Annals of Library and Information Studies
Vol. 69, September 2022, pp. 191-202
DOI: 10.56042/alis.v69i3.58719

Using text analysis to study doctoral-level library and information science research trends in India

Vinit Kumar[a] and Khusbu Thakur[b]

Department of Library and Information Science, Babasaheb Bhimrao Ambedkar University, Lucknow, U.P, India,
[a]E-mail: mailvinitkumar@gmail.com, [b]E-mail: thakurkhusbu02@gmail.com

Received: 30 December 2021; revised: 25 July 2022; accepted: 26 July 2022

This study attempts to identify research trends in doctoral-level library and information science research in India. The study uses topic modelling with Latent Dirichlet Allocation to identify research trends by collecting relevant sections from the full-text of theses awarded in the last ten years by Indian universities in the discipline of library and information science. The topic modelling results show that the entire corpus can be classified into ten topics, including information communication technology and its application in libraries. The topic-wise trends indicate that 'ICT and its application in libraries' are still the prime themes of choice among doctoral-level research in Indian LIS schools, followed by studies on 'Information seeking behaviour'. The growth-wise trend analysis suggests the decline in the interest in 'Bibliometrics/Scientometrics' and 'Webometrics/Website evaluation studies'. The findings of the study will be of help to academics studying the development of LIS research, as well as aspiring doctoral-level scholars who are to identify a topic for their research work.

Keywords: Trends analysis, Text mining, Topic modelling, LIS Research, Doctoral Research

Introduction

In the last ten years, we have seen a paradigm shift in the variety of topics chosen by Library and Information Science (LIS) researchers to answer problems related to information behaviour, library management, measuring research productivity, and several other areas. LIS research is primarily advanced through three channels: the publication of research papers in established academic journals, presentations at conferences and seminars, and the submission of thesis and dissertations. Finding a suitable research topic or understanding the research trend can be difficult for a new researcher. It takes a lot of effort on the part of the entrant researchers to identify such trends. They have to go through a lot of published literature and read some of it in depth. It will save them a lot of time and effort if they have access to trends about research topics that are gaining popularity and those that are losing popularity.

Recent advances in text mining tools and techniques can potentially process large amounts of textual data to identify trends, valuable insights, and other information. Because of the rapid growth of electronic text data, there are significant challenges in organising, managing, and disseminating information. There is a massive amount of electronic textual data in the academic world, and this textual data is available in a variety of formats. It could be structured, unstructured, or semi-structured.

Text mining is a method for analysing unstructured text to extract meaningful patterns from unstructured text data using various text mining tools and techniques. Text mining is a technique for identifying text patterns and analysing trends in unstructured textual data. Text mining is becoming increasingly popular in academia, and there is a constant increase in textual academic content such as electronic scholarly journal articles, e-newspapers, electronic theses and dissertations (ETDs).

The evolution of research trends is an important area of study in every discipline as it helps understand the growth and future trajectory of a field. To understand the research trends in the LIS domain in India, we used text mining on a corpus containing textual data from 1271 LIS theses submitted in the Shodhganga repository (an Indian ETDs digital repository) between 2010 and 2019.

Using topic modelling technique, we analyse the research trends based on LIS theses. The study explores the topics studied in the library and information science theses from 2010 to 2019, using the text mining technique.

Review of literature

The current study aims to identify trends in doctoral-level research conducted in India in library and information science (LIS) using topic modelling. Most researchers use techniques like content analysis, scientometric analysis, co-word occurrence, keyword frequency, and text analysis to determine the direction of current research. We examined research trends in India and abroad.

Dutta and Mondal[1] applied content analysis to investigate the progression of the number of LIS theses conferred in India from 1950 to 2017. The findings of this research indicated that the areas of LIS that have received the most attention from academic researchers are bibliometrics, information systems, information sources and services, and community information services. Dora and Kumar[2] analysed the research trends in the LIS between 2004 and 2005 using co-occurrence analysis, keyword frequency, and burst detection methods. The authors revealed that bibliometrics and scientometrics were the most popular topics in the field of LIS.

Using qualitative analysis, Chandrashekara & Ramasesh[3] examined the research trends of LIS doctoral dissertations awarded between 1985 and 2005. They discovered that the most popular topics in LIS were bibliometrics/scientometrics/informatics, library management, university libraries, indexing, information-seeking behaviour, and library and information services. Chakrabarti, Mondal, and Maity[4] conducted a trend analysis of LIS doctoral dissertations awarded between 1979 and 2018 and found that the majority of researchers focused on topic related to user study, information-seeking behaviour, information sources and services.

Pandita & Singh[5] analysed doctoral degrees awarded in LIS in India during 2010–2014, and they reported that most doctoral-level studies were undertaken in the area of application of information technology in libraries. Some authors have investigated the research trends of scholarly research papers such as Mittal[6] examined the research trends of 1,408 scholarly research papers written by Indian authors and included in the Library and Information Science Abstracts (LISA) database from January 1990 to June 2010. Using co-word occurrences, the author identified library practice, user services, cataloguing, user studies, university libraries, and public libraries as core research areas. While topics such as the World Wide Web, open access, and Web 2.0 were found to be the trending topics in the field of LIS.

Abdoulaye[7] looked at the research trends of master's theses awarded by the International Islamic University Malaysia from 1994 to 2004 and found that information technology, information needs, and library management were the most popular research areas among master's students in library and information science. Samdan and Bhatti[8] looked at research trends in 18 LIS doctoral theses awarded in Pakistan between 1947 and 2010. The study found that academic libraries, school libraries, library education, information technology, information seeking behaviour, library management, user education, publishing, history of Islamic libraries, classification, and bibliometrics studies were the most popular research topics among Pakistani doctoral researchers.

Mahapatra & Sahoo[9] examined the research topics of thirty-one doctoral theses in LIS awarded by Indian universities from 1997–2003. Information needs, information seeking behaviour, user satisfaction, and evaluation of information resources and services are among the most recently focused areas among doctoral-level researchers. Mundhiala, Sahoo, Dash & Mohanty[10] worked on trend analysis of the library and information science doctoral research from 2004-2018. The study revealed that information needs and seeking behaviour, information sources, services, and systems are the broad areas of this discipline.

