DESIDOC Journal of Library & Information Technology, Vol. 44, No. 2, March 2024, pp.114-123, DOI : 10.14429/djlit.44.2.19312

© 2024, DESIDOC

Analysing Library and Information Science Articles Using Topic Modeling Approaches: A Study With Scopus Indexed Indian Journals

Debasis Majhi and Bhaskar Mukherjee*

Department of Library and Information Science, Banaras Hindu University, Varanasi- 221 005, India
**E-mail: mukherjee.bhaskar@gmail.com*

ABSTRACT

Identifying trends in research through co-citation or content analysis of journal contents is quite a common practice in LIS research. In this study, however, we proposed the Latent Dirichlet Allocation (LDA), a popular topic-modeling approach for identifying research trends of published articles in three scopus-indexed Indian LIS journals. A total of 1213 titles & their abstracts published between 2011 and 2022 have been considered. From these data, a corpus of frequently used 15 key phrases was identified from each journal using Count Vectorizer and then ten topics having higher coherence scores were extracted from each journal corpus using LDA techniques to understand to what extent these topics are different in these journals. The analysis of the study indicates that 'Library users' studies' especially in academic libraries; and 'bibliometric indicators for measuring research growth are a few common topics in these journals and, technological innovation; utilisation of electronic and print information resources; library management; or network analysis are some of the topics that are journal specific. From the t-SNE visualisation and pyLDAvis diagram, it was seen that the topics of *DJLIT* are significantly unique with discrete distributions than the other two journals. On analysing the growth of the top ten topics longitudinally, it was seen that research on digital libraries, analysing the global output, online search strategy, ranking universities, etc. are concurrent interests of research among researchers while academic library resources, including electronic resources and its use, open access are among diminishing research interests of authors. Since the topic-modeling approach can provide results devoid of bias, it can be used to identify research land scape longitudinally as well as obsolescence of topic in a domain.

Keywords: LDA; Topic modeling; Machine learning; Publication patterns-LIS-India; Trend analysis; LIS publications

1. INTRODUCTION

In the modern world, information is expanding at an unprecedented rate. Millions of new pieces of research are added daily to research databases worldwide, and most of the data is unstructured and unorganised. Retrieving dominant fields of research or plotting the research trend from these unstructured data is a tedious job. The principal motive behind any retrieval strategy is to return the smallest unit that includes the results based on the query. The retrieval precision, which represents the relevance of retrieved documents in response to a query, may not be effective for unstructured data, as it often lacks clear patterns or structures that can be easily extracted and analysed. Despite these challenges, machine learning and artificial intelligence techniques are emerging quite promising in processing and analysing unstructured data which seems difficult for other mechanical processes. Techniques such as sentiment analysis and topic modeling can be used to analyse unstructured data and reveal trends, patterns, and meanings in an effective manner.

Furthermore, when human indexers allot keywords manually in scientific research articles, the selected keywords may not always fully represent the content of the document. Authors typically select keywords to help others find their papers, but this process is often an afterthought and not extensively researched[1]. Other methods, such as expert opinions or quantitative analysis, can be used to predict the trend of a subject or research topic. However, topic prediction by experts may be biased because the panel and survey may not fairly reflect experts' opinions[2]. In this manner, it is desirable to utilise a machine-learning algorithm to reduce the role of human biasness[3].

1.1 Topic Modeling

Topic modeling is a powerful unsupervised technique for understanding text data or a statistical machine-learning technique used to uncover hidden topics and themes in a collection of documents[4]. By using this method, similar-content-related papers are automatically categorised into subjects. It can also be used to uncover and investigate long-term trends in research topics, find relevant issues, develop new research hypotheses,

Received : 20 July 2023, Revised : 10 January 2024
Accepted : 12 January 2024, Online published : 04 April 2024

identify relationships between documents and topics, and categorize documents based on their topics. The most well-known and frequently used topic modeling method, Latent Dirichlet Allocation (LDA) was employed to analyze large volumes of textual data and identify topics in the documents. It was originally formulated by Blei[5], *et al.* LDA assists in identifying topics in large collections of documents by extracting words that are likely to appear together and then grouping them into distinct subjects. The Python module Gensim helps create the LDA model by removing ambiguous syntax[6]. For this purpose, the present study will try to analyse the most frequently used phrases in publications through the LDA topic modeling approach to identify the main research areas of the selected three Indian LIS journals.

1.2 Indian LIS Journals

Although India has a considerable number of journals in the Library and Information Science (LIS) discipline, presently only three journals are indexed in the Scopus database. These journals are *Annals of Library & Information Science* (ALIS), *DESIDOC Journal of Library & Information Technology* (DJLIT), and *Journal of Scientometric Research* (JSCIRES). Of these three journals, except JSCIRES all publications belong to government-funded organisations and possess a distinguished publication history. On the other hand, JSCIRES (SJR 0.281), publishes 3 issues/year, started its journey in the year 2012, and was indexed in Scopus from 2019 onwards only. ALIS (SJR 0.221) was indexed in Scopus since 2011 and DJLIT (SJR 0.281) from 2012 onwards. The CiteScore of ALIS is 54 percentile, DESIDOC is 64 percentile, and JSCIRES is 58 percentile in Library & Information Science category of Scopus.

2. LITERATURE REVIEW

There are several studies on research patterns in LIS, conducted in the last few years. These earlier studies analysed, highlighted, and pointed out the quantity output, growth pattern, and main areas of research with broad and narrow subjects[7]. Bibliometric analysis has been applied to understand the LIS research trends in India[8]. But currently, topic modeling is being increasingly applied to discover hidden subject areas and identify the contemporary and future directions of a research field[9]. Topic modeling algorithms are also used for locating topics from large, unstructured collections of texts, by clustering words with similar meanings[10-12]. It has also been employed to extract themes from the abstracts of papers published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS) between 1991 and 2001, then classify the cold and hot subjects by time period[13]. In another research[12], 17,000 studies published in science were analysed. Further, Sun and Yin used topic modeling to analyze research patterns in the transportation industry by examining country participation from 22 prominent transportation journals[14]. Furthermore, topic modeling has been demonstrated by Mann[15], *et al.* by using topical n-grams on 300,000 publications from computer science to determine the impact of research papers. In 2020, 57 journal papers were gathered for a study on the evolution of information technology in CDW management from 2000 to 2019 to analyze trends over the past 20 years using the method of scientific mapping analysis[16]. In LIS, Sugimoto[17], *et al.* analysed the topic of 3,131 North American dissertations on LIS. Yan[18] discussed research dynamics and impact using topic modeling and citation data with 47,137 publications from JCR-indexed journals.

