correlate as they do not denote the same concept.Newman described a method for predicting which scientific papers will be highly cited in the future, even if they are not cited. It is a prediction exercise, and this view offers some new insights.

2. Review of Literature

In several industries where huge volumes of data are accessible, data mining techniques are utilised to unearth undiscovered or buried information. Nicholson and Stanton (2003) coined the term 'bibliomining' or data mining for libraries to characterise the combination of data warehouses, data mining, and bibliometrics. This phrase refers to studyinglibrary system transaction patterns, behavioural changes, and trends. Although the concept is not novel, the name "bibliomining" was coined to enable searching for the phrase's "library" and "data mining" in the context of physical libraries, as opposed to digital libraries. Bibliomining is a valuable strategy for locating relevant library material in historical data to enhance decision-making (Kao et al., 2003). To offer a comprehensive report on the library system, bibliomining must be utilised iteratively in conjunction with other measurement and assessment techniques; as strategic information is discovered, new questions may be addressed, restarting the process (Nicholson, 2003b).

As with any other approach to knowledge extraction, bibliomining must adhere to a systematic methodology to enable proper information discovery. The bibliomining technique begins with selecting relevant themes and acquiring data from internal and external sources (Nicholson, 2003b). The data are then gathered, cleansed, and anonymisedbefore being stored in a data warehouse. The bibliomining process comprises picking the most applicable analysis tools and techniques from statistics data mining, and bibliometrics to identify significant patterns in the acquired data (Nicholson, 2006a). Reports analyse and illustrate significant patterns. The mining approach will continue until the acquired data is reviewed and confirmed by important users, such as librarians and library managers (Shieh, 2010).

Bibliomining technologies are a new trend that might be used better to analyse behaviour patterns among library customers and staff and patterns of information resource use throughout the library (Nicholson & Stanton, 2006). Bibliomining is highly recommended for producing vital and necessary information for library management requirements, emphasising professional librarianship difficulties despite its reliance on databases (Shieh, 2010). Bibliomining may also offer a full perspective of the library process to evaluate staff performance, recommend areas for improvement, and anticipate future user requests (Prakash, Chand, & Gohel, 2004). The gathered data permits scenario analysis of the library system, which evaluates multiple factors that must be considered throughout the decision-making procedure (Nicholson, 2006a). Another application is to standardise formats and reporting so that data warehouses may be shared among library groups, enabling libraries to benchmark their data (Nicholson, 2006a). Therefore, applying data mining techniques in libraries is advantageous in enhancing the quality of interaction between a library and its patrons (Chang & Chen, 2006). This study will categorise the most popular articles and their journals for effective library decision-making.

International papers are not well represented among high-impact papers in research specialities, but they dominate highly cited papers from small countries, cities, and institutions within them (Olle Persson). Bao-Zhong Yuan and Jie Sun's study demonstrated that more top papers come from journals with higher IF and higher ranks in the WoS Category. So, authors can choose their ideal journal with a high impactfactor or Q1 in Category to publish their papers in the English language related to their research field.

Research Questions

This work primarily intends to track the highly cited articles in data mining as reflected in the Scopus database over a given period. Studying the characteristics of highly cited papers will help

users and scientists understand how to generate high-level papers based on optimum research. Only a specific period accounts for high-impact papers in a large span window. These high-impact papers contribute to the growth of the domain. A study of these papers in a large window indicates the scholarly papers. Typically, they are authored by many scientists, often involving international collaboration. (Dag W Aksnes 2003)

These high-impact papers are likely to be published in many journals, of which high-impact journals may account for most. These journals benefit from publishing high-cited papers as they increase the impact factors and other citation counts for them. The leading countries' scientists can be brought to the attention of the data mining community and research evaluation systems. We also studied the highly used phrases to understand the concepts that cause the growth of a domain. We also like to record the keywords used in these highly cited papers.

The study of the characteristics of highly cited papers is interesting. Most of the research carried out on this theme uses mainly bibliometric indicators. In this work, we intend to use text representations besides bibliometric data.