Several researchers used various techniques such as bibliometric analysis[11-17], citation analysis [18,19] and content analysis on a small sample of articles, theses, and dissertations in the domain of LIS to measure the research trend analysis, publication pattern, chronological distribution, topic trends analysis[20,21] and identifying the relationship between journal articles and author interests. Text mining techniques have also been used to pinpoint research trends across a range of academic fields, including computer science[22], biology[23], medicine[24-26], education[27], and social science disciplines[28-30].

Some studies in the literature apply topic modelling to LIS literature, especially doctoral theses and dissertations. Sugimoto *et al.*[31] analysed 3121 LIS doctoral dissertations from 1930–2009. Applying the LDA technique, they highlighted the core topics in LIS, such as library history, citation analysis, information-seeking behaviour, information retrieval, and information use.

Lamba & Madhusudhan[32] analysed topic research trends by applying LDA topic modelling on awarded doctoral theses in library and information science (2013–2017) and reported five core topics:

information literacy, user studies, scientometrics, library resources, and library services. Mazumder & Barui[33] aimed to discover the research topic from the title of 2132 doctoral theses in LIS and applied topic modelling through the LDA technique. The study found that library use, open-source, management, university library, information seeking behaviour, collection development, ICT, higher education and librarianship are the most popular topics in LIS.

The literature review revealed that most research trend studies used the content analysis or bibliometric/ scientometric methods on small samples of LIS doctoral theses. There are several limitations in evaluating research trends that rely on bibliometrics or manual content analysis approaches such as there is a chance of ambiguous topic identification in bibliometric studies because they primarily rely on author-supplied keywords and only the text of the title and abstract. Similar to quantitative analysis using bibliometric methods, manual content analysis requires significant intellectual effort to identify topics; as a result, it is helpful for studies with smaller sample sizes. The topic modelling technique is an unsupervised text analysis method that can be used for studies with large samples because it can quickly analyse the entire text, ultimately saving much human effort.

Objectives of the study

- To identify the dominant themes on which the Indian doctoral-level LIS researchers have conducted studies during 2010 to 2019; and
- To explore the emerging trends in the dominant themes on which research topics were chosen by the doctoral-level LIS research.

Scope

The study's scope is limited to theses awarded in the subject of library and information science in India between 2010 and 2019, and available in the Shodhganga repository, an Indian open access repository of ETDs of theses and dissertations awarded by Indian universities (https://shodhganga.inflibnet.ac.in/) as of April 15th, 2021. Only theses written in English were included in the study.

Methods

Due to the unavailability of electronic copies of all the theses awarded to date by all LIS schools on a single platform, we collected electronic versions of the awarded thesis from the Shodhganga repository. For the selected period of ten years (2010-2019), a total of 1425 theses were available in the repository, out of which 78 theses were excluded from the study because two of them had no pdf files attached, 53 theses were also excluded as they were written in a language different from the English language and 23 theses were duplicate entries and hence excluded. Moreover, 23 theses were found without the year of submission, although they were considered in this study. After applying inclusion and exclusion criteria, 1271 theses were selected for further analysis.

Table 1 shows the year-wise distribution of the total 1271 theses that have been retrieved from the Shodhganga repository for ten years (2010-2019).

Table 1 indicates that the maximum of 14.56% of LIS theses belong to the year 2018, followed by 13.14% in 2016, 14.48% in 2017, and 10.86% in 2019, respectively. Around 50% of the dataset belongs to the years 2015-2018.

The PDF files of the selected theses were converted to text files to extract the relevant contents from the sections "Title", "Aim of the study," "Objectives of the study," "Significance of the study," and "Hypotheses". Additionally, the other bibliographic details, such as the year of submission and the author names were collected for each thesis. We chose to include the relevant sections rather than just the title or abstract for two reasons: first, the abstract was not available in all of the theses available at the Shodhganga repository, and second, as Syed & Spruit[34] reported limitations in the appropriateness of the topic modelling results for small text content, we wanted to include more representative content in the dataset. Another reason for selecting the contents of the sections mentioned above was to have enough text to determine the theme or research area, as involving the entire full text, which includes lengthy tables with numerical data and interpretations with repetitive

Table 1 — Number of theses included in this study

Year	Number of theses	Percentage (%)
2010	51	4.01%
2011	63	4.96%
2012	104	8.18%
2013	118	9.28%
2014	119	9.36%
2015	139	10.94%
2016	167	13.14%
2017	184	14.48%
2018	185	14.56%
2019	138	10.86%
NA	3	0.24%
Total	1271	

sentences, could have resulted in noise in the data, resulting in ambiguous theme determination.

Some of the theses, for example, had textbook-like content in the research methodology chapter discussing the various types of data collection methods in detail, which has little relevance to the study at hand; if such data were included in the corpus, the topic modelling algorithm would have placed the thesis in themes such as 'interview' or 'questionnaire' which is not the subject area, the thesis has actually studied. Hence, we selected only those sections of the thesis which could provide enough information to determine the dominant themes. We chose a ten-year period, 2010-2019, to investigate current research trends in LIS schools. The PDF files of the theses awarded in the selected period were downloaded from 3 March 2021 to 15 April 2021.

The extracted content was saved as a dataset in a CSV file to develop a text corpus further. Python libraries such as Pandas, Gensim, NLTK, and pyLDAvis were used to text analyse the extracted contents. The text was pre-processed in the following order: changing the text to lowercase, removing stop words using the NLTK stop word list (for the English language) along with a customised stop word list consisting of frequently used words in the section descriptions such as hypothesis, aims, and objectives in most of the theses. The text corpus was also lemmatised using the WordNet lemmatiser, which is included in Python's NLTK package. Similarly, terms that appeared in more than 80% of the theses and rare terms that appeared in less than ten theses were removed from the corpus. The block diagram depicts the steps taken in the study, such as metadata collection, selection, pre-processing, transformation, topic modelling, evaluation, and analysis (Fig 1).