Krenn[19], *et al.* predicted research trends in semantic and neural networks with applications in quantum physics using machine learning techniques in 2019. The extraction of information from abstract or a large volume of articles using the machine learning process is useful for identifying current research trends[20]. In addition, these procedures can help researchers and academicians in conceiving nascent topic ideas for future research[21]. Each article provides crucial sources of theories and information, such as abstracts and keywords, however, extracting them is required for an effective analysis[22]. NLP techniques like Count Vectorizer and YAKE can be used to automatically extract the important keywords from a corpus[23]. Currently, the LDA topic modeling has been used for the analysis of computer linguistics[24], library and information management[25], economics[26], and other fields[27-29], to understand the research trends of a particular subject. As a result, topic modeling is being used extensively to analyse already published works containing bibliographic information such as research titles and abstracts to determine the research patterns across various academic disciplines. Therefore, in the present study, the topic modeling method has been utilised to discover LIS research patterns.

3. OBJECTIVES

The objectives of the present study are:

- To identify whether there is any considerable change in the number of articles published per volume and citation profile over a period in qualitative LIS journals in India.
- To identify frequently used words or phrases from the title and abstract of LIS journals.
- To identify the common and unique research themes of these journals.
- To identify the overall trend in research themes and the potential areas that need to focus more on research.

4. METHODOLOGY

4.1 Data Collection

This study quantitatively analysed Indian Scopus-indexed journals through content analysis focused on the title and abstract content. Although there are many ways to identify the top journals of any subject, we considered those Indian LIS journals that are indexed in the Scopus database. The query was performed in March 2023, and thus, the data obtained included all publications indexed in the Scopus database till 2022. The search string (Source title = DESIDOC Journal of Library and Information Technology) OR (Source title = Annals of Library and Information Studies) OR (Source title = Journal of Scientometric Research) was used. Initially, 1213 papers were retrieved, where 388 papers were retrieved from the *Annals of Library and Information Studies*, 639 papers from the *DESIDOC Journal of Library and Information Technolog*y, and 186 papers from the *Journal of Scientometric Research*. This dataset was exported in a ".CSV" file with title, abstract, and publication year fields.

4.2 Pre-processing

Unstructured data usually contains a lot of irrelevant information. Therefore, we started preprocessing the exported dataset to remove the unwanted, unnecessary information contained in the dataset. We used the *Natural Language Toolkit (NLTK)* Library for preprocessing the dataset. In this step, first, we removed any double spaces, numerals, and special characters from the text and changed it all to lowercase. After that, *tokenisation* was performed with the text corpus, and each word was considered as a single token. Then, using NLTK stop word English language, the stop words 'a', 'an', 'the', 'but', 'of', and 'in' that are inconsequential to the text are eliminated. Additionally, we expanded the NLTK stop-word list by developing a dictionary of stop-word lists. A recently created stop word list containing words that are irrelevant in scientific titles and abstracts (e.g. 'named', 'formerly', 'article', 'across', 'actually' etc.) was used and lemmatisation was performed to decrease word inflection or avoid distinct word feature forms.

4.3 Topic Model

In this study, the titles and abstracts of texts were extracted from 1213 articles to prepare the text corpus. Gensim Python Library was used for the Latent Dirichlet Allocation (LDA) topic modeling. Hyper parameter tuning was carried out on 75 % and 100 % corpus to find out the best model parameters. The LDA model was constructed using various values for the different k-number of topics, and the highest coherence (0.455806) values were chosen using the highest alpha and beta values of 0.91 with the 10 (ten) topics selected for a better result in the LDA model (Table 1). A sample view of the output is shown in Annexure-A and Table 4. We separately performed journal-wise topic modeling to understand the publication patterns of journals by fixing the number of topics to ten (10).Also, we have applied the Count Vectorizer procedures from the sklearn library and Ngram [ngram_range = (2,3)] for extracting the most used keywords from the corpus and the most significant 15 keywords are shown in Table 3.

Table 1. Coherence matrix using c_v coherence score

Validation_set	Topics	Alpha	Beta	Coherence score
100% Corpus	10	0.91	0.91	0.455806
100% Corpus	10	0.61	0.91	0.442392
100% Corpus	10	symmetric	0.91	0.438462
100% Corpus	10	0.31	0.91	0.432823
100% Corpus	10	0.61	0.61	0.42886
100% Corpus	10	0.01	0.61	0.428336
100% Corpus	10	0.01	0.91	0.425084

4.4 Visualisation

In this phase, we used the T-distributed Stochastic Neighbour Embedding (t-SNE) programme to visualise the publication data with default n_components= 2 and verbose= 1'. T-SNE enables the visualisation of the underlying local structure of high-dimensional data and identifies a data point's related neighbours in a low-dimensional map[28].

5. RESULTS

5.1 Trends in Publications

Table 2 gives the chronological distribution of Indian LIS publications that are indexed in the Scopus database. It was observed that for the last ten years or so there was not any considerable change in the number of articles published per issue. Each volume includes an average of 30 to 40 articles. The highest number of articles,i.e. 161 was published in 2021 and all three journals have a larger number of articles published in this year than the preceding year. At the individual level,the highest number of articles appeared in ALIS in 2014, DJLIT in 2012, and JSCIRES in 2022. However, citations per article were observed higher in different years for these three journals. Overall, the average citation per LIS article published in these three journals was 4.26 in the last ten years. However, DJLIT articles have a larger capacity of receiving citations than the other two journals. Comparing the citation profile of these three journals shows that of the total citations, a higher citation was received in ALIS for an article published in 2015 on internet of things (39) followed by scientometric analysis (37). On the other hand, the higher number of citations in the DESIDOC journal came from an article published in 2019 on the theme of plagiarism (37) followed by bibliometric analysis (35). In JSCIRES an article published in 2019 received 156 citations on the theme of bibliometric analysis using the R package followed by bibliometric analysis of machine learning research (13).

