DESIDOC Journal of Library & Information Technology, Vol. 42, No. 2, March 2022, pp. 105-113, DOI : 10.14429/djlit.42.2.17307

Bibliometric Analysis of Latent Dirichlet Allocation

Mohit Garg[#] and Priya Rangra[$,*]

[#]*Indian Institute of Technology, Delhi - 110 016, India*
[$]*Department of Library and Information Science, Central University of Himachal Pradesh, Shahpur - 176 206, India*
[*]*E-mail: priyarangra26494@gmail.com*

ABSTRACT

Latent Dirichlet Allocation (LDA) has emerged as an important algorithm in big data analysis that finds the group of topics in the text data. It posits that each text document consists of a group of topics, and each topic is a mixture of words related to it. With the emergence of a plethora of text data, the LDA has become a popular algorithm for topic modeling among researchers from different domains. Therefore, it is essential to understand the trends of LDA researches. Bibliometric techniques are established methods to study the research progress of a topic. In this study, bibliographic data of 18715 publications that have cited the LDA were extracted from the Scopus database. The software R and Vosviewer were used to carry out the analysis. The analysis revealed that research interest in LDA had grown exponentially. The results showed that most authors preferred "Book Series" followed by "Conference Proceedings" as the publication venue. The majority of the institutions and authors were from the USA, followed by China. The co-occurrence analysis of keywords indicated that text mining and machine learning were dominant topics in LDA research with significant interest in social media. This study attempts to provide a comprehensive analysis and intellectual structure of LDA compared to previous studies.

Keywords: Bibliometrics; Big data; Citation analysis; Latent dirichlet allocation

1. INTRODUCTION

Latent Dirichlet Allocation, commonly known as LDA, is a statistical topic modeling algorithm useful for discovering topics in the dataset. David M. Blei *et al.* introduced it as an unsupervised machine learning-based three-level hierarchical Bayesian model for finding the topics from the text corpora[1]. With the emergence of the Internet and social media technologies, a plethora of text data is generated every day. Several models like pLSA, HMM, HDP, etc., have been discussed in the literature for analyzing text data. However, scholars of the different domains have commonly used LDA to model topics from the text corpus. These studies were published in many reputed international journals. Some popular domains include bibliometric analysis[2-3], social media analysis like Facebook[4-5], Twitter[6-7], YouTube[8-10], Q&A sites[11-12], Stackoverflow[13-14], Analysis of open-ended questions in Surveys[15-16], etc.

The concept of bibliometric was defined by Alan Pritchard in 1969 as mathematical and statistical methods to quantify written communication[17]. Since then, these techniques have been popularly used to portrait the research productivity of disciplines, institutions, countries[18]. Many studies have been found in the literature related to different fields, institutions, and nations, but only a few studies were related to specific algorithms. Dejian, Zeshui & Xizhao[19] studied the trends of research of the most popular algorithm for Machine Learning, i.e., Support Vector Machines (SVM). The authors have used the web of science database and bibliometric techniques to present comprehensive progress and current situations of SVM in China.

Previously there has been some research in the literature related to Topic Modelling and LDA. Li & L[20] presented the bibliometric analysis of topic modeling studies from 2000 to 2017. It indicates that the popularity of LDA is increasing compared to other topic model algorithms like pLSA etc. Jelodar *et al.*[21] made a comprehensive survey of scholarly articles from 2003 to 2016 related to LDA-based topic modeling. The study pointed out the intellectual structure, applications, and various tools available for performing the LDA analysis. The other associated surveys related to topic modeling were explored by Chen[22], Daud[23], Sun[24].

The above reviews have presented some insight about the intellectual structure of LDA with some shortcomings too. The majority of research works were mainly related to Topic Modeling, and only one study was related to LDA with analysis of the small number of researches. The present study overcomes this by analysing the large number of studies that have cited LDA in their research.

2. OBJECTIVES

The main aim of this study is to examine the comprehensive research trends of LDA-based publications. The objectives of the present study are:

- To study the growth of scientific interest in LDA;

Received : 07 July 2021, Revised : 18 December 2021
Accepted : 14 January 2022, Online published : 28 February 2022

- To identify influential Authors, Institutions, and countries in referring LDA;
- To find out the significant topics and subject domain of the LDA based research;
- To determine the publication sources, where publications are most concentrated;
- To study the Collaboration Network of Institutions and Countries.

3. METHODOLOGY

The analysis of the present study is based on the Scopus bibliographic database from Elsevier. In general, the data for bibliometric analysis is extracted based on searching the keyword in Title, Keywords, Abstract, or a combination of all three. Previous research on SVM has also used a similar methodology[19]. But this is not feasible as per the objectives of the present study due to the issues discussed below.

Suppose we use only the phrase "Latent Dirichlet Allocation.", in that case, there will be possibilities of missing documents that have used the word LDA.

Moreover, if we use both "Latent Dirichlet Allocation" and LDA in that case, there will be the possibility of inclusion of other documents which have used the keyword LDA in different contexts like Laser Diffraction Analysis (LDA), Linear Discriminant Analysis (LDA), etc. That is why when we searched with Title-Abstract-Keywords (LDA OR "Latent Dirichlet Allocation"), there were more than 32,000 results retrieved. Also, sometimes authors use attractive titles in the article and discuss the algorithm used in the main body or methodology of the article.

Because of these many reasons, we have analysed the citations of the original research paper entitled "Latent Dirichlet Allocation". Citations are the important criterion of examining the impact and are considered popular methods for identifying core documents[25]. Though, citation-based analysis has limitations in exploring the reason for citing that study. Still, citation-based research is commonly used to portray a domain's scientific progress. Still, citation-based researches are commonly used to portrait the scientific progress of a given domain.

At first, we have searched the Scopus by the query Title ("Latent Dirichlet Allocation") and sorted the results by the highest cited documents. The citations of document "Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. The Journal of Machine Learning Research, 3, 993-1022." was selected. Then 18841 documents published in the English language were selected after excluding publications of 2021 and 1984.