Topic modelling is one of the methodologies used in natural language processing and text analysis for text summarisation. It entails using statistical and mathematical modelling tools to extract important themes or concepts from a collection of text documents. These statistical models are probabilistic in nature as the extracted topics are based on the probability of terms present in a document and in a

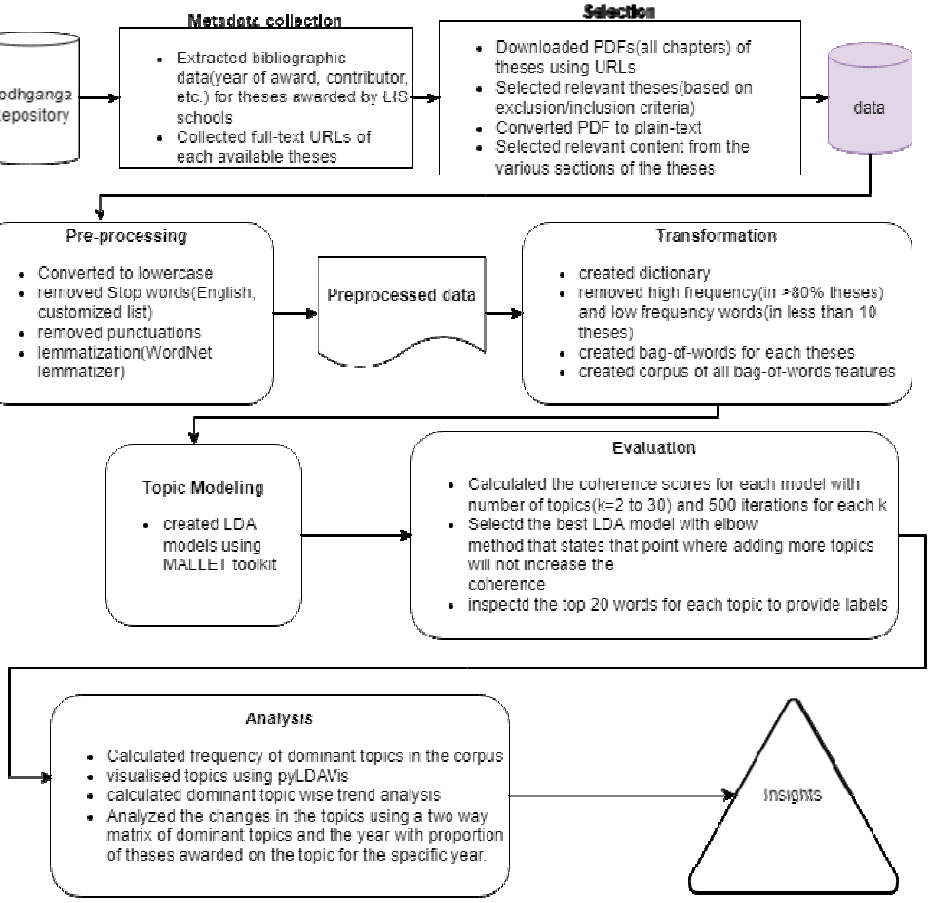

Fig. 1 — Block diagram showing the steps undertaken in the study

collection of documents. The extracted themes are referred to as topics, and each can be thought of as a bag or collection of words from the documents that collectively constitute the theme of the document. In other words, themes are hidden in the text, and topic modelling aids in the discovery of these associated latent semantic structures that aid in the summarisation of a vast collection of documents[35].

The topic models employ statistical techniques such as Singular Value Decomposition (SVD) and Latent Dirichlet Allocation (LDA) to uncover the latent themes. The LDA topic model used in this study is a Bayesian probabilistic topic model based on the idea that documents contain various themes in different proportions, and thus it captures the heterogeneity of the concepts in a text. Given a set of documents, the LDA topic model generates a document-topic matrix with topic probability for each document and a topic-term matrix with term probability for each topic[36]. For a concise introduction to LDA, interested readers may refer to Blei et al.'s seminal paper[37].

For the entire dataset, we performed topic modelling using Latent Dirichlet Allocation (LDA) deploying Python's Gensim library, and the MALLET toolkit implementation, considering the popularity of LDA[38]. We calculated the coherence scores for the number of topics ranging from two to thirty topics with five hundred iterations for each number of topics in order to decide on the number of topics (Fig. 2).

We discovered that topics 5, 8 and 10 had the highest coherence scores; considering the elbow method that states that the point where adding more topics will not increase the coherence, we selected the number of optimal topics as 10. Other parameters were left as they were in the Gensim library's documentation as a standard setting. The authors chose the labels for the identified topics corroborating with experience in the LIS domain and considering the top 20 terms in each topic until the authors were in full agreement. The changes in the topics were analysed using a two way matrix of dominant topics and the year with the proportion of theses awarded on the topic for the specific year.

Results

Topic modelling - Analysis of LIS theses during 2010-2019

We used the Latent Dirichlet Allocation (LDA) algorithm to identify the dominant topics chosen by LIS doctoral researchers over the last ten years. LDA is an unsupervised learning algorithm that classifies documents in a set based on the occurrence of terms in a bag of words for each item and generates a classification model based on term probability. The topic modelling technique in text mining is one of the most powerful techniques for discovering hidden topics in textual data. Table 2 summarises the ten dominant topics generated using Python's Gensim library with the MALLET toolkit. Each topic was then labelled as an appropriate theme by the authors in agreement. For each topic, the table shows the number and proportion of documents in the dataset. In addition, the top ten terms for each topic are listed based on their probability.