5.2 Significant Key Phrases in the Title and Abstract Fields

Table 3 displays the most frequently used keywords in the Title and Abstract fields in the selected three journals arranged in decreasing order. We have applied

Table 2. Trend in publications by year

Year	ALIS (quarterly)		DJLIT (six issues/year)		JSCIRES (three issues/year)		Total	
	Articles	CPP	Articles	CPP	Articles	CPP	Article	CPP
2011	36	8.53	-		-		36	8.53
2012	29	6.00	69	3.83	-		98	4.47
2013	27	5.44	66	5.05	-		93	5.16
2014	45	6.47	63	5.90	-		108	6.14
2015	38	3.66	54	7.17	-		92	5.72
2016	32	4.22	51	3.69	-		83	3.89
2017	32	4.19	60	5.90	-		92	5.30
2018	29	3.34	61	4.52	-		90	4.14
2019	18	3.83	54	3.93	36	4.83	108	4.21
2020	27	2.96	55	2.96	44	2.02	126	2.63
2021	43	0.67	59	1.03	59	0.75	161	0.83
2022	32	0.00	47	0.32	47	0.09	126	0.15

*Citation data is based on from publication year up to 2022, CPP=Average Citation per publication

Table 3. Frequency of significant key phrases in titles and abstracts of scientific articles

ALIS		DJLIT		JSCIRES	
Keywords	Frequency	Keywords	Frequency	Keywords	Frequency
University library	60	Library service	109	Bibliometric analysis	45
Open access	53	Library professional	102	Scopus database	29
Digital library	42	Open access	91	Open access	25
Library service	29	Academic library	66	Technology innovation	19
Impact factor	27	Digital library	60	Scientometric analysis	17
Academic library	26	Library website	40	Machine learn	16
Indian journal	22	High education	39	Network analysis	15
Library website	21	Public library	36	Artificial intelligence	15
Institutional repository	21	Scopus database	34	Impact factor	14
Library professional	20	Library user	33	Citation count	13
Communication policy	20	Library management	33	Social network	12
Scopus database	19	Global publication	33	Publication citation	11
Indian library	18	Authorship pattern	33	Data repository	11
Authorship pattern	17	Social networking	32	Sector innovation	10
Search engine	15	Service quality	32	Scientific production	10

the Count Vectorizer procedures from the sklearn library and Ngram [ngram_range = (2,3)] for extracting the most used keywords from the corpus.

The title and abstract are the most essential attributions that use precise words to represent the most important content of the article. As a result, the most frequently matched words are usually grouped from the title and abstract data.

It was observed that library service, especially academic library service is one of the most frequently used terms in the title and abstract field LIS articles. Key phrases like ‘open access’, and ‘Scopus database’ are only two terms that appeared in the title and abstracts corpus of all three journals. On the other hand, terms like authorship pattern’, ‘bibliometric analysis’, ‘digital library’, ‘institutional repository’, ‘library professionals’, ‘library services’ etc. are such terms which were journal specific. To see how such areas are translated into research topics or themes, we have applied LDA techniques.

5.3 Topic Representation for Selected Journals Using LDA Algorithms

Using the title and abstract of papers with a topic probability of more than 0.8 and by using the top 10 highly weighted terms for each topic, we labeled 10 topics for each journal. Each topic was given a name based on the words assigned as shown in Table 4. The terms that LDA has assigned for each journal are mentioned in Annexure- A.

Table 4. Extraction of topics from scopus indexed journals using LDA algorithms

ALIS	No. of articles	DJLIT	No. of articles	JSCIRES	No. of articles
Academic library users, use and user studies	139	Library and its users and their professional development	281	Bibliometric analysis of journals	39
Bibliometric indicator and research growth analysis	95	Bibliometric indicator and research growth analysis	171	Growth of institutional scientific publication	28
Citation analysis of psychological research	34	Uses of electronic and print Information resources	72	citation impact on research collaboration	24
Ontology-based semantic technology	27	Awareness of plagiarism in academia	31	Bibliometric indicator and research growth analysis	20
Digital education systems	23	Digital library/ smart library	26	Scientometric analysis of technology & innovation	18
Network patterns of university-industry collaboration	22	Growth of institutional scientific publication	22	Citation impact of science journals	15
Pharmacology research trends	19	Digital preservation system	17	Developing indicators for environment and health	12
Library classification and web retrieval	18	Citation analysis of bibliographic data sources	16	Technology and innovation in government platform	9
Role of UGC for LIS professional development	12	Open access consortium policy	14	Global research trend on cancer	7
Academic profile and career choice in LIS	11	Digital tools for digitising the document	14	The quality indicators of information technology in higher education	6

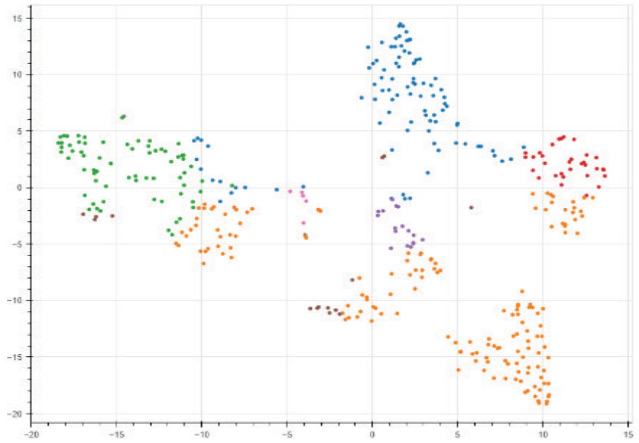


Figure 1. t-SNE clustering of ten LDA topics of ALIS.

