DLINE JOURNALS

Journal of Information & Systems Management

ISSN: 2230 – 8776

JISM 2024; 14 (4)

https://doi.org/10.6025/jism/2024/14/4/165-170

Uncovering the Trends in Indian Research: Latent Dirichlet Allocation (LDA) Analysis of UGC-funded Publications

Devanath P R
International Centre for Theoretical Sciences
Department of Library andInformation Science
Tumkur University,
Karnataka. India
devanath.pr@icts.res.in

Rupesh Kumar A
Department of Library and
Information Science
Tumkur University, Karnataka
a.rupeshkumar@gmail.com

Received: 31 May 2024
Revised: 10 July 2024
Accepted: 22 July 2024

ABSTRACT

This study traced the trends in Indian-funded research using UGC-funded publications indexed in Web of Science. Using R programming language, we employed Latent Dirichlet Allocation (LDA) and n-gram modelling of titles of 6094 publications made during 2023. The analysis indicated three important areas of research: Anticancer compounds, Nanoparticles/Carbon Nanoparticles, and Complex Oxide Materials. The high degree of prevalence of the terms "analysis", "synthesis", and "characterization" in the clusters of terms indicated the predominance of the experimental method of research circling the three areas of research stated above. The n-gram modelling revealed "molecular docking" and "molecular docking studies" as the most frequent bigram and trigram. It further showed that "breast cancer" among the types of cancer is the most crucial area of research. UGC mainly funded experimental research on "cancer" during 2023. Extensive research on abstracts and full publications texts could offer a more comprehensive insight into prevailing research trends.

Keywords: Indian-funded research, UGC, Text Mining, Topic modelling, Latent Dirichlet Allocation, LDA

1. Introduction

University Grants Commission (UGC) is crucial in providing financial assistance to universities and colleges in India while ensuring their adherence to regulations. A significant portion of this assistance is distributed as fellowships

and grants. Evaluating the outcomes of the investments is crucial to understanding their impact and productivity, and prioritising the efficient allocation of resources is essential. Therefore, this study aims to identify the research trends in the publications that result from the financial assistance provided by the UGC.

Recently, topic modelling has become an invaluable tool for uncovering valuable insights in digital text (Muchene & Safari, 2021). One among them is Latent Dirichlet Allocation (LDA), a probabilistic generative model of documents proposed by Blei (2003). It is a type of topic model that infers latent topics and their probability distributions in many documents (Tomojiri et al., 2022). We applied the LDA topic modelling method to uncover the latent topics. We conducted text analysis through the n-grams method to highlight the most frequent terms in the UGC-funded research publications.

2. Review of related literature

The concept of topic modelling is not new; it began in the 1990s (Churchill & Singh, 2022), and has been widely embraced by numerous research fields to reveal hidden insights within extensive datasets. A study by Sugimoto et al. (2011) presented a comprehensive historical analysis of research topics in Library & Information Science (LIS) using Latent Dirichlet Allocation (LDA), which highlighted the evolution of research on LIS topics from 1930 to 2009. Lamba & Madhusudhan (2019) Used the LDA method to identify core topics within the DESIDOC Journal of Library & Information Technology and predicted the topics of future research articles using text mining classifiers. Ahmed et al. (2022) used the LDA model and presented statistics on the debt dynamics of Pakistan's economy. Tomojiri et al. (2022) carried out an extensive study and found the temporal trends and geographical distribution of research topics in anthropogenic marine debris using the LDA model. With an exhaustive literature search, it was found that there is considerable scope for applying LDA to uncover latent topics and research areas funded by Indian funding agencies. The present study illustrates this using UGC-funded publications for 2023 as a case.

3. Research Questions

The present study aims to answer the following questions:

· What are the priority areas of research that UGC funded during the year 2023?

· What do authors of UGC-funded publications employ the most prominent research method?

· What are the most frequently appearing terms in the titles of UGC-funded publications?

4. Data and Methods

4.1. Data

We collected records of 6094 UGC-funded research publications made during the year 2023 from the Web of Science Core Collection (Clarivate Analytics, 2024). We employed advanced search using the authors' country affiliation (India) to find publications by Indian authors and then refined the search to find research publications funded by UGC. The data was then exported as a Comma Separated Values (CSV) file. The titles of the publications were selected for the analysis.

4.2. Methods

The 'R' programming language was used as a platform for the analysis and visualization (R Core Team, 2024). First, the data were preprocessed to bring different forms of a word to their root level in the dataset. This involves converting the text to lowercase and removing all punctuation marks and numbers. Stop words were removed from the dataset. Furthermore, we Lemmatized the corpus using "*textstem*" library package for keeping the words that belong only to nouns, adjectives, verbs, adverbs, person, language, quantity, and geographical location. (Mazumder & Barui, 2021). This preprocessed data was then employed

for the LDA model to determine the hidden topics in the dataset. We extracted ten clusters of topics. Each cluster consisted of ten terms. We also extracted bigrams and trigrams to find the most frequently discussed topics.