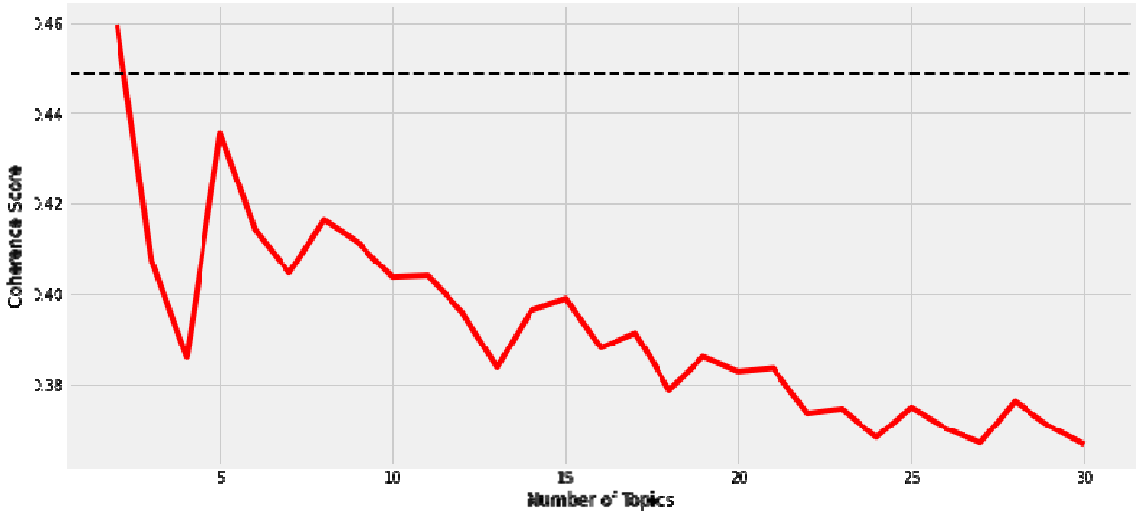

Fig 2 — Coherence analysis showing scores and number of topics

Table 2 — Distribution of theses based on dominant topics

Dominant Topic	Theme	Number of theses	% Total theses	Topic terms (top 20 terms based on the probability)
1.	Awareness of library resources	182	14.1	resource, electronic, faculty_member, scholar, internet, access, faculty, academic, usage, user, awareness, print, institution, accessing, consortium, online, technology, database, satisfaction, search
2.	Bibliometric/Scientometric Study	174	13.48	publication, literature, output, growth, scientific, author, journal, pattern, analysis, country, trend, citation, productivity, institution, India, published, subject, contribution, article, distribution
3.	Information seeking behaviour	164	12.7	college, user, service, engineering, medical, satisfaction, source, facility, seeking_behaviour, quality, seeking, staff, art_science, affiliated, academic, education, tamil_nadu, student, perception, reference
4.	ICT and its application in libraries	124	9.6	ICT, service, public, digital, development, community, knowledge, government, system, infrastructure, communication_technology, centre, network, region, district, sharing, role, problem, traditional, rural
5.	Case study on different types of libraries	119	9.22	university, state, central, agricultural, Kerala, India, user, agriculture, tamil_nadu, Delhi, problem, ass, Gujarat, social_networking, department, Maharashtra, work, uttar_pradesh, limitation, surveyed
6.	Open access and scholarly communication	99	7.67	science, journal, researcher, social, open_access, source, subject, book, work, lis, citation, publishing, reference, scientist, discipline, problem, form, list, analysis, thesis
7.	Webometric Study/Evaluation of Websites	98	7.59	web, website, content, system, law, tool, data, standard, model, design, quality, analysis, question, specific, problem, evaluation, Indian, feature, open, institution
8	Library management	119	9.22	professional, management, librarian, technology, software, application, skill, institution, change, working, work, lis_professional, knowledge, environment, staff, academic, challenge, job_satisfaction, competency, related
9.	Reading habits of different groups in society	112	8.68	student, education, literacy, school, learning, woman, health, medium, knowledge, teacher, reading_habit, reading, source, relationship, child, gender, people, understand, society, social
10.	Collection development	100	7.75	collection, development, institute, India, technology, national, automation, material, Assam, policy, problem, organisation, survey, method, document, system, product, select.

Table 2 indicates the ten dominant topics on which the LIS schools in India awarded PhD degrees in the last ten years (2010-2019). The dominant topics are derived from the output of the LDA algorithm. It indicates that during the last ten years, the majority of theses(14.1%) in LIS were awarded on topics revolving around the theme of "Awareness of library resources" involving topics measuring the awareness, satisfaction, and usage of print and electronic resources available in libraries. Followed by bibliometric/scientometric studies (13.43%), information seeking behaviour (12.7%), ICT and its application in libraries (9.6%), case study on different types of libraries (9.22%), open access and scholarly communication (7.67%), webometric study/evaluation of websites (7.59%), library management (9.22%), reading habits of different groups in society (8.68%), and collection development (7.75%).

For further visualisation of topics, we used pyLDAVis, to create an inter tropical distance map using multidimensional scaling, which helps in visualising the LDA topic model in a 2D map where the topics are displayed as circles and distances between the circles show the relatedness and the number of terms being shared by each topic. An overlap of circles shows that the topics are highly related and share common terms. Figure 3 shows that topics 4 and 7 are related, while topics 3 and 5 are associated with each other.