3. Methodology

Standard bibliometric techniques were used to conduct this investigation. The author used the Scopus Index database to get the information. There are 34,431 publications about data mining that are available in English language. The top 20 articles cited the most out of the 34,431 articles received for the research have been chosen. The researcher used Microsoft Excel and MS Office to tabulate the data and make interpretationsfor data analysis. In the table titled "Top Productive Year", the Researcher displayed the years in which the top 100 cited articles were published in the data analysis and interpretation section.

4. Data Analysis and Interpretation

4.1. Top Productive Year

The table revealed that the most productive year in data mining. The study demonstrates the years of articles published from 1995 to 2017 in Scopus-indexed journals. The result shows that 2002 and 2014 were the most productive years, with 10 and 8 articles published in Scopus-indexed journals, respectively. Only 1 article was published in the years 1995 and 1997.

S. No.	Year of Publication	Total No. of Publications
1	1995	1
2	1996	2
3	1997	1
4	1998	4
5	1999	4
6	2000	3
7	2001	4
8	2002	10
9	2003	6
10	2004	4

11	2005	6
12	2006	5
13	2007	3
14	2008	3
15	2009	6
16	2010	4
17	2011	6
18	2012	4
19	2013	6
20	2014	8
21	2015	2
22	2016	4
23	2017	3

Table 1. Top productive years of highly cited papers

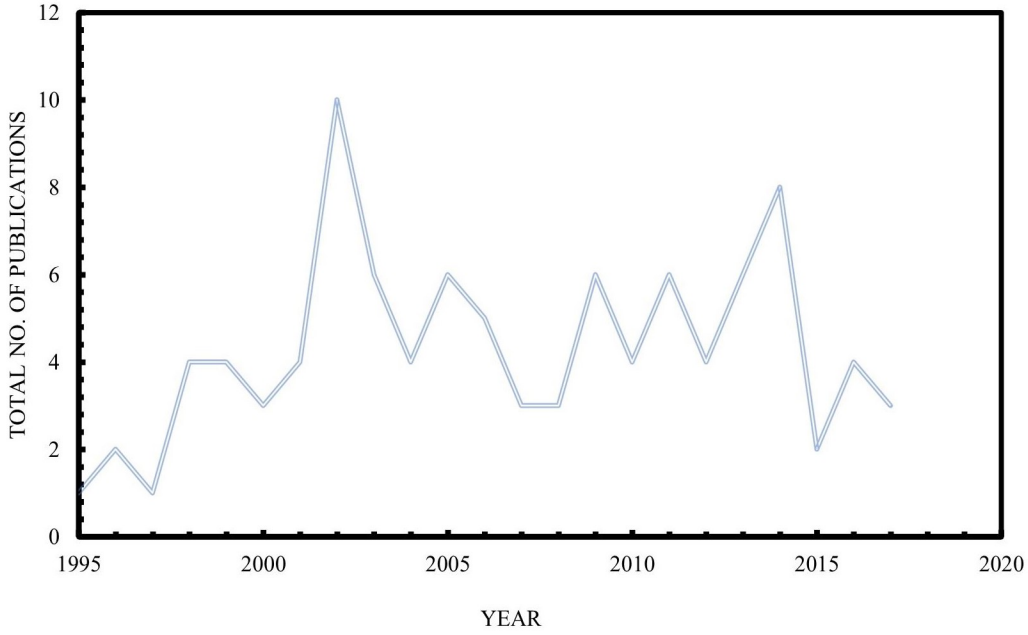

Figure 1. Top productive year

4.2. Top Twenty cited Articles with Authors, Year and Title

Table 2 depicts the top twenty cited articles with authors, years, and titles in the field of data mining. This paper shows that more than one author appears in the selected top 20 cited articles. According to the table, the "Top 10 algorithms in data mining" title with 3400 citations were published in 2008 and took first rank and second place in "From data mining to knowledge discovery in databases" title with 2704 citations that were published in 1996.