Later, it was observed that the document of type Conference Paper, Article, Book Chapter, and Review compositely represent 99.33 per cent of the publications. Therefore, finally, 18715 publications were selected for the analysis consisting of Conference Paper (11041), Article (7016), Book Chapter (351), Review (307).

The Scopus provides bibliographic, citation, and keywords information for 2000 results and only citation information for more than 2000 results. The objectives of our study will suffice only after the analysis of bibliographic and keywords information of each publication. For this, we have downloaded year-wise multiple files of less than 2000 results consisting of data on bibliographic, citation, and keyword. Then all these files were clubbed together into one file of 18715 results using windows command prompt.

The eighteen-year data from 2003 to 2020 were grouped into three classes of six-year each, i.e., 2003 to 2008, 2009 to 2014, and 2015 to 2020. The open-source software R was used to visualise the year-wise distribution of publications. The co-occurrence network of keywords and co-authorship network of authors, institutions, and countries were analysed using VOSviewer software[26].

4. RESULTS

A total of 18715 documents of English language published in Conference Proceedings, Journals, Book Chapters have cited the Latent Dirichlet Allocation. Table 1 shows the annual distribution of publications cited LDA. The journey of LDA citing publications starts from the year 2003 when the LDA was first published in the Journal of Machine Learning and Research. However, it begins with 05 publications in the year 2003. After that, rapid interest was observed in LDA, with the number of publications going to more than four times of publications in 2004. The annual average publication citing LDA was approx. 1040 (~1039.72). There was an upward trend in referring LDA in the publications throughout the period under study.

Table 1. Annual distribution of LDA cited publications

Year	TP	Cumulative	%	Cumulative %	Six-Year TP
2020	2350	2350	12.56	12.56	
2019	2378	4728	12.71	25.26	
2018	2192	6920	11.71	36.98	
2017	2036	8956	10.88	47.85	12666
2016	2073	11029	11.08	58.93	
2015	1637	12666	8.75	67.68	
2014	1455	14121	7.77	75.45	
2013	1287	15408	6.88	82.33	
2012	1051	16459	5.62	87.95	
2011	785	17244	4.19	92.14	5559
2010	575	17819	3.07	95.21	
2009	406	18225	2.17	97.38	
2008	190	18415	1.02	98.4	
2007	126	18541	0.67	99.07	
2006	87	18628	0.46	99.54	
2005	60	18688	0.32	99.86	490
2004	22	18710	0.12	99.97	
2003	5	18715	0.03	100	

The first six years from 2003 to 2008 had a minimum number of publications (490, 2.62 %), with an annual growth rate of 130.67 per cent. The highest number of publications in this period was in 2008 (190,1.01 %).

The next six years from 2009 to 2014 had rapid publication growth with almost eleven times of the previous period. The approx. 30 per cent (~29.70) of the total publications with an annual average growth rate of 43.54 per cent were published in this period (5559, 29.70 %). The initial three years of this period had publications under 1000, with 406 in 2009, 575 in 2010, and 785 in 2011. The next three years had publications more than 1000. The year 2012 was the first year after 2003 in the citation journey of LDA when the cited publications crossed the 1000 mark (1054, 5.63 %). The 1287 publications cited LDA in 2013 and 1455 publications in 2014.

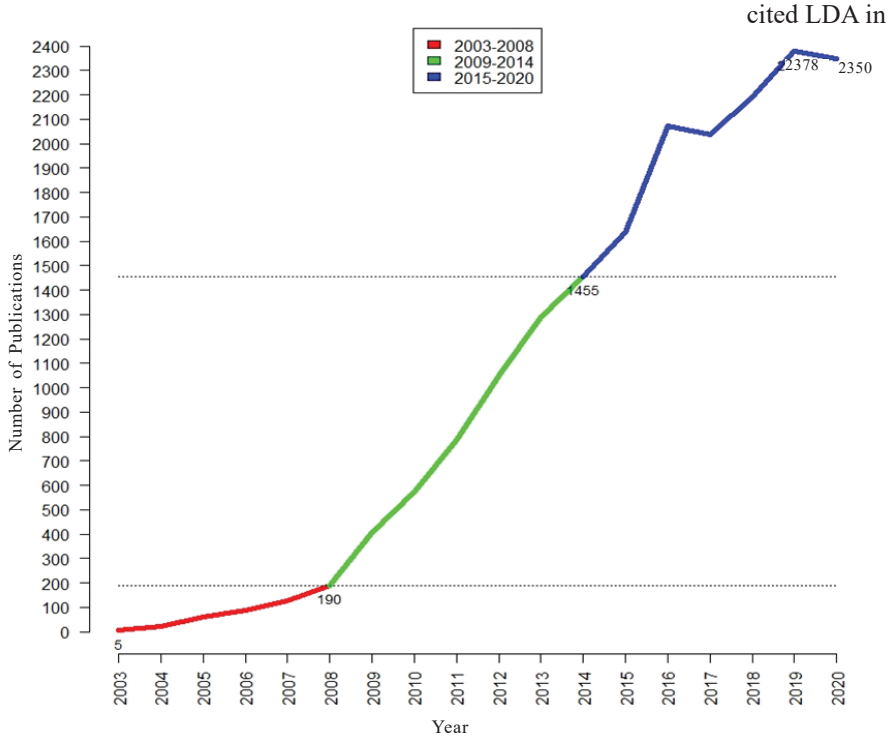

Figure 1. Year-wise distribution of publication cited LDA.

The interest in LDA accelerated from 2015 to 2020, with 68 %(~67.69) of total publications was published in this period (12666). In this period, 2015 observed publications less than 2000, with 1637 publications (8.75 %). From 2016 to 2020, the publications were above 2000, with a combined total of 11029 publications(58.93 %).

The highest number of publications in the entire period from 2003 to 2020 was published in 2019 (2378,12.71 %).