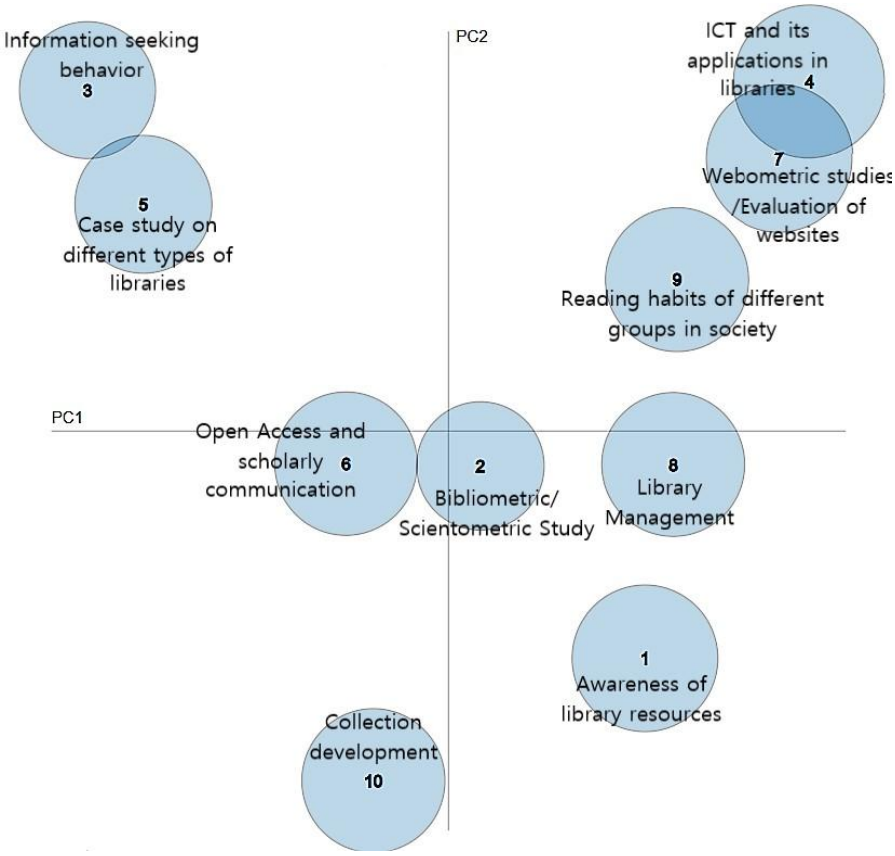

Fig. 3 — Visualisation of the dominant topics in LIS doctoral theses

A significant proportion (13.48%) of theses were found to be in the area of bibliometricss/ scientometrics. The third dominant topic emerged out to be studies on "Information Seeking Behaviour" with 12.7% of awarded theses in the dataset belonging to this category. It suggests that the information-seeking behaviour of library users and their perception towards library services are important concerns for LIS doctoral-level researchers.

In the last decade, the LIS schools awarded 9.6% of theses in the area of "Information Technology and its application in libraries" which includes studies measuring the aspects such as library automation and use of ICT in the delivery of library services.

The other themes, "Case studies on different types of libraries" constituted 9.22% of theses, followed by topics related to "Library Management" with 9.22% of theses, "Reading habits of different groups in society" with 8.68%, "Collection development" with 7.75%, "Open access and scholarly communication" with 7.67% of theses. The studies related to the theme of "Webometric Study/Evaluation of Website" accounted for 7.59% of theses in the dataset (Fig. 3).

Trend analysis

We calculated the number of theses belonging to each dominant topic for each year to understand which of the dominant topics identified in the topic modelling were dominant during the ten-year period. In addition, for comparison, we calculated the proportion of theses awarded on a specific dominant topic to the total number of theses awarded that year.

Table 3 and Fig. 4 indicate that topic number four with the theme "ICT and its application in libraries" had the highest proportion of awarded theses in 2013 and consecutively for the last four years (2016-2019), totalling to 5 (50%) times in the selected time period, which indicates that research topics revolving around this theme are popular among the doctoral LIS researchers. Similarly, theme three, "Information seeking behaviour" had the highest share of awarded thesis in 2011, 2014, 2015, totalling 3 (30%) times in the selected time period, indicating that the second choice of researchers is research topics in this theme. However, themes two and seven, "Bibliometric/ scientometric" and "Webometric and website evaluation" had maximum proportions in the year 2012 and 2010, respectively.

Table 3 — Dominant topic-wise trends

Dominant Theme	Topic No.	Year (% of theses awarded) 2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Awareness of library resources	1	3 (5.88%)	8 (12.69%)	16[a] (15.38%)	11 (9.32%)	8 (6.72%)	15 (10.79%)	11 (6.58%)	11 (5.97%)	19 (10.27%)	18 (13.04%)
Bibliometric/Scientometric Study	2	7 (13.72%)	0 (0%)	20[b][a] (19.23%)	9 (7.62%)	10 (8.40%)	7 (5.03%)	20 (11.97%)	17 (9.23%)	12 (6.48%)	14 (10.14%)
Information seeking behavior	3	2 (3.92%)	8[b] (12.69%)	12 (11.53%)	12 (10.16%)	23[b] (19.32%)	29[b][a] (20.86%)	24 (14.37%)	25 (13.58%)	22 (11.89%)	14 (10.14%)
ICT & Its application in libraries	4	6 (11.76%)	5 (7.93%)	5 (4.80%)	20[b] (16.94%)	16 (13.44%)	21 (15.10%)	29[b] (17.36%)	32[b][a] (17.39%)	29[b] (15.67%)	19[b] (13.76%)
Case study on different types of libraries	5	7[a] (13.72%)	8 (12.69%)	5 (4.80%)	9 (7.62%)	5 (4.20%)	16 (11.51%)	11 (6.5%)	11 (5.97%)	14 (7.56%)	13 (9.42%)
Open Access and scholarly communication	6	5 (9.80%)	3 (4.76%)	4 (3.84%)	14[a] (11.86%)	5 (4.20%)	9 (6.47%)	17 (10.17%)	18 (9.78%)	10 (5.40%)	10 (7.24%)
Webometric Study/Evaluation of Websites	7	8[b] (15.68%)	6 (9.52%)	17[a] (16.34%)	17 (14.40%)	17 (14.28%)	18 (12.94%)	18 (10.77%)	22 (11.95%)	28 (15.13%)	12 (8.69%)
Library Management	8	3 (5.88%)	6 (9.52%)	7 (6.73%)	13 (11.01%)	19[a] (15.96%)	9 (6.47%)	14 (8.38%)	16 (8.69%)	14 (7.56%)	13 (9.42%)
Reading habits of different groups in society	9	3 (5.88%)	6 (9.52%)	7 (6.73%)	3 (2.54%)	8 (6.72%)	7 (5.03%)	11 (6.58%)	25 (13.58%)	26[a] (14.05%)	16 (11.59%)
Collection development	10	7 (13.72%)	13[b][a] (20.63%)	11 (10.57%)	10 (8.47%)	8 (6.72%)	8 (5.75%)	12 (7.18%)	7 (3.80%)	11 (5.94%)	9 (6.52%)

[a]*highest proportion in the concerned theme over the study period.*
[b]*highest proportion in the concerned year among all the themes.*