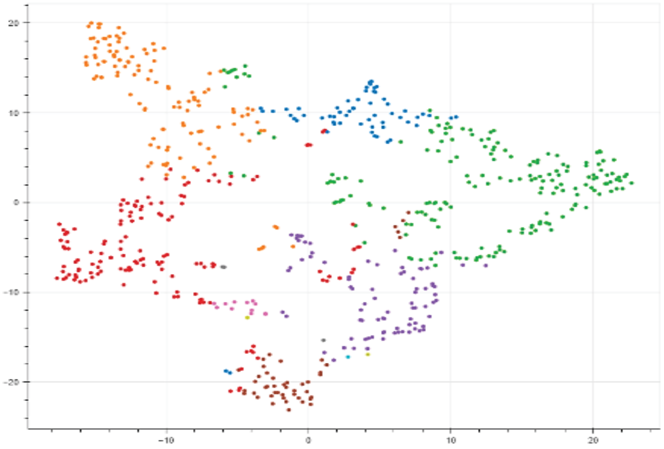


Figure 2. t-SNE clustering of ten LDA topics of DJLIT.

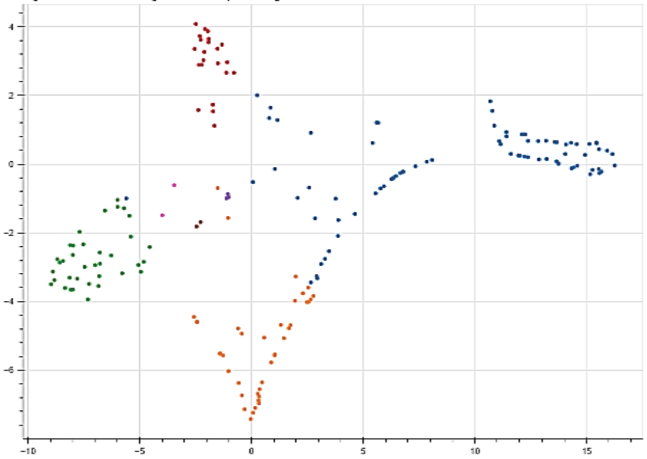
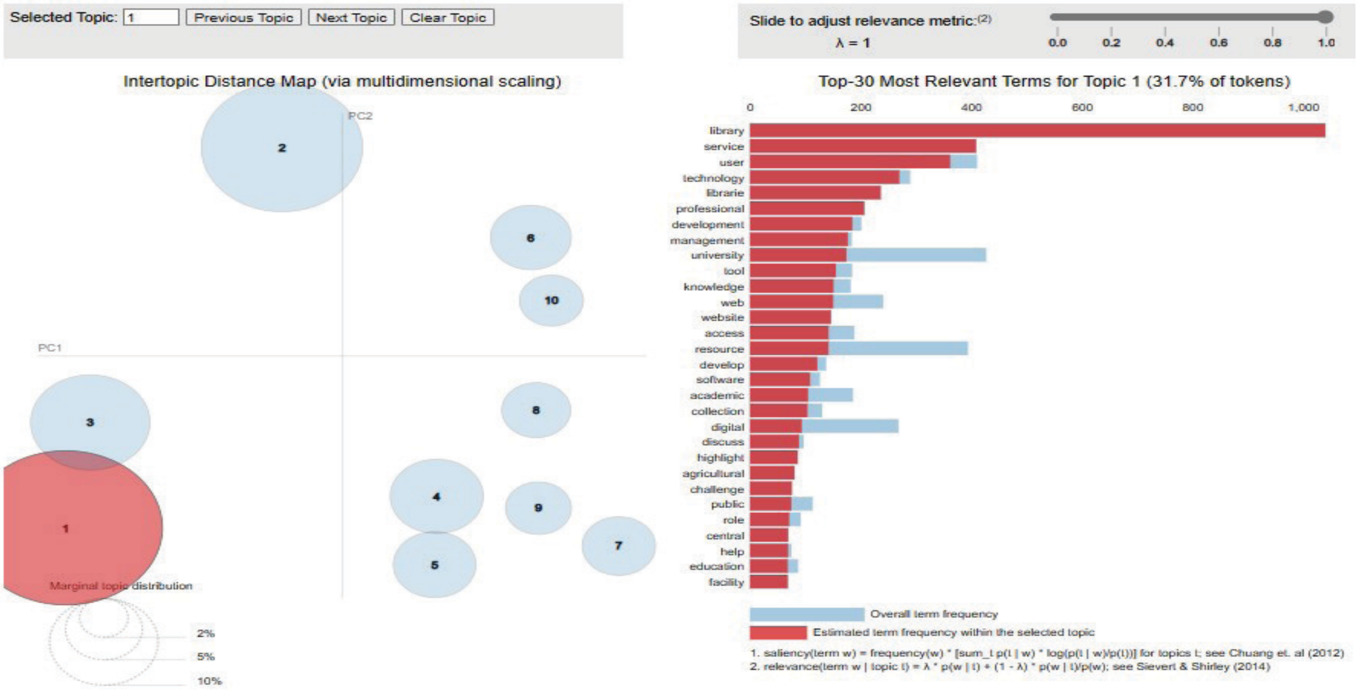


Figure 3. t-SNE clustering of ten LDA topics of JSCIRES.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 4. Topic visualisation based on title and abstract words using DJLIT data.