Figure 1 shows the year-wise distribution of publication cited LDA. The three intervals of six years are represented by Red (2003-2008), Green (2009-2014), and Blue (2015-2020). The upward trends in referring LDA can be seen from the Fig. 1. The red line was the minimal publication zone with the exponential growth from 2008. The green and blue lines show the increased interest in LDA, with the highest number in 2019 with 2378 publications.

Table 2 shows the publication venues that cited LDA with a minimum of 100 publications.

Table 2. Publication venues cited LDA with minimum 100 publications

Source	TP	Cumulative	%	Cumulative %	Source Type
Lecture Notes In Computer Science Including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics	1491	1491	7.97	7.97	Book Series
ACM International Conference Proceeding Series	378	1869	2.02	9.99	Conference Proceeding
Coeur Workshop Proceedings	284	2153	1.52	11.51	Conference Proceeding
International Conference On Information And Knowledge Management Proceedings	248	2401	1.33	12.83	Conference Proceeding
Communications In Computer And Information Science	226	2627	1.21	14.04	Book Series
Proceedings Of The ACM SIGKDD International Conference On Knowledge Discovery And Data Mining	220	2847	1.18	15.22	Conference Proceeding
IEEE Access	211	3058	1.13	16.34	Journal
Expert Systems With Applications	126	3184	0.67	17.02	Journal
Neurocomputing	126	3310	0.67	17.69	Journal
IEEE Transactions On Knowledge And Data Engineering	124	3434	0.66	18.35	Journal
Advances In Intelligent Systems And Computing	118	3552	0.63	18.98	Book Series
Knowledge-Based Systems	117	3669	0.63	19.61	Journal
Plus One	110	3779	0.59	20.20	Journal
Journal Of Machine Learning Research	109	3888	0.58	20.78	Journal

Table 3. Type of subject cited LDA

Subject	TP	%
Computer Science	15035	80.34
Engineering	3771	20.15
Mathematics	3766	20.12
Social Sciences	2560	13.68
Decision Sciences	1743	9.31
Business, Management and Accounting	1048	5.6
Arts and Humanities	892	4.77
Medicine	597	3.19

Table 4. Most frequent authors cited LDA

Authors	TP	%
Xing, E.P.	67	0.36
Blei, D.M.	61	0.33
Tang, J.	58	0.31
Han, J.	52	0.28
Phung, D.	50	0.27
Xiong, H.	48	0.26
Boyd-Graber, J.	47	0.25
Venkatesh, S.	47	0.25
Zhu, J.	47	0.25
Dascalu, M.	44	0.24
Trausan-Matu, S.	44	0.24
Carin, L.	42	0.22
Moens, M.F.	40	0.21
Chen, E.	39	0.21
Takasu, A.	38	0.2
Li, T.	37	0.2
Chien, J.T.	36	0.19
Chua, T.S.	35	0.19
Ding, Y.	35	0.19
Iwata, T.	35	0.19
Li, J.	35	0.19
Liu, B.	35	0.19
Lo, D.	33	0.18
Li, Y.	32	0.17
Orchid, M.	32	0.17
Xu, C.	32	0.17
Zhai, C.X.	32	0.17
Chen, B.	31	0.17
Gatica-Perez, D.	31	0.17
He, Y.	31	0.17
McCallum, A.	31	0.17
Mimno, D.	31	0.17
Liu, H.	30	0.16

The result shows that the Book Chapter published in the three Book Series constituted 9.81 per cent of total publications (1835). Most publications were from the Book Series titled "Lecture Notes In Computer Science Including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics "(1491, 7.97 %)". The four Conference Proceedings constituted 6.05 per cent of total publications.

These proceedings were "ACM International Conference Proceeding Series" (378,2.02 %)," Coeur Workshop Proceedings" (284,1.52 %), "International Conference On Information And Knowledge Management Proceedings" (248, 1.33 %), and "Proceedings Of The ACM SIGKDD International Conference On Knowledge Discovery And Data Mining" (220, 1.18 %). The analysis found that each of the seven journals has more than 100 publications cited LDA.

Table 2 also presents the top seven journals, with a minimum of 100 publications that refereed LDA. The majority of the journals were from Engineering and Computer Science domain.

As per Scopus, the LDA-cited publications were distributed among 26 subject categories. Table 3 presents the top eight subject categories having minimum publications of 500. It was observed from the results that the majority of the publications are from computer science and engineering. However, the interest from Social Science, Arts, Business Management was also reported.

Table 4 shows the top authors who have cited LDA frequently. The author Xing, E.P. tops the list with 67

Table 5. Most frequent institutions cited LDA

Affiliation	TP	%
Chinese Academy of Sciences	542	2.9
Tsinghua University	512	2.74
Carnegie Mellon University	463	2.47
University of Illinois Urbana-Champaign	257	1.37
Peking University	237	1.27
Ministry of Education China	215	1.15
Zhejiang University	214	1.14
Wuhan University	211	1.13
Beijing University of Posts and Telecommunications	208	1.11
University of Chinese Academy of Sciences	208	1.11
Beihang University	182	0.97
Shanghai Jiao Tong University	177	0.95
Microsoft Research	176	0.94
Massachusetts Institute of Technology	165	0.88
Stanford University	160	0.85
National University of Singapore	158	0.84
Institute of Automation Chinese Academy of Sciences	156	0.83
Pennsylvania State University	151	0.81

publications, followed by Blei, D.M., the author who introduced LDA with 61 publications.

The top 18 institutions with more than 150 publications are listed in Table 5. The institutions from China top the list in citing LDA, Chinese Academy of Sciences with 542 publications, followed by Tsinghua University with 512 publications. After China, the institutions from the United States had the most interest in citing LDA, Carnegie Mellon University with 463 publications, followed by the University of Illinois Urbana-Champaign with 257 publications. The majority of the institutions were from China, except two from the United States. It reflects that Chinese institutions were actively referring to LDA in their research.