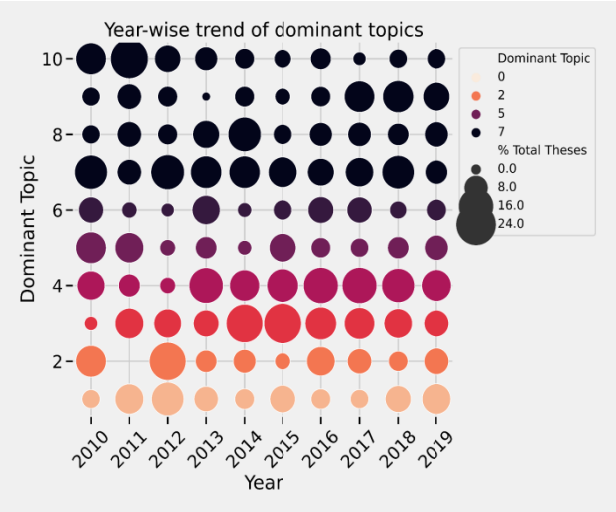

Fig. 4 — Year-wise trend of dominant topics of LIS theses

When we compare the themes over the study period based on the proportion of theses awarded in the corresponding theme, theme one 'Awareness of library resources', theme two 'Bibliometric/ Scientometric study' and theme seven 'Webometric Study/Evaluation of Websites' had the highest proportion of theses in 2012. Similarly, themes three 'Information seeking behaviour' had the highest proportion in 2015 and theme four 'ICT & its application in libraries' in 2017, while theme six 'Open Access and scholarly communication had the highest proportion (peak) in 2013, theme eight 'Library management' in 2014, theme nine 'Reading habits of different groups in society' in 2018, and theme ten 'Collection development' in 2011(Fig. 4). This year-wise comparison shows the year when a particular theme was at its peak level of popularity.

The year-wise growth of dominant topics

We calculated the percentage growth of theses proportions during the study period to investigate the growth rate of theses awarded for each theme. The percentage growth is calculated by dividing the increase from the previous year's value by the last year's value multiplied by hundred. It indicates the percentage increase in the values from the previous year to the present year and helps compare the year-on-year growth.

Table 4 — Year-wise growth of dominant topics

Dominant Theme	Topic No.	% Growth									AAGR
		2010-2011	2011-2012	2012-2013	2013-2014	2014-2015	2015-2016	2016-2017	2017-2018	2018-2019	
Awareness of library resources	1	116%	21%	-39%	-28%	61%	-39%	-9%	72%	27%	20%
Bibliometric/Scientometric Study	2	-100%	0%	-60%	10%	-40%	138%	-23%	-30%	56%	-5%
Information seeking behavior	3	224%	-9%	-12%	90%	8%	-31%	-5%	-12%	-15%	26%
ICT & Its application in libraries	4	-33%	-39%	253%	-21%	12%	15%	0%	-10%	-12%	18%
Case study on different types of libraries	5	-8%	-62%	59%	-45%	174%	-44%	-8%	27%	25%	13%
Open Access and scholarly communication	6	-51%	-19%	209%	-65%	54%	57%	-4%	-45%	34%	19%
Webometric Study/Evaluation of Websites	7	-39%	72%	-12%	-1%	-9%	-17%	11%	27%	-43%	-1%
Library Management	8	62%	-29%	64%	45%	-59%	30%	4%	-13%	25%	14%
Reading habits of different groups in society	9	62%	-29%	-62%	165%	-25%	31%	106%	3%	-18%	26%
Collection development	10	50%	-49%	-20%	-21%	-14%	25%	-47%	56%	10%	-1%

According to Table 4, theses on the themes, ‘Awareness of library resources’ and ‘Information seeking behaviour’, experienced the most significant growth in 2010-2011, while theme seven, ‘Webometric studies and Evaluation of websites’ underwent the greatest growth rate in 2011-12. Similarly, three of the ten selected themes, ‘ICT and its application in libraries’, ‘Open access and scholarly communication’, and ‘Library management’ experienced the largest increase in 2012-13, while theme nine ‘Reading habits of different groups in society’ experienced the strongest growth in 2013-14. Furthermore, theme five, ‘Case study of different types of libraries,’ experienced the largest growth in 2014-15, while theme two, ‘Bibliometrics/ Scientometeric study,’ experienced the strongest increase in 2015-16. Similarly, theme ten, ‘Collection development’ experienced the greatest growth in 2017-18.

To compare the growth rate for each theme over the study period, we calculated the Average Annual Growth Rate (AAGR). AAGR is calculated by taking the average of percentage growth rates over a time period. From Table 6, it can be seen from the data that the maximum growth is observed in theses awarded in themes three and nine (both 26%), followed by theme one (20%), theme six (19%), theme four (18%), theme eight (14%), and theme five (13%). In contrast, an annual average decline is observed in theme 7 (-1%), theme 10(-1%), and theme two (-5%).

Based on this analysis and Fig. 5, we can conclude that, in terms of percentage growth, all themes except “Bibliometrics/Scientometrics” and “Webometrics and Website evaluation” are on the rise. It can be deduced that these topics, while popular from 2012 to 2016, have matured, and research scholars are now venturing more into traditional areas of librarianship such as “Information seeking behaviour” and “Reading habits of different groups in society.” Similarly, theses on topics such as “ICT Applications in Libraries,” "Awareness of Library Resources," and "Open access and scholarly communications" have grown steadily

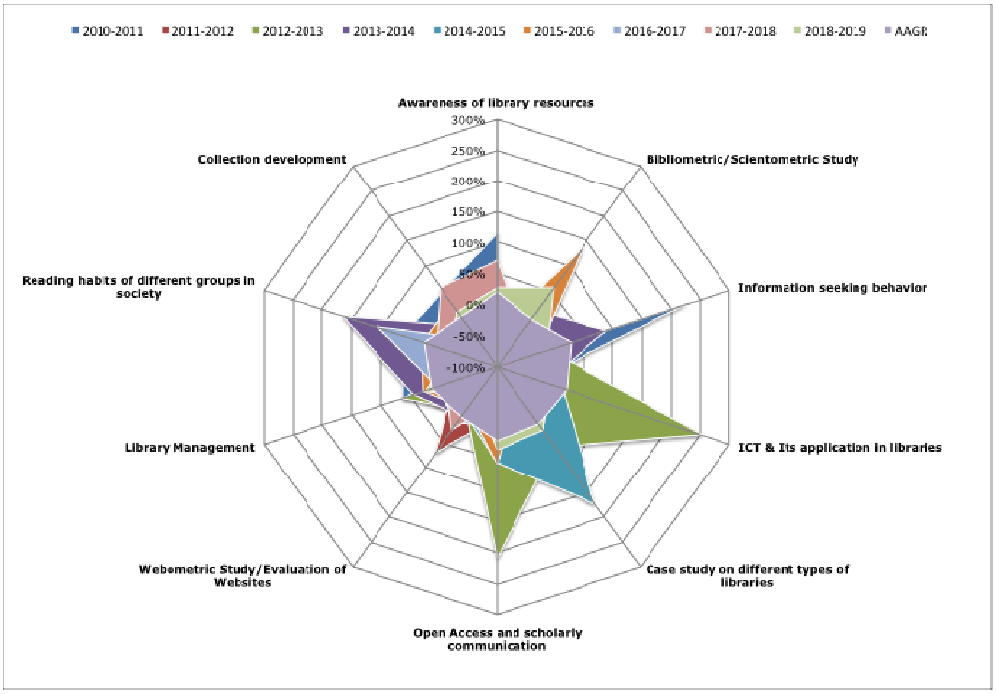

Fig. 5 — Growth rate of theses for each theme during (2010-20)