Table 5. Research trends in themes by years

T#	Theme	Total	AL	DJ	SC	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
1	Academic library services, Users, digital library	162	35	114	13	3	15	10	12	12	20	12	13	12	7	24	22
2	Global scientometric output, Indian science & citation	151	35	79	37	3	7	8	14	7	3	9	15	26	18	28	13
3	Indian library science professional evaluation	144	51	71	22	5	11	9	12	9	9	19	12	12	16	14	16
4	Bibliometric analysis, science mapping, collaboration	125	41	46	38	3	6	9	6	13	7	4	10	8	24	19	16
5	Online search strategy, ranking university,	137	30	86	21	6	13	14	12	8	11	5	6	13	15	18	16
6	Academic library resources usage	135	64	55	16	6	17	15	12	9	9	16	10	7	14	12	8
7	Trends in electronic resources in library	112	36	66	10	7	14	10	13	7	7	8	10	7	11	12	6
8	Impact of computer usage in libraries, challenges	107	31	54	22	0	6	12	11	11	7	6	3	11	9	18	13
9	Ontological framework	75	21	43	11	1	5	3	9	8	7	8	4	7	2	13	8
10	Growth of open access, and use of books by students	65	19	36	10	2	4	3	7	8	3	5	7	5	10	3	8

T#-Theme number, DJ-DJLIT, SC- JSCIRES, AL-ALIS

Here, topic labels represent an interesting subject, research perspective, or related areas of subjects. But this phase is very important to eliminate the fuzziness of the generated topic and to determine what the topic is all about. For example, in DJLIT corpus [“repository” + “internet” + “datum” + “digital” + “read” + “legal” + “day” + “reading” + “mobile phone” + “blog”] is about Digital Library/ smart library. Similarly, the first topic of JSCIRES [“publication” + “bibliometric” + “journal” + “science” + “topic” + “document” + ‘”network” + “datum”] discusses Bibliometric analysis of journals.

From the table, it is visible that the highest number of publications in ALIS are related to ‘Academic Library Users, Use and User Studies’ (139), whereas in DJLIT it was ‘Library and its Users and their Professional Development (281), and in JSCIRES it was ‘Bibliometric analysis of journals’ (39). Other themes are mentioned in the table.

In order to learn the relationship between weighted terms of a topic a d-dimensional map that reflects the similarity between high-dimensional terms, t-SNE visualisation techniques have been used. From Figure 1 to 3, it is clear that the t-SNE visualisation for DJLIT data is significantly unique with discrete distributions. In t-SNE, similar publications points can overlap with each other[30] but here, the overlapping of the subjects is very less. The same is also visualised in Figure 4 of the pyLDAvis diagram. Similarly, topic distribution in JSCIRES results is linear, and as shown in the t-SNE diagram, subject overlap is uncommon. This is the indication that the ten subjects that were generated are highly different from one another, with a high similarity distance. Each topic represents a new study area and creates distinctive research trends.

5.4 Growth of Research Themes in Indian LIS Journals

Finally, we have adjusted all the publications of the three journals in the time window 2011-2022 to analyse the overall trend in research themes. Table 5 displays the result. To understand overall trends, we have combined all the titles and abstract of the journals into a single .csv file and imported it in the python based LDA module in such a way that each title is assigned by a broad theme. By doing so through Gensim, the coherence score came 14.34. Since we have decided to group all themes into only 10 broad themes, each title was assigned by these 10 broad themes only.

As indicated in the Table 5, few themes are losing interest among authors and some other themes are gaining interest among researchers. Overall, researches on digital library, analysing global output, online search strategy, ranking universities etc. are concurrent interest of research among researchers whiles academic library resources, including electronic resources and its use, open access are among diminishing research interests of authors.

6. DISCUSSION

The basic intention of this study was to identify whether the scope of Indian LIS research as reported in the scholarly journals is diverse and multi-dimensional or it is restricted to fewer sub-fields only. We have selected three Scopus-indexed journals in this study since there are no journals indexed in the Web of Science core collection till now. A total of 1213 titles and their abstracts have been examined that appeared between 2011 to 2022. Since a considerable number of earlier researches have been conducted on co-citation analysis, content analysis, or author-keywords analysis, we applied a Python-based machine algorithm approach in this study. LDA topic modeling algorithms were applied to the whole title and abstract corpus and identified major key phrases that appeared in the title and abstracts of Indian LIS journal articles. By applying this technique, we framed 10 dominant areas of published research based on the high weight of ten keywords in each of these three LIS journals. Before applying the LDA approach for excavating results, we have seen the publication and citation trend of these journals.

From the publication and citation profile of the published article, we do not observe any significant difference however, from key phrase analysis we observed that sub-fields like user-studies and scientometrics are predominant fields of published research in the journals like ALIS and DJLIT & JSCIRES. While the largest amount of research has been reported in the areas of academic library users, use, and user studies, followed by bibliometric indicators and citation analysis in the ALIS the research activities on libraries, their users, and their professional growth, as well as the usage of print and electronic resources get leading importance in DJLIT. This may be because Indian libraries are moving towards completely digital, authors are still paying interest in discussing the policies, and strategies to ensure effective storage, organisation, and retrieval of digital content in libraries and challenges on digital preservation. On the other hand, JSCIRES has published most of the research in the areas of bibliometric analysis of journals, institution, and the influence of citations on research collaboration. The fact that relatively less research has been conducted in the fields of information retrieval, library management or library classification may be an indication that these fields are losing interest among researchers. However, an egligible amount of publications in the areas like smart library systems, machine learning techniques, or artificial intelligence, ontologies, metadata elements, or semantic web applications is a matter of concern.

Since scientometric analysis and citation analysis are quite common fields of published research in all three LIS journals of India, editors of these journals may now expect the authors to contribute advanced scientometric applications in research using state-of-the-art tools like R, python, or application of machine learning approach in scientometric analysis. From the trend analysis of titles from 2011 through 2022 indicates the changing dimention of LIS research. A more depth study may reveals how changes in research front are occuring under each sub-fields changes in research front are occuring.

