World Digital Libraries 12(1): 33–89 (2019)
DOI: 10.18329/09757597/2019/12103

Metadata Tagging and Prediction Modeling: Case Study of DESIDOC Journal of Library and Information Technology (2008–17)

M. Lamba
Research Scholar, Department of Library and Information Science,
University of Delhi, Delhi-110007
(E): lambamanika07@gmail.com

M. Madhusudhan
Associate Professor and former Deputy Dean (Academics),
Department of Library and Information Science, University of Delhi, Delhi
(E): mmadhusudhan@libinfosci.du.ac.in

Abstract

The present paper describes the importance and usage of metadata tagging and prediction modeling tools for researchers and librarians. 387 articles were downloaded from DESIDOC Journal of Library and Information Technology (DJLIT) for the period 2008–17. This study was divided into two phases. The first phase determined the core topics from the research articles using Topic-Modeling-Toolkit (TMT), which was based on latent Dirichlet allocation (LDA), whereas the second phase employed prediction analysis using RapidMiner toolbox to annotate the future research articles on the basis of the modeled topics. The core topics (tags) were found to be digital libraries, information literacy, scientometrics, open access, and library resources for the studied period. This study further annotated the scientific articles according to the modeled topics to provide a better searching experience to its users. Sugimoto, Li, Russell, *et al.* (2011), Figuerola, Marco, and Pinto (2017), and Lamba and Madhusudhan (2018) have performed studies similar to the present paper but with major modifications.

Keywords: Metadata tagging, DESIDOC Journal of Library and Information Technology (DJLIT), Latent Dirichlet allocation (LDA), Information retrieval, Naive Bayes, Prediction modeling, Support Vector Machine (SVM), Text mining, Topic modeling

Introduction

Information in the form of text and images is generated in an enormous amount and stored in archives by many organizations. This poses the challenge of managing such loads of data and extracting the appropriate knowledge for decision-making. New tools and techniques are required to manage this explosion of electronic documents in a better way. Machine learning and statistics in a decade have developed new techniques for finding patterns of words in a large collection of documents for effective information retrieval and one of these techniques is popularly known as topic modeling. Topic modeling has a wide range of applications, such as tag recommendation, text categorization, keyword extraction, and similarity search. It can be broadly applied in the fields of text mining, information retrieval (IR), and statistical language modeling. In the early 2000s, performance of the prediction model in IR helped us to answer the following question (Cronen-Townsend, Zhou, and Croft, 2002): Is it possible to predict how good a result returned by an IR system can be even before presenting it to the user or running it on the IR system?

This study provides a method to identify the disciplinary identity of research in India by finding the main topics (tags) in DESIDOC Journal of Library and Information Technology (DJLIT) for the epoch of 2008–17. 'DJLIT is a peer-reviewed, open access, bi-monthly journal. It publishes original research and review papers related to library science and IT applied to library activities, services, and products. This journal is meant for librarians, documentation and information professionals, researchers, students, and others interested in the field. Furthermore, it is one of the prestigious journals in the field of LIS (Library and Information Science) in India' (DJLIT, 2019). Latent Dirichlet Allocation (LDA) modeling technique was used to model the topics (tags), and Support Vector Machine (SVM) and Naive Bayes were used as the predictors to predict the performance of created models. The paper has both methodological and applicative objectives: (i) to identify the main topics (tags) of the scientific articles, (ii) to annotate the scientific articles according to the modeled topics, and (iii) to statistically evaluate the predictive models. The performance of the built models was analysed by the statistical evaluation measures (accuracy, precision, and recall). This study addresses gaps in the literature by utilizing a technique that explores the disciplines and uses them as tags in context to the Indian journals. This work has a broad application to those interested in information retrieval, semantic web, and linked data.

Review of Related Literature

There are many articles existing on this subject, and a few important studies have been reviewed with regard to the application of LDA, topic modeling, and text mining. A classic research paper on LDA is by Blei, Ng, and Jordon (2003), who showed efficient approximate inference techniques based on vibrational methods and an EM (expectation–maximization) algorithm for empirical Bayes parameter estimation. They reported results in document modeling, text classification, and collaborative filtering, comparing to a mixture of unigrams model and the probabilistic LSI (Latent Semantic Indexing) model. Some of the selected articles that applied algorithms other than LDA in their study for topic modeling are those by Muresan and Harper (2004), Zhang, Sim, Su, *et al.* (2011), Hurtado, Agarwal, and Zhu (2016), and Nikolenko, Koltcov, and Koltsova (2017).