Table 6 presents the list of 18 countries with more than 200 LDA citing publications. The United States tops the list with 6166 publications, 5220 publications in China, and 1137 publications in the United Kingdom. As discussed above, China and the United States had a much higher interest in referring to LDA than other countries.

Table 6. Most frequent countries cited LDA

Country	TP
United States	6166
China	5220
United Kingdom	1137
Japan	982
Germany	865
India	822
Australia	798
Canada	719
Singapore	536
France	529
Italy	519
Hong Kong	499
South Korea	492
Spain	328
Taiwan	326
Netherlands	309
Switzerland	273
Russian Federation	204

Figure 2 shows the collaboration between different institutions that had cited LDA. This collaboration network was visualised using VOSViewer. Each circle, also known as nodes, represents the institution and the edge connecting one node to another shows the collaboration between the two institutions. The thickness of the edge defines the strength of the collaborations, i.e., the more the collaboration between the institutions thicker the edge will be between them. The analysis showed that Carnegie Mellon University(CMU) has the most collaborations with other institutions, including universities like Princeton, California, Harvard, Tsinghua, etc., and popular research labs like Google, Amazon, Oracle, etc. The other institutions after CMU are Singapore Management University, University of Chinese Academy, Microsoft Research, etc. The result shows that these institutions are continuously referring to LDA.

The previous section presented the collaboration network of institutions cited LDA. As one country have multiple institutions, it's essential to investigate the country-wise network analysis to identify the inter-region collaboration. Figure 3 illustrates the collaboration networks of countries.

It can be seen from the Fig. 3 that the thickness of the edge between the United States and China is much higher than any other edge. This reflects the strong collaboration between the two countries and jointly publishing research.

It is also essential to know the researchers' aim referring the LDA in their study. The one way of knowing this can be achieved with the help of author keyword analysis. The author's keywords are the words mentioned by the author to portrait the broader area of research. Research hot-spots can be understood by examination of keyword co-occurrence[27].

The Network of keyword co-occurrence analysis was plotted using Vosviewer. Each node represents the keyword, and the edge between two nodes represents the co-occurrence of those two keywords.

The node's size represents the frequency of keyword, which means the larger the size of the node, the more the frequency of keyword. The thickness of the edge shows a more substantial relationship of keyword co-occurrence.

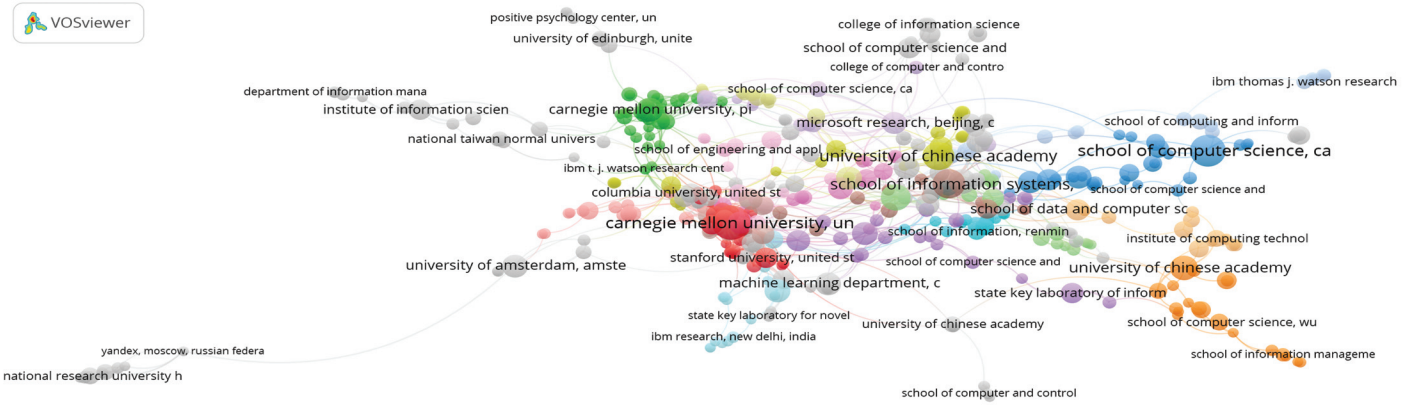

Figure 2. Collaboration network of institutions.