7. CONCLUSION

This study introduces a novel method “topic modeling” for detecting emerging patterns using the LDA topic model algorithms and Count Vector keywords extraction. These methods provide several important advantages for trend analysis. First, only the title and abstract of publications are required. Second, because of the minimal human involvement in these processes, the outcome cannot be skewed. Third, the title and abstract are the most important tags of any article, thus extracting the most prominent terms or phrases will assist readers to comprehend the subject at hand. The performance is measured by the c_v coherence scores for LDA. Pre-processing is vital for both approaches since it efficiently decreases dimensionality and removes the majority of extraneous words from the input's unstructured text using the Stop words removal process. In addition, the t-SNE distribution of themes was carried out. According to Van der Maaten and Hinton[31], the t-SNE visualisation method places each data point on a two- or three-dimensional map to visualise high-dimensional data. The findings of this study indicate that academic library users, use and user studies, bibliometric indicator and research growth analysis, citation impact on research collaboration, utilisations of electronic and print information resources, digital library, network patterns of university-industry collaboration, open access policy in LIS education, scientometric analysis of technology & innovation are the main topics extracted from the corpus-based on Scopus indexed Indian LIS journals. The fact that there is a lack of publications on cataloging, classification, indexing systems, information retrieval, library management, ontology, and semantic web, emphasises the need to focus more on research in these areas.

Although topic modeling techniques have many implications for evaluating research output, this work exclusively examined the LIS research trends of Indian publications that are indexed in Scopus. Also, new subjects can be found by applying topic modeling techniques such as LDA with Bag of Words. Additionally, LDA techniques help to research automated literature reviews from a large number of papers with the least amount of time and effort. Further, using the train test algorithms, topic modeling offers predictions of research data and identifies the future trends of research.

Analysis of publication patterns using Topic modeling is useful for various purposes, including assisting researchers in identifying popular topics and providing information to individuals who are interested in the current status of LIS research in India.

REFERENCES

1. Lu, Wei.; Liu, Zhifeng.; Huang, Yong.; Bu, Yi.; Li, Xin. & Cheng, Qikai. How do authors select keywords? A preliminary study of author keyword selection behavior, *J. Informetr.*, 2020, **14**(4). doi: 10.1016/j.joi.2020.101066 (Accessed on 30 December 2023).
2. Bengisu, M. & Nekhili, R. Forecasting emerging technologies with the aid of science and technology databases. *Technol. Forecast. Soc. Change,* 2006, **73**(7), 835–44. doi: 10.1016/j.techfore.2005.09.001.
3. Mahesh, B. Machine learning algorithms -A review. *Int. J. Sci. Res.*, 2019, **9**(1). doi: 10.21275/ART20203995.
4. Vayansky, I. & Kumar, S.A.P. A review of topic modeling methods. *Inf. Syst.,* 2020, **94**, 101582. doi: 10.1016/j.is.2020.101582.
5. Blei, D.; Ng, A. & Jordan, M. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 2003, **3**, 993–1022. doi: 10.1162/jmlr.2003.3.4-5.993.
6. Sharma C.; Batra I.; Sharma S.; Malik A.; Hosen ASMS. & Ra IH. Predicting trends and research patterns of smart cities: A semi-automatic review using Latent Dirichlet Allocation (LDA). *IEEE Access*, 2022, **10**, 121080–121095. doi: 10.1109/ACCESS.2022.3214310.
7. Maity, B.K. & Hatua, S.R. Research trends of library management in LIS in India since 1950–2012. *Scientometrics*, 2015, **105**(1), 337–46. doi: 10.1007/s11192-015-1673-8.
8. Sahu, R. & Parabhoi, L. Bibliometric study of library and information science journal articles during 2014-2018: LIS research trends in India. *DESIDOC J. Lib. Inf. Technol.*, 2020, **40**(6), 390–395. doi: 10.14429/djlit.40.6.15631.
9. Jiang, H.; Qiang, M. & Lin, P. A topic modeling based bibliometric exploration of hydropower research. *Renew. Sust. Energ. Rev.*, 2016, **57**, 226–237. doi: 10.1016/j.rser.2015.12.194.
10. Griffiths, T.L. & Steyvers, M. Finding scientific topics. *Proc. Natl. Acad. Sci.,* 2004, **101**(suppl_1), 5228–5235. doi: 10.1073/pnas.0307752101.
11. Steyvers, M. & Griffiths, T. Probabilistic topic models. *In* Handbook of latent semantic analysis. NJ, Laurence Erlbaum, 2007, 427-448.
12. Wang, C.; Blei, D. & Heckerman, D. Continuous time dynamic topic models. *In* Proceedings of the twenty-fourth conference on uncertainty in Artificial Intelligence, 2008, Arlington, Virginia, USA, AUAI Press, 2008. pp. 579–586. (UAI'08).
13. Grimmer, J. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Polit. Anal.,* 2010, **18**(1), 1–35. doi: 10.1093/pan/mpp034.
14. Sun, L. & Yin, Y. Discovering themes and trends in transportation research using topic modeling. *Transp. Res. Part C: Emerging. Technol.*, 2017, **77**, 49–66. doi: 10.1016/j.trc.2017.01.013.
15. Mann, G.S.; Mimno, D. & McCallum, A. bibliometric impact measures leveraging topic analysis. *In* Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries. New York, USA: Association for Computing Machinery, 2006. 65–74. (JCDL '06). doi: 10.1145/1141753.1141765.
16. Li, C.Z.; Zhao, Y.; Xiao, B.; Yu, B.; Tam, V.W.Y., & Chen, Z. *et al.* Research trend of the application of information technologies in construction and demolition waste management. *J. Clean. Prod.,* 2020, **263**, 121458. doi: 10.1016/j.jclepro.2020.121458
17. Sugimoto, C.R. *et.al*. The shifting sands of disciplinary development. *J. Assoc. Inf. Sci. Technol.*, 2011, **62**, 185-204.
18. Yan, E. Research dynamics, impact, and dissemination. *J. Assoc. Inf. Sci. Technol.*, 2015, **66**, 2357-2372.
19. Krenn, M. & Zeilinger, A. Predicting research trends with semantic and neural networks with an application in quantum Physics. *Proc. Natl. Acad. Sci.*, 2020,