Rank	Authors	Year	Title	Total no. of Citations
1.	Wu X., Kumar V., Ross Q.J., Ghosh J., Yang Q., Motoda H., McLachlan G.J., Ng A., Liu B., Yu P.S., Zhou Z.-H., Steinbach M., Hand D.J., Steinberg D.	2008	Top 10 algorithms in data mining	3400
2.	Fayyad U., Piatetsky-Shapiro G., Smyth P.	1996	From data mining to knowledge discovery in databases	2704
3.	Götz S., García-Gómez J.M., Terol J., High-throughput functional Williams T.D., Nagaraj S.H., Nueda M.J., Robles M., Talón M., Dopazo J., Conesa A.	2008	mining with the Blast2GO suite	2668
4.	Agrawal R., Srikant R.	2000	Privacy-preserving data mining	2322
5.	Wu X., Zhu X., Wu G.-Q., Ding W.	2014	Data mining with big data	1892
6.	Agrawal R., Gehrke J.,, Gunopulos D., Raghavan P.	1998	Automatic subspace clustering of high dimensional data for data mining applications	1860
7.	Alcalá-Fdez J., Fernández A., Luengo J., Derrac J., García S., Sánchez L., Herrera F.	2011	KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework	1649
8.	Ramaswamy S., Rastogi R.,. Shim K	2000	Efficient algorithms for mining outliers from large data sets	1469
9.	García S., Fernández A., Luengo J., Herrera F.	2010	Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power	1364
10.	Buczak A.L., Guven E.	2016	A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection	1275

11.	Keim D.A.	2002	Information visualization and visual data mining	1128
12.	Jia F., Lei Y., Lin J., Zhou X., Lu N.	2016	Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data	1126
13.	Menzies T., Greenwald J., Frank A.	2007	Data mining static code attributes to learn defect predictors2	1032
14.	Alcalá-Fdez J., Sánchez L., García S., del Jesus M.J., Ventura S., Garrell J.M., Otero J., Romero C., Bacardit J., Rivas V.M., Fernández J.C., Herrera F.	2009	KEEL: A software tool to assess evolutionary algorithms for data mining problems	1010
15.	Stelzer G., Rosen N., Plaschkes I., Zimmerman S., Twik M., Fishilevich S., Iny Stein T., Nudel R., Lieder I., Mazor Y., Kaplan S., Dahary D., Warshawsky D., Guan-Golan Y., Kohn A., Rappaport N., Safran M., Lancet D.	2016	The GeneCards suite: From gene data mining to disease genome sequence analyses	986
16.	Romero C., Ventura S.	2007	Educational data mining: A survey from 1995 to 2005	961
17.	Demšar J., Curk T., Erjavec A., Gorup C., Hoèevar T., Milutinoviè M., Možina M., Polajnar M., Toplak M., Stariè A., Štajdohar M., Umek L., Žagar L., Žbontar J., Žitnik M., Zupan B.	2013	Orange: Data mining toolbox in python	918
18.	Hall M.A., Holmes G.	2003	Benchmarking Attribute Selection Techniques for Discrete Class Data Mining	898
19.	Parpinelli R.S., Lopes H.S., Freitas A.A.	2002	Data mining with an ant colony optimization algorithm	796
20.	Robbins P.F., Lu Y.-C., El-Gamil M., Li Y.F., Gross C., Gartner J., Lin J.C., Teer J.K., Cliften P., Tycksen E., Samuels Y., Rosenberg S.A.	2013	Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumour- reactiveT cells	782

Table 2. Top twenty cited articles with authors, year, and title

4.3. Most Cited Journals

Table 3 depicts the top-cited journals on data mining. The table explores twenty journals' names that are highly cited. According to this table, the "Knowledge and Information Systems" journal secured the first position in data mining.