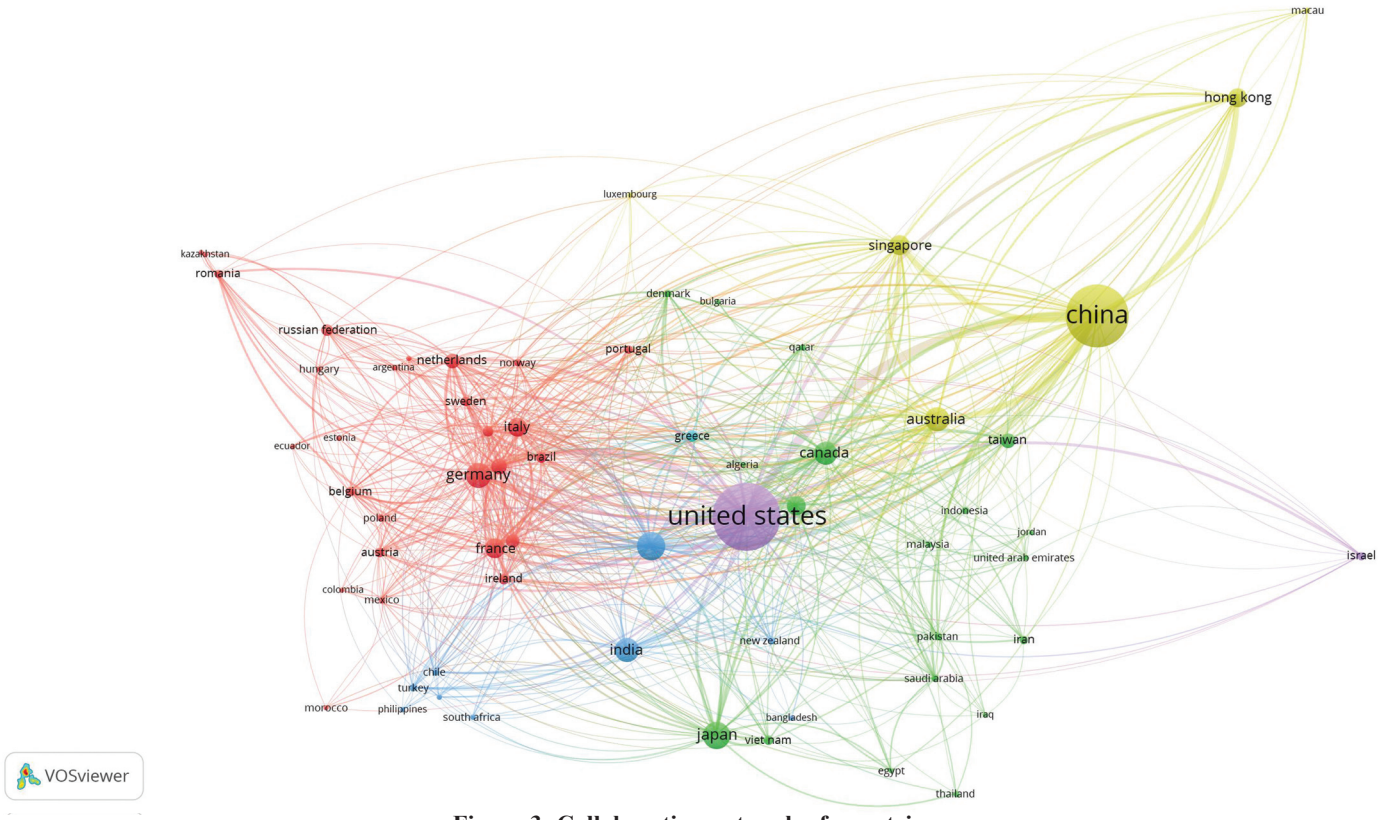

Figure 3. Collaboration network of countries.

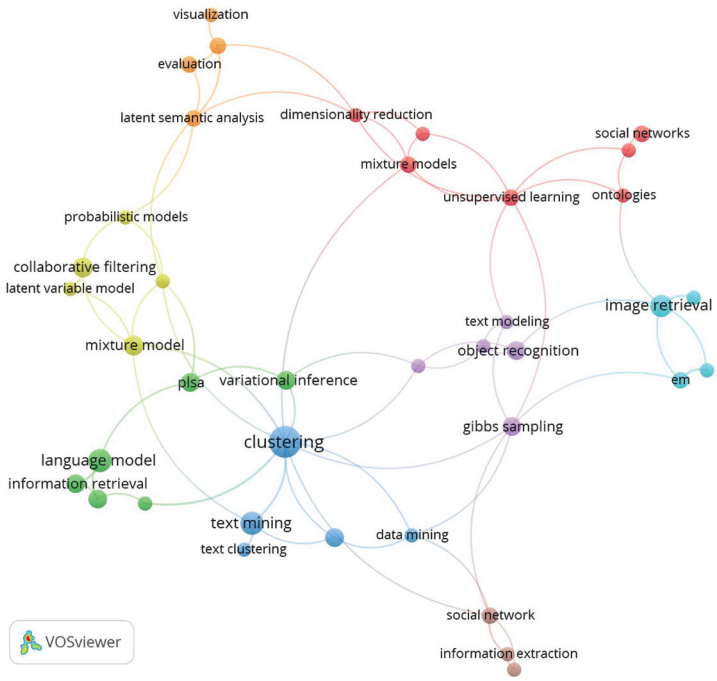

Figure 4. Co-occurrence network of author keywords from 2003 to 2008.

As we have seen above, there was not much growth in the first six years and then upward growth in the next six years from 2009 to 2014, and later in 2015 to 2020. Similarly, it is essential to visualise the evolution of themes of LDA researches over a specific time. Therefore, the co-occurrence of author keywords was examined in three different periods, i.e., 2003 to 2008, 2009 to 2014, and 2015 to 2020. The minimum occurrence of keywords was set to 3. The typical author keywords like Latent Dirichlet Allocation, LDA, topic models, topic modeling, topic model, topic detection were excluded for presenting main research topics.

Figure 4 shows the co-occurrence network of author keywords from 2003 to 2008. It showed that all the keywords were divided into eight clusters represented by different colors. The keyword "clustering" has the highest frequency and link strength of 15. It is not surprising that the keyword "clustering" is in the center of the graph as both clustering and LDA are unsupervised machine learning algorithms. The other nodes of bigger size during 2003-2008 after "clustering" are "language model," "text mining," and "image retrieval ." It showed that most of the studies were related to analysing text data sets.

Figure 5 presents the keyword co-occurrence from 2009 to 2014. A total of 12 clusters were identified in VosViewer, comprising 104 keywords. The cluster in green color with the node "image retrieval" had the highest total link strength of 42. The "image retrieval" had the thick edge with "image annotation" and "automatic image annotation," reflecting the use of LDA as the technique for the analysis of images. The big nodes of other clusters were "unsupervised learning" (Freq 20, Link Strength 22), "clustering" (Freq 19, Link Strength 21), "information retrieval" (Freq 19, Link Strength 28), and "text mining" (Freq 18, Link Strength 20). It showed that scholars' interest is progressively increasing towards text mining.

The cluster "collaborative filtering" having an edge with "social media," "sentiment analysis," "opinion mining,"

"Twitter" shows the increasing interest in LDA for analysis of social media data.