117(4), 1910–1916. doi: 10.1073/pnas.1914370116

20. Tiwari, P.; Chaudhary, S.; Majhi, D. & Mukherjee, B. Comparing research trends through author-provided keywords with machine extracted terms: A ML algorithm approach using publications data on neurological disorders. *Iberam. J. Sci. Meas. Com.*, 2023, **3**(1). doi: 10.47909/ijsmc.36.
21. Reisenbichler, M. & Reutterer, T. Topic modeling in marketing: Recent advances and research opportunities. *J. Bus. Econ.,* 2019, **89**(3), 327–56. doi: 10.1007/s11573-018-0915-7.
22. Abuhay, T.M.; Nigatie, Y.G. & Kovalchuk, S.V. Towards predicting trend of scientific research topics using topic modeling. *Procedia Comput. Sci.,* 2018, **136**, 304–310. doi: 10.1016/j.procs.2018.08.284.
23. Mukherjee, B. & Majhi, D. Automatic extraction of significant terms from the title and abstract of scientific papers using the machine learning algorithm: A multiple module approach. *Ann. Libr. Inf. Stud.,* 2023, **70**(1), 33–40. doi: 10.56042/alis.v70i1.71272.
24. Wang, J.; Fan, Y.; Zhang, H. & Feng, L. Technology hotspot tracking: Topic discovery and evolution of China's blockchain patents based on a dynamic LDA model. *Symmetry,* 2021, **13**, 415. doi: 10.3390/sym13030415.
25. Wu, Z.; Xie, P.; Zhang, J.; Zhan, B. & He, Q. Tracing the trends of general construction and demolition waste research using LDA modeling combined with topic intensity. *Front Public Health,* 2022, **10**, 899705. doi: 10.3389/fpubh.2022.899705.
26. Piepenbrink, A. & Nurmammadov, E. Topics in the literature of transition economies and emerging markets. *Scientometrics*, 2015, **102**(3), 2107–2130. doi: 10.1007/s11192-014-1513-2.
27. Wu, Z.; He, Q.; Yang, K.; Zhang, J. & Xu, K. Investigating the dynamics of China's green building policy development from 1986 to 2019. *Int. J. Env. Res. Public Health*, 2020, **18**, 196. doi: 10.3390/ijerph18010196.
28. Wu, Z.; Zhang, Y.; Chen, Q. & Wang, H. Attitude of Chinese public towards municipal solid waste sorting policy: A text mining study. *Sci. Total Environ.,*2021, **756**, 142674. doi: 10.1016/j.scitotenv.2020.142674,
29. Wu, Z.; He, Q.; Chen, Q.; Xue, H. & Li, S. A topical network based analysis and visualisation of global research trends on green building from 1990 to 2020. *J. Clean. Prod.*, 2021, **320**, 28818. doi: 10.1016/j.jclepro.2021.128818.
30. Kobak, D. & Linderman, G.C. Initialisation is critical for preserving global data structure in both t-SNE and UMAP. *Nat. Biotechnol.,* 2021, **39**(2), 156–157. doi: 10.1038/s41587-020-00809-z.
31. Van der Maaten, L. & Hinton, G. Viualising data using t-SNE. *J. Mach. Learn. Res.*, 2008, **9**(86), 2579–2605. https://www.jmlr.org/papers/v9/vandermaaten08a.html (Accessed on 8 July 2023).

CONTRIBUTORS

Mr Debasis Majhi is a Senior Research Scholar in the Department of Library and Information Science, Banaras Hindu University, Varanasi, India. He has published research articles in reputed journals. His areas of research interest include Semantic analysis, Text analysis, ICT applications in libraries, Library automation, Trend analysis, and Digital libraries.
He has collected, organised, processed, and analysed the data for this paper.

Prof Bhaskar Mukherjee is the Professor of the Department of Library and Information Science, at Banaras Hindu University, Varanasi, India. He has published more than 80 research articles in national and international journals. His research interests are in Scientometrics, Open access, Journal evaluation techniques, Information storage and retrieval, Knowledge organisation, etc. He conceptualised the idea and assisting in technical writing and review of the paper.

Appendix-A: Top 10 highly weighted terms of LDA topic modeling from selected three scopus-indexed Indian LIS journals