5. Results and Discussion

5.1. Priority Research Areas

We use LDA to answer the first research question. Table 1 lists ten topics with ten terms in each cluster. Unlike keyword or word frequency analysis, LDA uncovers latent topics by providing clusters of closely associated terms. It may be broadly noted from the cluster of topics given in the table that the primary research area of UGC-funded publications has been Biochemical and Molecular applications with a focus on anticancer compounds. From a closer analysis of clusters of terms generated, we infer that UGC-funded research during 2023 revolves around the following topics:

1. Chemical Synthesis and Molecular Investigations
2. Novel Anticancer Compounds and Structural Investigations
3. Nanoparticles Property and Evaluation
4. Molecular Modeling and Potential Applications of Synthesized Compounds
5. Analysis and Synthesis of Carbon Nanoparticles
6. Characterization of Complex Oxide Materials for Potential Cancer Applications
7. Synthesis of Acid-Based Compounds for Potential Cell Applications
8. Complex Optical Nanoparticles and their Applications
9. Effects of Acid-Base Interactions on Energy-Related Reactions
10. Complex Metal Oxide Materials for Potential Cell Applications

Topic-1	Topic-2	Topic-3	Topic-4	Topic-5	Topic-6	Topic-7	Topic-8
synthesis	synthesis	Effect	synthesis	synthesis	property	property	Complex
use	effect	Use	effect	property	effect	use	Use
base	use	structural	base	use	base	cell	characterizatio
acid	activity	Activity	potential	cell	complex	synthesis	nanoparticle
complex	novel	nanoparticle	analysis	potential	synthesis	acid	Optical
molecular	cancer	Base	model	india	characterization	potential	Activity
investigation	structural	Property	molecular	effect	cell	model	Cell
india	base	Efficient	evaluation	nanoparticle	oxide	activity	Effect
energy	performance	evaluation	temperature	carbon	analysis	structural	Efficient
performance	enhance	Detection	role	analysis	cancer	base	molecular

Table 1. Top Ten topics with related terms

From the LDA analysis, we deduced that the priority research areas of UGC-funded publications made during 2023 were:

1. Anticancer compounds

2. Nanoparticles / Carbon Nanoparticles

3. Complex Oxide Materials

5.2. Predominant Research Method

The second research question pertains to the most predominant research method. Figure 1 shows the topic modelling of UGC-funded publications with the degree of strength of each term in a topic. It may be observed that the term *synthesis* is the most representative term in several clusters, indicating its importance as a method employed in research, next only to the term *property*. Together, they provide insight into the focus of research in UGC-funded publications. Overall, the terms *analysis, synthesis, and characterization* indicate the predominance of an experimental research method circling around the three areas of research indicated above.

Figure 1. Topic Modelling of UGC-funded Publications

5.3. Frequency of Terms

To answer the third research question, we used the n-grams model to find the most frequent terms in the titles of UGC-funded publications. We generated the most frequent bigrams (two-worded terms) and trigrams (three-worded terms). Figure 2 shows the cluster of bigrams.

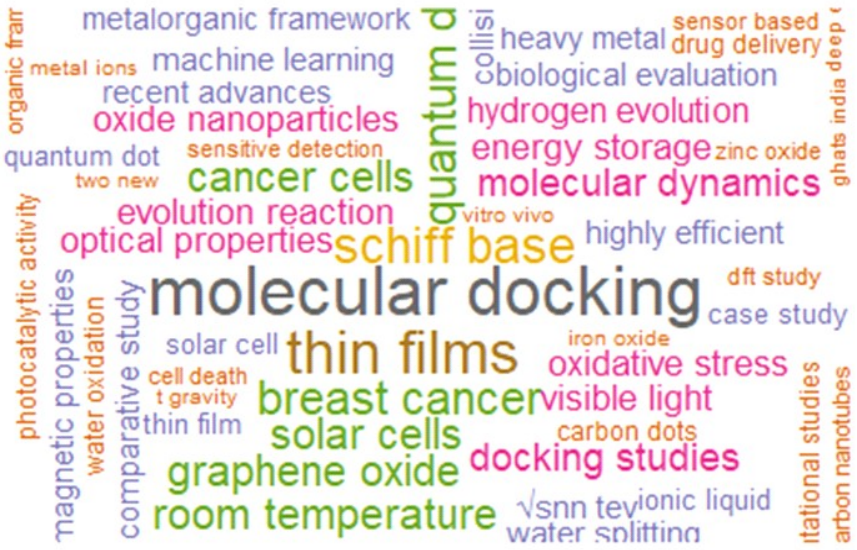

Figure 2. Bigram Model of UGC-funded Publications

As may be noted from the figure, "molecular docking" is the most frequent bigram, followed by "thin films". "Molecular docking" is well supported by the bigrams "docking studies" and "molecular dynamics". We may also observe another set of bigrams showing approximately equal frequency: "oxide nanoparticles", "hydrogen evolution", "evolution reaction", "energy storage", "optical properties", "oxidative stress", and "visible light". Similar sets with approximately the same frequency may also be noticed ("cancer cells", "breast cancer", etc.).