Figure 6 shows the keyword co-occurrence network from 2015-2020. The Network contains 22 clusters represented by different colors. The most significant node represented by blue color was "text mining" (Freq 621, Total Link Strength 1421). The other keyword of this cluster includes "Bibliometric," "Bibliometric Analysis," "Citation Network," "Informetric," "Scientometrics," reflecting the latest trends of using topic modeling techniques in bibliometric analysis. The trend of bibliometric research based on text mining and topic modeling can be seen in the literature of Journals related to Library and Information Sciences. The keyword "machine learning" (Freq 536, Total Link Strength 1346) had the second most occurrence after "text mining" followed by "Social Media" (Freq 468, Total Link Strength 1066), "Natural Language Processing" (Freq 451, Total Link Strength 1095) and "Sentiment Analysis" (Freq 419, Total Link Strength 1019). It showed that there had been growing interest in social media-based studies. Interestingly, the sixth most frequent occurrence of the keyword "Twitter" shows the analysis of the Twitter dataset. One reason for expansion of LDA use in different research themes can be the availability of tools for extracting datasets through various APIs and for analysis like Mallet[28], LDA Analyser[29], Genism[30], Matlab Topic Modelling[31], Yahoo_LDA[32], LDA in R[33].

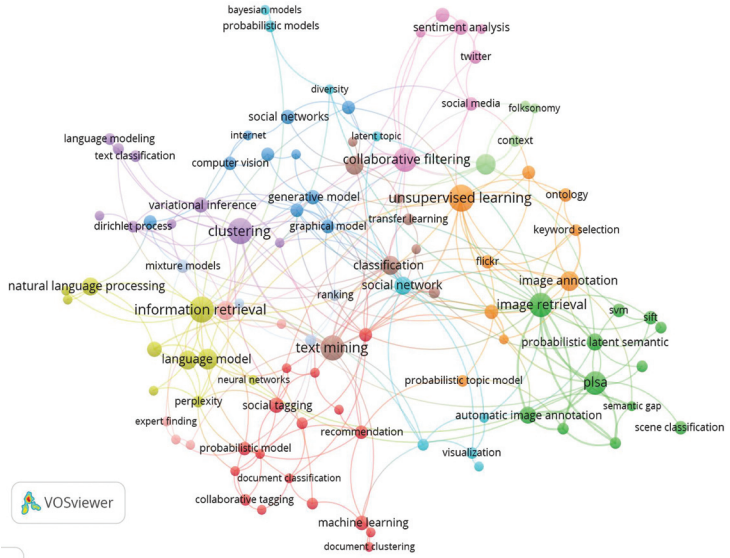

Figure 5. Co-occurrence network of author keywords from 2009 to 2014.

5. LIMITATIONS

The results of the present study may deviate due to some limitations. Firstly, the use of citation of LDA as data may limit the current study by considering additional studies which might not have used precisely the LDA for the analysis. Secondly, the selection of Scopus as the data source, as every database has its strength and weakness[34]. The results may deviate from analysis of other databases like Web of Science,

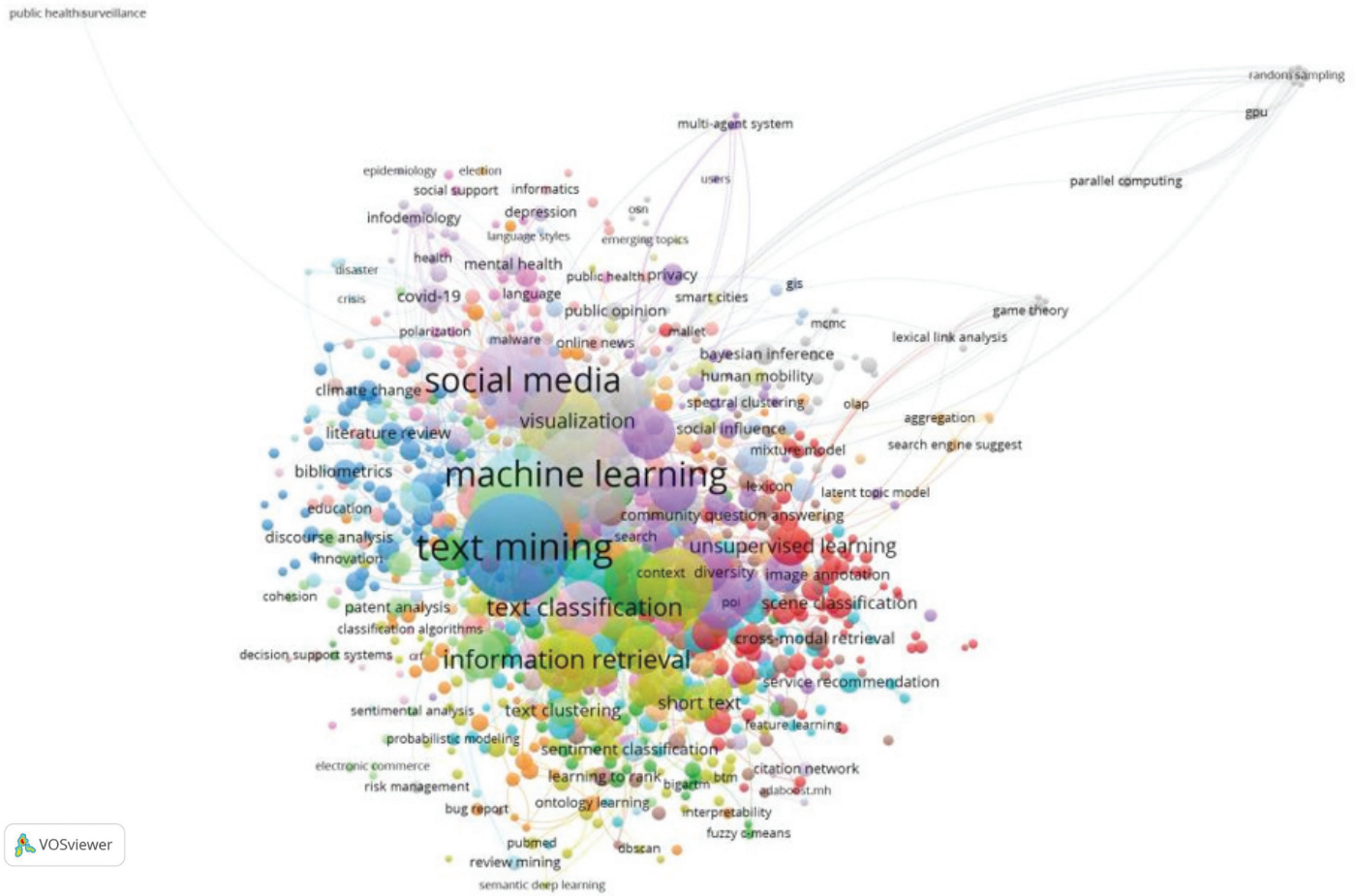

Figure 6. Co-occurrence network of author keywords from 2015 to 2020.