ALIS	DJLIT	JSCIRES
0.052*”library” + 0.035*”resource” + 0.023*”university” + 0.017*”service” + ‘’0.016*”communication” + 0.014*”user” + 0.012*”science” + 0.010*”identify” + ‘’0.010*”librarian” + 0.009*”institution”	‘0.078*”library” + 0.031*”service” + 0.027*”user” + 0.020* “technology” + ‘’0.018*”librarie” + 0.016* “professional” + 0.014*”development” + ‘’0.013* “management” + 0.013* “university” + 0.012*”tool”	0.023*”publication” + 0.017*”field” + .017*”bibliometric” + 0.014*”trend” ‘’+ 0.014*”journal” + 0.014*”science” + 0.012* “topic” + 0.012*”document” + ‘’0.011*”network” + 0.011* “datum”
0.040*”publication” + 0.040*”science” + 0.039* “citation” + 0.021*”output” + ‘’0.020*”impact” + 0.018*”country” + 0.017*”communication” + 0.015*”datum” + ‘ ‘0.014*”rank” + 0.013*”database”	‘0.040*”publication” + 0.030*”science” + 0.021*”patent” + 0.021*”country” + ‘’0.019*”citation” + 0.019*”institution” + 0.017*”datum” + 0.015*”impact” + ‘’0.013*”growth” + 0.013*”indian”	‘0.020*”country” + 0.017*”citation” + 0.015*”publication” + 0.013*”science” + 0.013*”distribution” + 0.013*”institution” + 0.012*”original” + ‘’0.011*”output” + 0.011*”source” + 0.010*”reproduction”
0.044*”professional” + 0.034*”reference” + 0.030* “error” + 0.023*”attitude” ‘’+ 0.019*”consortia” + 0.019*”key” + 0.016*”preservation” + 0.016*”accuracy” + 0.013*”responsible” + 0.009*”impart”	0.054*”student” + 0.048*”resource” + 0.030*”university” + 0.022*”database” ‘’+ 0.021*”usage” + 0.021*”respondent” + 0.020*”electronic” + ‘’0.019*”questionnaire” + 0.017*”online” + 0.016*”print”	0.028*”citation” + 0.026*”collaboration” + 0.023* “researcher” + ‘’0.022*”scientific” + 0.019*”network” + 0.016*”publication” + 0.015*”impact” ‘’+0.013*”correlation” + 0.012*”field” + 0.012*”university”
0.032*”concept” + 0.019*”trend” + 0.018*”seek” + 0.016*”source” + ‘’0.016*”news” + 0.013*”document” + 0.013*”semantic” + 0.012*”wide” + ‘’0.012* “ontology” + 0.012*”search”	‘0.023*”feature” + 0.020*”college” + 0.014*”datum” + 0.014* “current” + ‘’0.014*”virtual” + 0.014*”plagiarism” + 0.010* “relationship” + ‘’0.010*”programme” + 0.009* “knowledge” + 0.009*”explain”	‘0.034*”citation” + 0.022*”country” + 0.016*”datum” + 0.013*”indian” + ‘’0.013*”policy” + 0.010*”publication” + 0.010*”journal” + 0.009*”indicator” + 0.008*”science” + 0.008*”productive”
0.047*”student” + 0.026*”indicator” + 0.016* “majority” + 0.015*”develop” + ‘’0.013*”school” + 0.012*”software” + 0.012*”undergraduate” + 0.012* “change” + 0.011*”technique” + 0.010* “challenge”’	‘0.029*”repository” + 0.021*”internet” + 0.018*”datum” + 0.016*”digital” + ‘’0.015*”read” + 0.014*”legal” + 0.013* “day” + 0.012*”reading” + ‘’0.012*”mobile_phone” + 0.012*”blog”	‘0.018*”bibliometric” + 0.014*”learn” + 0.012*”innovation” + 0.010*”model” + ‘’0.010*”technology” + 0.010*”water” + 0.009*”country” + ‘ ‘0.009*”scientometric” + 0.007*”field” + 0.007*”publication”
0.087*”indian” + 0.029*”collaboration” + 0.023* “pattern” + 0.016*”network” ‘’+ 0.015*”statistic” + 0.013*”crop” + 0.013*”private” + 0.012*”agricultural” ‘’+ 0.012*”period” + 0.010*”prolific”	‘0.026*”web” + 0.026*”contribution” + 0.019*”cite” + 0.019*”journal” + ‘’0.017*”traditional” + 0.016*”publisher” + 0.015*”language” + ‘’0.015*”scientist” + 0.014*”institutes” + 0.013*”degree”	‘0.026*”citation” + 0.024*”journal” + 0.015*”science” + 0.013*”impact” + ‘’0.012*”readership” + 0.011*”reference” + 0.011*”library” + 0.010*”feature” ‘’+ 0.010*”datum” + 0.009*”coverage”
0.037*”potential” + 0.022*”scheme” + 0.013*”apocynin” + 0.010*”full” + ‘’0.010*”raise” + 0.010*”globe” + 0.009* “cannabinum” + 0.009*”apocynum” + ‘’0.009*”cgp” + 0.007*”cell”	‘0.046*”preservation” + 0.045*”digital” + 0.026*”medical” + ‘’0.022*”catalogue” + 0.021*”mobile” + 0.018*”archive” + 0.014*”health” + ‘’0.014*”record” + 0.013*”account” + 0.011*”practice”	‘0.021*”development” + 0.018*”health” + 0.010*”indicator” + ‘’0.009*”evolution” + 0.009*”pollution” + 0.009*”bric” + 0.009*”word” + ‘’0.008*”knowledge” + 0.007*”community” + 0.007*”university”
0.030*”classification” + 0.022*”distribution” + 0.016* “url” + ‘’0.016*”medical” + 0.015*”catalogue” + 0.013*”difference” + 0.013*”mining” + ‘’0.013* “principle” + 0.011*”description” + 0.011*”web”	0.153*”journal” + 0.047*”citation” + 0.017*”database” + ‘ ‘0.015*”classification” + 0.013*”bibliographic” + 0.011* “multi” + ‘’0.010*”science” + 0.009*”volume” + 0.009*”rate” + 0.009*”document”	0.059*”innovation” + 0.031*”language” + 0.025*”country” + ‘’0.011*”publication” + 0.011*”platform” + 0.010*”woman” + 0.010*”government” + 0.009*”mobile” + 0.009*”bilateral” + 0.009*”issue”
0.064*”lis” + 0.024*”newspaper” + 0.024*”skill” + 0.020*”level” + ‘’0.016*”site” + 0.015*”ugc” + 0.014* “development” + 0.012*”college” + ‘’0.011*”job” + 0.011* “attribute”	‘0.046*”open_access” + 0.029*”lis” + 0.016*”consortium” + 0.015*”public” + ‘’0.014*”demand” + 0.013*”teacher” + 0.013*”firm” + 0.011*”scientific” + ‘’0.011*”policy” + 0.011* “national”	0.038*”patent” + 0.018*”cancer” + 0.018*”country” + 0.018* “disease” + ‘’0.012*”burden” + 0.011*”public” + 0.010* “biosimilar” + 0.010*”trend” + ‘’0.009*”relative” + 0.008* “prostate_cancer”
0.049*”scientist” + 0.016*”facility” + 0.016*”doctoral” + 0.013*”age” + ‘’0.013*”laboratory” + 0.012*”work” + 0.012*”career” + 0.012*”position” + ‘’0.011*”wise” + 0.011*”cybercafe”	0.033*”digital” + 0.022*”document” + 0.013*”art” + 0.012* “search” + ‘’0.011*”museum” + 0.011*”image” + 0.010* “device” + 0.010*”generally” + ‘’0.010*”tool” + 0.009* “design”	0.034*”university” + 0.020*”country” + 0.019*”rank” + 0.018* “ranking” + ‘’0.014*”indicator” + 0.014*”regional” + 0.013* “software” + ‘’0.009*”optogenetic” + 0.009*”tool” + 0.008* “power”