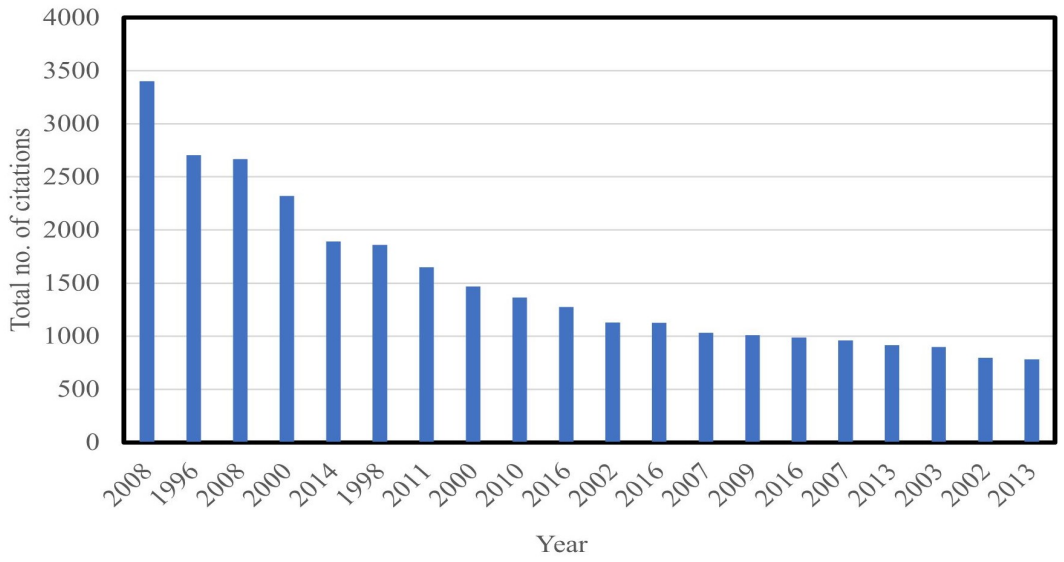

Figure 2. Top twenty cited articles with year

S. No.	Name of Journal
1	Knowledge and Information Systems
2	AI Magazine
3	Nucleic Acids Research
4	SIGMOD Record (ACM Special Interest Group on Management of Data)
5	IEEE Transactions on Knowledge and Data Engineering
6	Journal of Multiple-Valued Logic and Soft Computing
7	Information Sciences
8	IEEE Communications Surveys and Tutorials
9	IEEE Transactions on Visualization and Computer Graphics
10	Mechanical Systems and Signal Processing
11	IEEE Transactions on Visualization and Computer Graphics
12	Mechanical Systems and Signal Processing
13	IEEE Transactions on Software Engineering
14	Soft Computing
15	Current Protocols in Bioinformatics
16	Expert Systems with Applications
17	Journal of Machine Learning Research
18	IEEE Transactions on Knowledge and Data Engineering
19	IEEE Transactions on Evolutionary Computation
20	Nature Medicine

Table 3. Most Cited Journal

4.4. Country with the Highest Productivity

Table 4 reveals the country with the highest productivity in data mining. The United States is a highly productive country, placing in rank 1 in the field of data mining. Spain and China secured second and third places, respectively.

4.5. Most occurred Phrases in Data Mining

The reflection of the content is measured in terms of phrases. These phrases contain the significant words such as data mining, data management, data acquisition, and algorithm as identified from Scopus.

Table 5 demonstrates the most used keywords in data mining. The keywords most commonly used were in Scopus-indexed journals.

Rank	Name of country
1	United States
2	Spain
3	China
4	Israel
5	South Korea
6	Germany
7	Slovenia
8	New Zealand
9	Brazil
10	Canada
11	Australia
12	Japan

Table 4. Country with the Highest Productivity

4.6. Most occurred Phrases in Data Mining

The reflection of the content is measured in terms of phrases. These phrases contain the significant words such as data mining, data management, data acquisition, and algorithm as identified from Scopus.

Table 5 demonstrates the most used keywords in data mining. The keywords most commonly used were in Scopus-indexed journals.

Rank	Keywords
1	Data Mining
2	Data Handling

3	Data Acquisition
4	Algorithm
5	Machine learning
6	Decision trees
7	Nonhuman
8	Gene expression
9	Feature extraction
10	Pattern recognition

Table 5.The most used Keywords in Data Mining

4.7. The average Number of Pages in Published Articles

Table 6 revealed an average page number in published articles. According to the table, approximately 15 pages of each author's article were written.

Total 298/20= 14.9 ***(Approx. 15 pages)***

The average page numbers of top-cited articles in data mining are 14.9. (Approx. 15 pages).