etc. Thirdly, the time frame of data extraction from Scopus. The bibliographic databases face the challenge of delay in the publication by the publishers. The data for the present research was extracted from Scopus on 29th May 2021; the number of publications may increase if data is extracted later in the year for the same search selection criteria.

6. CONCLUSION

The present study sketched the intellectual development of LDA by analysis of 18715 publications that have cited LDA. This study extends earlier research, which was limited to a small core of LDA research.

The results highlight not much interest in the initial Six years (2003 to 2008). However, from 2008, exponential growth was observed, with an upward trend of scientific interest in LDA in the next two six years block (2009 to 2014, 2015 to 2020). It shows that LDA has become popular in the last decade. The majority of the research activity was from Computer Science, followed by Engineering and Mathematics. However, significant interest was also reported from social science. The analysis showed that the USA and China dominated the number of publications and collaboration networks with other countries. The source of type book series titled "Lecture Notes In Computer Science Including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics" had published the highest 1491 publications. The majority of the researches were published in Book Series followed by Conference Proceedings and Journals. The co-occurrence network analysis of keywords shows that the published literature addresses the topics related to "text mining," "machine learning," "Social Media," etc. The result showed a growing interest in LDA-based bibliometric analysis for modeling the topics of scientific literature. The present study considered citations of an original published source of LDA as the data source for bibliometric analysis. Therefore, scholars can further test this approach on different algorithms. This research serves as a comprehensive guide for researchers to stay updated with the development and applications of the LDA.

REFERENCES

1. Blei, D.M.; Ng, A.Y. & Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 2003, **3**, 993-1022.
2. Figuerola, C.G.; Marco, F.J.G. & Pinto, M. Mapping the evolution of library and information science (1978–2014) using topic modeling on LISA. *Scientometrics*, 2017, **112**(3), 1507-1535.
   doi: 10.1007/s11192-017-2432-9.
3. Choi, S. & Seo, J. An exploratory study of the research on caregiver depression: using bibliometrics and LDA topic modeling. *Issues Ment. Hesalth Nurs.*, 2020, **41**(7), 592-601.
   doi: 10.1080/01612840.2019.1705944.
4. Sinha, A.; Li, Y. & Bauer, L. What you want is not what you get: Predicting sharing policies for text-based content on facebook. *In* Proceedings of the 2013 ACM workshop on Artificial intelligence and security, November 2013,pp. 13-24.
   doi: 10.1145/2517312.2517317.
5. Wang, Y.C.; Burke, M. & Kraut, R.E. Gender, topic, and audience response: An analysis of user-generated content on Facebook. *In* Proceedings of the SIGCHI conference on human factors in computing systems, April 2013, pp. 31-34.
   doi: 10.1145/2470654.2470659.
6. Majumdar, A. & Bose, I. Do tweets create value? A multi-period analysis of Twitter use and content of tweets for manufacturing firms. *Int. J. of Prod. Eco.*, 2019, **216**, 1-11.
   doi: 10.1016/j.ijpe.2019.04.008.
7. Ikegami, Y.; Kawai, K.; Namihira, Y. & Tsuruta, S. Topic and opinion classification based information credibility analysis on twitter. *In* 2013 IEEE International Conference on Systems, Man, and Cybernetics, October 2013, pp. 4676-4681.
   doi: 10.1109/SMC.2013.796.
8. Negi, S.; Balasubramanyan, R. & Chaudhury, S. Discovering user-communities and associated topics from YouTube. *In* 2014 22nd International Conference on Pattern Recognition, August 2014, pp. 1958-1963.
   doi: 10.1109/ICPR.2014.342.
9. Vlachos, E. & Tan, Z.H. Public perception of android robots: Indications from an analysis of YouTube comments. *In* 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), October 2018, pp. 1255-1260.
   doi: 10.1109/IROS.2018.8594058.
10. Vasconcelos, M.; Pereira, E.; Guimarães, S.; Ribeiro, M. H.; Melo, P. & Benevenuto, F. Analyzing YouTube Videos Shared on Whatsapp in the Early COVID-19 Crisis. *In* Proceedings of the Brazilian Symposium on Multimedia and the Web, November 2020 ,pp. 25–28.
    doi: 10.1145/3428658.3431090.
11. Hoang, D.T.; Nguyen, N.T.; Phan, H.T. & Hwang, D. An approach for recommending group experts on question and answering sites. In Modern Approaches for Intelligent Information and Database Systems, edited by Andrzej Sieminski, Adrianna Kozierkiewicz, Manuel Nunez & Quang Thuy Ha . Springer, 2018, pp. 27-37.
    doi: 10.1007/978-3-319-76081-0_3.
12. Maity, S.K.; Kharb, A. & Mukherjee, A. Analyzing the linguistic structure of question texts to characterise answerability in quora. *IEEE Trans. Comput. Soc. Syst.*, 2018, ***5***(3), 816-828.
    doi: 10.1109/TCSS.2018.2859964.
13. Ali, R.H. & Linstead, E. Modeling topic exhaustion for programming languages on StackOverflow. *In* The 32nd International Conference on Software Engineering and Knowledge Engineering, SEKE 2020, KSIR Virtual Conference Center, July 9-19, 2020, USA. 2020. pp. 400-405.
    doi: 10.18293/SEKE2020-107.
14. Bandeira, A.; Medeiros, C.A.; Paixao, M. & Maia, P.H. We need to talk about microservices: An analysis from the discussions on StackOverflow. *In* 2019 IEEE/ACM 16th International Conference on Mining Software

Repositories (MSR), May 2019, pp. 255-259.
doi: 10.1109/MSR.2019.00051.