The trigram model shown in Figure 3 further justifies the bigram model. "Molecular docking studies" is the most frequently used trigram in the titles of UGC-funded publications, followed by "density functional theory" and "oxygen evolution reaction". The n-gram modelling of titles reveals that "breast cancer" among the types of cancer is the most crucial area of research.

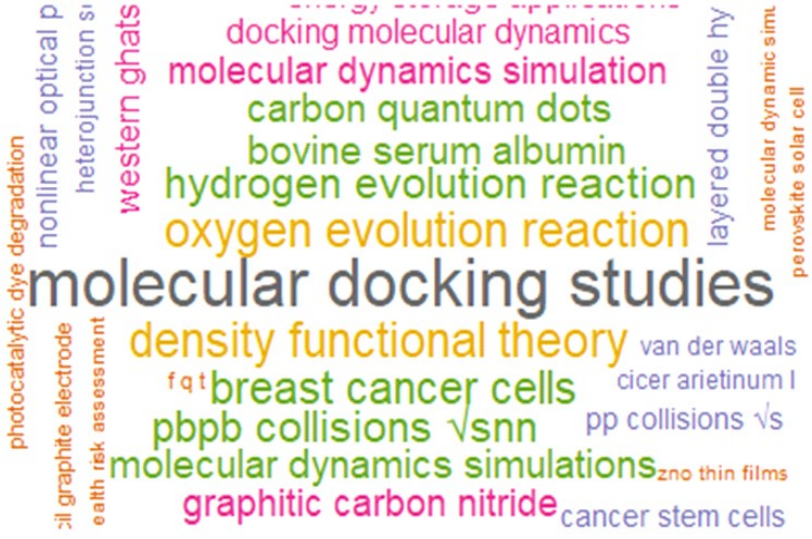

Figure 3. Trigram Model of UGC-funded Publications

6.Conclusion

The study presents different approaches to identify the topical trends in research using Web of Science-indexed UGC-funded publications of 2023 as a case. The results of the present study reveal that UGC has funded mainly experimental research on "cancer". We argue that the Latent Dirichlet Allocation (LDA) method coupled with n-gram modelling of publications can uncover the latent topics in research. The support for cancer research is important because cancer is among the leading causes of death worldwide. (*Understanding Cancer*, 2024). In-depth analyses of abstracts and full-text publications can provide more efficient insight into research trends and patterns.

References

[1] Ahmed, F., Nawaz, M., Jadoon, A. (2022). Topic modeling of the Pakistani economy in English newspapers via latent Dirichlet allocation (LDA). *SAGE Open, 12*(1). https://doi.org/10.1177/21582440221079931

[2] Analytics, C. (2024). Certain data included herein are derived from Clarivate™ (Web of Science™). © Clarivate. All rights reserved. [Dataset]. https://clarivate.com/

[3] Blei, D. M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[4] Churchill, R., Singh, L. (2022). The evolution of topic modeling. *ACM Computing Surveys, 54*(10s), 1–35. https://doi.org/10.1145/3507900

[5] Lamba, M., Madhusudhan, M. (2019). Metadata tagging and prediction modeling: Case study of DESIDOC Journal of Library and Information Technology (2008-17). *World Digital Libraries, 12*(1), 33–89. https://doi.org/10.18329/09757597/2019/12103

[6] Mazumder, S., Barui, T. (2021). Discovering topics from the titles of the Indian LIS theses. *Library Philosophy and Practice*. https://digitalcommons.unl.edu/libphilprac/4574/

[7] Muchene, L., Safari, W. (2021). Two-stage topic modeling of scientific publications: A case study of University of Nairobi, Kenya. *PLOS ONE, 16*(1), e0243208. https://doi.org/10.1371/journal.pone.0243208

[8] Sugimoto, C. R., Li, D., Russell, T. G., Finlay, S. C., Ding, Y. (2011). 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), 185–204. https://doi.org/10.1002/asi.21435

[9] Team, R. C. (2024). *R: A language and environment for statistical computing (4.4.1)* [R]. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[10] Tomojiri, D., Takaya, K., Ise, T. (2022). Temporal trends and spatial distribution of research topics in anthropogenic marine debris study: Topic modeling using latent Dirichlet allocation. *Marine Pollution Bulletin, 182*, 113917. https://doi.org/10.1016/j.marpolbul.2022.113917

[11] Understanding cancer. (2024). *National Cancer Institute*. https://www.cancer.gov/about-cancer/understanding/statistics#:~:text=Cancer%20is%20among%20the%20leading,million%20cancer%2Drelated%20deaths%20worldwide