Page Start	Page End	Total No. of Page
1	37	37
37	53	17
3420	3435	15
439	450	11
97	107	10
94	105	11
255	287	32
427	438	11
2044	2064	20
1153	1176	23
1	8	7

303	315	12
1301	13033	33
135	146	11
2349	2353	4
1437	1447	10
321	332	11
747	752	5
2832	2842	10
46	54	8
Total		**298**

Table 6. The average number of pages in published articles

5. Conclusion

Nicholson and Stanton (2003) coined the term "bibliomining," or "data mining for libraries," to refer to the fusion of data warehousing, data mining, and bibliometrics. This phrase is used to track transaction patterns, behaviour alterations, and library system trends. The most popular research technique in any discipline is the analysis of research output using bibliometrics. The most referenced papers in the field of data mining are methodically identified and categorised in this study. The study's conclusions have several advantages for data mining academics and practitioners. The findings also assist new researchers in learning from the types of contributions, strategies, and research techniques used in highly-cited articles to write higher-quality studies that are more likely to receive high citations. Researchers can identify the most cited researchers to collaborate with, seek advice from, etc. Practitioners can identify the highest quality work in particular areas of Data Mining and aim at utilising techniques, tools, or findings reported in those studies thanks to the classifications, which also help established and new researchers identify the active and more influential topics.

References

[1] Yuan, Bao-Zhong., Sun, Jie. (2020). Mapping the scientific research on maize or corn: a bibliometric analysis of top papers during 2008–2018. *Maydica Electronic Publication*, 65(17).

[2] Bollen, Johan., Luce, Richard. (2002). Evaluation of digital library impact and user communities by analysis of usage patterns. *D-Lib Magazine*, 8(6).

[3] Chellappandi, P., Vijayakumar, C. S. (2018). Bibliometrics, Scientometrics, Webometrics / Cybermetrics, Informetrics and Altmetrics - An Emerging Field in Library and Information Sci ence Research. *Shanlax International Journal of Education*, 7(1), 5-8.

[4] Aksnes, Dag W. (2003). Characteristics of highly cited papers. *Research Evaluation*, 12(3), 159–170.

[5] Frater, J. L. (2021). The top 100 cited papers in the field of iron deficiency in humans: a bibliometric study. *BioMed Research International*, 1-9.

[6] Garousi, Vahid., Fernandes, João M. (2015). Highly-cited papers in software engineering: The top-100. *Information and Software Technology*, 71, 108–128.

[7] Girija, N., & Srivatsa, S. K. (2006). A research study: Using Data Mining in knowledge base business strategies. *Information Technology Journal*, 5(3), 590–600.

[8] Kolling, Marcus L., Furstenau, Leandro B., Sott, Maicon K., Rabaioli, Bárbara, Ulmi, Pedro H., Bragazzi, Nicola L., Tedesco, Luís P. C. (2021). Data Mining in Healthcare: Applying Strategic Intelligence Techniques to Depict 25 Years of Research Development. *International Journal of Environmental Research and Public Health*, 18, 3099, 1-20.

[9] Newman, M. E. J. (2014). Prediction of highly cited papers. *Europhysics Letters*, 105(2).

[10] Nicholson, Scott., Stanton, Jennifer M. (2003). Gaining strategic advantage through Bibliomining: data mining for management decisions in corporate, special, digital, and traditional libraries. *In Organizational Data Mining: Leveraging Enterprise Data Resources for Optimal Performance*. Hershey, PA: Idea Group Publishing.

[11] Persson, Olle. (2010). *Are highly cited papers more international? Scientometrics*, 83(2), 397–401.

[12] Ramageri, B. M. Data mining techniques and applications. *Indian Journal of Computer Science and Engineering*, 1, 301-305.

[13] Singh, Anil K., Kumar, Vinod. (2021). Citation analysis of library and information science doctoral theses awarded by universities in India with JabRef reference management software. *Library Philosophy and Practice*.

[14] Sumathi, S., Sivanandam, S. N. (2006). *Introduction to Data Mining and its Applications*. Springer.

[15] *Lecture notes on data mining & data warehousing course code*: bcs-403.