15. Chen, S.; Vidden, C.; Nelson, N. & Vriens, M. Topic modelling for open-ended survey responses. *Appl. Mark. Anal.*, 2018, **4**(1), 53-62.
doi: 10.1080/2573234X.2019.1590131.
16. Baburajan, V.; e Silva, J.D.A. & Pereira, F.C. Opening up the conversation: Topic modeling for automated text analysis in travel surveys. *In* 2018 21st International Conference on Intelligent Transportation Systems (ITSC), November 2018, pp. 3657-3661.
17. Pritchard, A. Statistical bibliography or bibliometrics. *J. Doc.*, 1969, **25**(4), 348-349.
18. Dwivedi, S.; Kumar, S. & Garg, K.C. Scientometric profile of organic chemistry research in India during 2004–2013. *Current Science*, 2015,869-877.
doi: 10.18520/v109/i5/869-877.
19. Yu, D.; Xu, Z. & Wang, X. Bibliometric analysis of support vector machines research trend: a case study in China. *Int. J. Mach. Learn. Cybern.*, 2020, **11**(3), 715-728.
doi: 10.1007/s13042-019-01028-y.
20. Li, X. & Lei, L. A bibliometric analysis of topic modelling studies (2000–2017). *J. Inf. Sci.*, 2021, **47**(2), 161-175.
doi: 10.1177/0165551519877049.
21. Jelodar, H.; Wang, Y; Yuan, C.; Feng, X.; Jiang, X.; Li, Y. & Zhao, L. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimed. Tools Appl.*, 2019, **78**(11), 15169-15211.
doi: 10.1007/s11042-018-6894-4.
22. Chen, T.H., Thomas, S.W., & Hassan, A.E. A survey on the use of topic models when mining software repositories. *Empirical Software Eng.*, 2016, **21**(5), 1843-1919.
doi: 10.1007/s10664-015-9402-8.
23. Daud, A.; Li, J.; Zhou, L. & Muhammad, F. Knowledge discovery through directed probabilistic topic models: a survey. *Front. Comput. Sci. China*, 2010, **4**(2), 280-301.
doi: 10.1007/s11704-009-0062-y.
24. Sun, X.; Liu, X.; Li, B.; Duan, Y.; Yang, H. & Hu, J. Exploring topic models in software engineering data analysis: A survey. *In* 2016 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), May 2016, pp. 357-362.
doi: 10.1109/SNPD.2016.7515925.
25. Haridasan, S. & Kulshrestha, V.K. Citation analysis of scholarly communication in the journal Knowledge Organization. *Library Review*, 2007, **56**(4), 299-310.
doi: 10.1108/00242530710743525.
26. Van Eck, N.J. & Waltman, L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 2010, **84**(2), 523-538.
doi: 10.1007/s11192-009-0146-3.
27. Li, H.; An, H.; Wang, Y.; Huang, J. & Gao, X. Evolutionary features of academic articles co-keyword network and keywords co-occurrence network: Based on two-mode affiliation network. *Physica A: Stat. Mech. Appl.*, 2016, **450**, 657-669.
doi: 10.1016/j.physa.2016.01.017.
28. McCallum, A.K. Mallet: A machine learning for language toolkit, 2002. *http://mallet. cs. UMass. edu*. (Accessed on 26 April 2021).
29. Zou, C. & Hou, D. LDA analyser: A tool for exploring topic models. *In* 2014 IEEE International Conference on Software Maintenance and Evolution, September 2014, pp. 593-596.
doi: 10.1109/ICSME.2014.103.
30. Rehurek, R. & P. Sojka. Gensim-statistical semantics in python. 2011. https://www.fi.muni.cz/usr/sojka/posters/rehurek-sojka-scipy2011.pdf (Accessed on 26 April 2021).
31. Steyvers, M. & T. Griffiths, Matlab topic modeling toolbox 1.4, 2011. https://www.mathworks.com/help/textanalytics/gs/getting-started-with-topic-modeling.html. (Accessed on 26 April 2021).
32. Ahmed, A.; Aly, M.; Gonzalez, J.; Narayanamurthy, S. & Smola, A.J. Scalable inference in latent variable models. In Proceedings of the fifth ACM international conference on Web search and data mining, February, 2012. pp. 123-132.
doi: 10.1145/2124295.2124312.
33. Chang, J. lda: Collapsed Gibbs sampling methods for topic models. 2011, R. https://rdrr.io/cran/lda/#:~:text=Implements%20latent%20Dirichlet%20allocation%20(LDA,Gibbs%20sampler%20written%20in%20C. (Accessed on 26 April 2021).
34. Falagas, M.E.; Pitsouni, E.I.; Malietzis, G.A. & Pappas, G. Comparison of PubMed, Scopus, Web of Science and Google Scholar: Strengths and weaknesses. *FASEB J.* 2008, **22**(2), 338-42.
doi: 10.1096/fj.07-9492lsf.

CONTRIBUTORS

Mr Mohit Garg is working as Assistant Librarian at Central Library, Indian Institute of Technology, New Delhi. His areas of interest are Application of ICT, Research Methodology, Information Retrieval, Data Science, and Machine learning.
He contributed to conceptualizing the present study, collection of related literature, and methodology

Ms Priya Rangra is a Research Scholar in Department of Library and Information Science, Central University of Himachal Pradesh. Her research interests include Knowledge Organisation, Information Seeking Behaviour, Bibliometrics, Accessible Information, and Quantitative Analysis.
Her contribution to the current study is the data analysis and preparation of the final draft of the paper.