over the last ten years. Another significant finding is that, even though "ICT applications in libraries" was a popular topic over the last two decades, LIS researchers still prefer to study the impact of ICT in libraries, which may indicate that several repetitive studies have been conducted in this area.

Discussion

The topic modelling results indicate that the doctoral level; researchers preferred studies in areas such as "Awareness of library resources", "Bibliometric/Scientometric Study", "Information Seeking Behaviour", "ICT & Its application in libraries", "Case study on different types of libraries", "Open Access and scholarly communication", "Webometric Study/Evaluation of Websites", "Library Management", "Reading habits of different groups in society", and "Collection development". These findings of the study further indicate that the studies on the awareness of library resources is still the prime topic of concern for doctoral-level research in Indian LIS schools. Further, more than half of the theses awarded by Indian LIS schools in the last ten years were in the fields of "Awareness of information resources," "Bibliometrics/Scientometrics," or "ICT application in libraries." These findings are in line with the results of most of the earlier studies, with the exception that the present study identified some new themes such as awareness of library resources, open access and scholarly communication, and reading habits of different groups in society, whereas some of the important themes such as information retrieval[31], cataloguing and classification[6,7,9], history of libraries[7-8], knowledge management[2], information literacy[32], LIS education[8,21], community information services[2] were not so dominant to be included in top ten themes. This suggests that doctoral level researchers are not interested in typical librarianship topics like cataloguing, classification, and information retrieval. The emergence of electronic resources, the open access movement, and changes in scholarly practices all provided new problems that scholars desired to examine, explaining the researchers' shift in focus to new topics.

Looking at the year-by-year trends in the ten dominant topics of LIS doctoral theses, it can be concluded that PhD-level studies continue to research ICT and its application in libraries, which can be explained by the tremendous impact on librarianship as a result of ICT implementation. Similarly, the findings of growth trends indicate that there is a shift in researchers from areas such as bibliometrics and webometrics, which gained popularity due to the easy availability of data from indexing databases, to areas such as information seeking behaviour and studying the reading habits of different groups in society.

Conclusion

The study concludes that there has been a shift of researchers to areas that have direct involvement in designing library policies, such as reading habits and information seeking behaviour. Doctoral-level researchers must abandon repetitive topics that ultimately degrade the quality of research and impede the growth of the field. The decline in some themes in recent years will have impact on the LIS research in the coming years and probably will lead to the emergence of new areas of focus. However, the growth of doctoral-level research also depends on a number of other variables, including the availability of research infrastructure, competent research supervisors, and the calibre of admitted doctoral students.

Topic modelling provides a faster and efficient approach to classify the collection of documents in topics. Since topic modelling using LDA is based on the likelihood of terms in a document, there is a risk of losing the themes due to the lack of representative documents in the dataset. There is a possibility that the topics with the fewest dissertations will be left unnamed or amalgamated with comparable topics or themes. Similar to manual content analysis, our method of classifying theses using automated unsupervised learning algorithms will almost surely have quality flaws. As future work, we intend to assess the effectiveness of automated versus manual content analysis on a more representative sample of LIS theses. Additionally, analogous studies comparing the research trends of two or more nations might be taken up.

The findings of the study will have broader implications for all the stakeholders, such as research supervisors, departmental research committees, research scholars and researchers interested in analysing the evolution of LIS research. It will also assist prospective doctoral-level researchers and their research supervisors in identifying a topic for their thesis.

References

1 Dutta S and Mondal D, Library & Information Science Education in the Universities of India: growth and development of research, *Library Philosophy and Practice (e-journal)*, 4676 (2020) 24-32. DOI: https://digitalcommons.unl.edu/libphilprac/4676

2 Dora M and Anil K H, An empirical analysis of the research trends in the field of library and information science in India–2004-2015, *COLLNET Journal of Scientometrics and Information Management*, 11 (2) (2017) 361-378.

3 Chandrashekara M and Ramasesh C P, Library and information science research in India, In *Asia–Pacific Conference on Library & Information Education & Practice*, 2009 p. 530-537.

4 Chakrabarti K, Mondal D and Maity A, A Trend Analysis of the Doctoral Dissertations in LIS Research in West Bengal, India during 1979-2018, *Library Philosophy and Practice (e-journal)*, 4149 (2020) 1-30. DOI: https://digitalcommons.unl.edu/ libphilprac/4149

5 Pandita R and Singh S, Doctoral theses awarded in library and information science in India during 2010-2014: A Study, *DESIDOC Journal of Library & Information Technology*, 37 (6) (2017) 379-386.

6 Mittal R, Library and information science research trends in India, *Annnals of Library and Information Studies*, 58 (4) (2011) 319-325.

7 Abdoulaye K, Research trends in library and information science at the International Islamic University Malaysia, *Library Review*, 51 (1) (2002) 32 - 37.

8 Samdan R A and Bhatti R, Doctoral research in library and information science by Pakistani professionals: An analysis, *Library Philosophy and Practice (e-journal)*, 649 (2011). DOI: https://digitalcommons.unl.edu/libphilprac/649

9 Mahapatra R K and Sahoo J, Doctoral dissertations in library and information science in India. *Annnals of Library and Information Studies,* 51 (1) (2004) 58-63.

10 Mundhial S, Sahoo J, Dash N K and Mohanty B, Indian Doctoral Research in the Field of Library and Information Science: An Empirical Analysis. *International Information & Library Review*, 54 (1) (2022) , 1-16.

11 Panahi S, Iotf M and Ouchi A, Global Research Trends and Hot Topics on Library and Information Science: A Bibliometric Analysis, *Library Philosophy and Practice (e-journal),*7073 (2022) 2-21. DOI: https://digitalcommons.unl.edu/libphilprac/7073

12 Sharma R, Sonkar S K and Kushwaha A K, Bibliometric study of the Ph. D. theses in Library and Information science of Babasaheb Bhimrao Ambedkar University Lucknow, *Library Philosophy and Practice (e-journal),* 5119 (2021). DOI: https://digitalcommons.unl.edu/libphilprac/5119

13 Kumbhar K N, A Bibliometric Study of Ph. D. Awarded Theses in Department of Library and Information Science, *Pearl: A Journal of Library and Information Science*, 13 (1) (2019) 80-85.

14 Gogoi M, Library And Information Science Research (Doctoral Theses) In India: A Bibliometric Study Through Inflibnet Shodhganga, In Proceeding of the paper presented at the conference on 11th Convention PLANNER, Tripura University, Tripura, 15-17 November 2018, p. 85-89.

15 Simte T P and Phuritsabam B, A Bibliometric Study of Research Trend In Library And Information Science In North-Eastern Region of India 1989-2018, In proceeding of the paper presented at International Conference on Innovations in Multidisciplinary Research, Imphal, 23-24 November 2021, p. 6-17.

16 Khaparde V and Ambedkar B, Growth and development of electronic theses and dissertations (ETDs) in India, *Journal of Library and Information Sciences*, 2 (1) (2014) 99-116.

17 Singh, S P and Babbar P, Doctoral Research in Library and Information Science in India: Trends and Issues, *DESIDOC Journal of Library & Information Technology*, 34 (2) (2014).

18 Jaffri S, Shahzad, K and Tariq M, Citation Analysis of Doctoral Theses Submitted During 2007 to 2016 in Pakistani Library Schools. *Library Philosophy and Practice (e-journal)*, 6004 (2021) 1-32. DOI: https://digitalcommons.unl.edu/libphilprac/6004

19 Raza, S N and Warraich N F, Citation Analysis of Information Management Graduates' MPhil and PhD Theses in University of the Punjab, Lahore-Pakistan. *Pakistan Journal of Information Management & Libraries,* 23 (2021) 96-117.

20 Han X, Evolution of research topics in LIS between 1996 and 2019: An analysis based on latent Dirichlet allocation topic model. *Scientometrics*, 125 (3) (2020) 2561-2595.

21 Figuerola C G, García Marco F J and Pinto M, Mapping the evolution of library and information science (1978–2014) using topic modeling on LISA. *Scientometrics*, 112 (3) (2017) 1507-1535.

22 Tunalı V and Bilgin T T, Text mining and social network analysis on computer science and engineering theses in Turkey. In *Proceedings of the paper presented at the conference on 15th International Conference on Computer Systems and Technologies,* 4 April 2014, p.187-193.

23 Papanikolaou N, Pavlopoulos G A, Theodosiou T and Iliopoulos I, Protein–protein interaction predictions using text mining methods. *Methods,* (2015) 47-53.

24 Salloum S A, Al-Emran M, Monem A A and Shaalan K, Using text mining techniques for extracting information from research articles. In *Intelligent natural language processing: Trends and Applications,* 2018 p. 373-397.

25 Zhou J and Fu B Q, The research on gene-disease association based on text-mining of PubMed. *BMC Bioinformatics*, 19 (1) (2018) 1-8.

26 Zhai X, Li Z, Gao K, Huang Y, Lin L and Wang L, Research status and trend analysis of global biomedical text mining studies in recent 10 years. *Scientometrics*, 105 (1) (2015) 509-523.

27 Ferreira-Mello R, Andre M, Pinheiro A, Costa E and Romero C, Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9 (6) (2019), 1-49.

28 Gao W, Text analysis of communication faculty publications to identify research trends and interest. *Behavioral & Social Sciences Librarian, 36* (1) (2017) 36-47.

29 DiMaggio P, Adapting computational text analysis to social science (and vice versa). *Big Data & Society*, 2 (2) (2015).

30 Nagarkar S P and Kumbhar R, Text mining: An analysis of research published under the subject category 'Information Science Library Science'in Web of Science Database during 1999-2013. *Library Review,* 64 (3) (2015) 248-262.

31 Sugimoto C R, Li D, Russell T G, Finlay S C, and Ding Y, The shifting sands of disciplinary development: Analyzing North American Library and Information Science dissertations using latent Dirichlet allocation, *Journal of the American Society for Information Science and Technology*, 62 (1) (2011) 185-204.

32 Lamba M and Madhusudhan M, Metadata tagging of library and information science theses: Shodhganga (2013-2017), *ETD 2018 Taiwan Beyond the Boundaries of Rims and Oceans: Globalizing Knowledge with ETDs*, 2018.

33 Mazumder S and Barui T, Discovering Topics from the Titles of the Indian LIS Theses, *Library Philosophy and Practice (e-journal)*, 5924 (2021) 1-23. DOI: https://digitalcommons.unl.edu/libphilprac/5924

34 Syed S and Spruit M, Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation, *IEEE International Conference on Data Science and Advanced Analytics (DSAA),* 2017 p.165–174. https://doi.org/10.1109/DSAA.2017.61.

35 Sarkar, D Text analytics with Python, 2nd edn (Apress Publication: Banglore), 2019, p. 1-659. Available at https://doi.org/10/1007/978-1-4842-4354-1_2

36 Doig, Christine Introduction to topic modelling in python, PyGotham (2015) Available at https://chdoig.github.io/pygotham-topic-modeling (Accessed on 22 November 2021)

37 Blei D M, Ng A Y and Jordan M I, Latent dirichlet allocation. *Journal of Machine Learning Research*, 3 (Jan) (2003) 993-1022.

38 Khusbu T and Kumar V, Application of text mining techniques on scholarly research articles: methods and tools, *New Review of Academic Librarianship* 28(3) (2022) 279-308. Available at https://doi.org/10.1080/ 13614533. 2021.1918190