Some of the selected articles that showed the implementation of LDA in their studies are by Wu, Kuang, Hong, *et al.* (2019), who proposed a model based on Knowledge Organization System (KOS) and LDA and applied it in detecting burst topic and its semantic information relationship in the cancer field. Experiments showed that the model played an important role in topic recognition, evolution recognition,

and visualization. Furthermore, they showed that the application of KOS combined with LDA can effectively remove the noisy concept from semantic layer and showed a good effect. Ma, Li, Ou, *et al.* (2018) presented a method of evaluating the competitiveness of research institutions based on research topic distribution. They used a LDA topic model to obtain a paper-topic distribution matrix to objectively assign the academic impact of papers (e.g., number of citations) to research topics. Then the method was used to calcula te the competitiveness of each research institution on each research topic with the help of an institution-paper matrix. Finally, the competitiveness and the research strength and weakness of the institutions were defined and characterized. Figuerola, Marco, and Pinto (2017) provided an overview of the bibliometric study of the domain of library and information science (LIS), with the aim of giving a multidisciplinary perspective of topical boundaries and main areas and research tendencies. Based on a retrospective and selective search, they obtained the bibliographical references (title and abstract) of academic production on LIS in the database LISA (Library and Information Science Abstracts) for the period 1978–2014, which ran to 92,705 documents. In the context of the statistical technique of topic modeling, they applied Latent Dirichlet Allocation to identify the main topics and categories in the corpus of documents analysed. The quantitative results revealed the existence of 19 important topics, which could be grouped together into four main areas: processes, information technology, library, and specific areas of the information application. Yau, Porter, Newman, *et al.* (2014) investigated the methods, including LDA and its extensions, for separating a set of scientific publications into several clusters. To evaluate the results, they generated a collection of documents that contained academic papers from several different fields and examined whether papers in the same field would be clustered together. They further explored potential scientometric applications of such text analysis capabilities. Yezheng *et al.* (2019) proposed an interactive strategy to generate high-quality topics with clear meanings by integrating subjective knowledge derived from human experts and objective knowledge learned by LDA. The proposed interactive latent Dirichlet allocation (iLDA) model developed deterministic and stochastic approaches to obtain subjective topic-word distribution from human experts, combined the subjective and objective topic-word distributions by a linear weighted-sum method, and provided the inference process to draw topics and words from a comprehensive topic-word distribution. The proposed model was a significant effort to integrate human knowledge with LDA-based models by the interactive strategy. The experiments on two real-world corpora showed that the proposed iLDA model could draw high-quality topics with the assistance of subjective knowledge from human experts. It was robust under various conditions and offered fundamental supports for the applications of LDA-based topic modeling.

The studies that applied prediction modeling with respect to topic modeling are by Özmutlu and Çavdur (2005), who proposed an artificial neural network to identify automatically topic changes in a user session by using the statistical characteristics of queries, such as time intervals and query reformulation patterns. Benton, Paul, Hancock, *et al.* (2016) considered survey prediction from social media. They used topic models to correlate social media messages with survey outcomes and to provide an interpretable representation of the data. Rather than relying on fully unsupervised topic models, they used existing aggregated survey data to inform the inferred topics, a class of topic model supervision referred to as collective supervision. They introduced and explored a variety of topic model variants and provided an empirical analysis, with conclusions of the most effective models for this task.

As textual data is unstructured in nature, there are very few predictive models that can successfully check the performance of the model. SVM is one of the few predictive models that can analyse textual data and give good results and has been used in this study. SVM is a form of linear classifiers. Linear classifiers in the context of text documents are models that make a classification decision based on the value of the linear combinations of the features of the document. One advantage of the SVM method is that it is quite robust to high dimensionality; that is, learning is almost independent of the dimensionality of the feature space (Allahyari, Pouriyeh, Assefi, *et al.*, 2017). It rarely needs feature selection since it selects data points (support vectors) required for the classification (Hotho, Nürnberger, and Paaß, 2005). Joachims (1998) found text data to be an ideal choice for SVM classification due to sparse high-dimensional nature of the text with few irrelevant features. SVM methods have been widely used in many application domains, such as pattern recognition, face detection, and spam filtering (Allahyari, Pouriyeh, Assefi, *et al.*, 2017).

Naive Bayes is the second predictive model that is used in this study for textual data prediction analysis. It is based on the Bayesian classification and represents both a supervised learning method and a statistical method for classification. It is also among the most successful known algorithms for learning to classify text documents like SVM. It is particularly suited when the dimensionality of inputs is high. Parameter estimation for Naive Bayes models uses the method of the maximum likelihood. In spite of over-simplified assumptions, it often performs better in many complex real-world situations. One of the advantages of using this model is that it requires a small amount of training data to estimate the parameters. This probabilistic approach makes assumptions about how data (words in documents) are generated and proposes a probabilistic predictive model based on these assumptions. Then it uses a set of training examples to estimate the parameters of the model. It models the distribution of documents in each class using a probabilistic model by assuming that the distribution of different terms is independent of each other. Even though the Naive Bayes assumption is false in many real-world applications, it performs surprisingly well. There are two main models commonly used for Naive Bayes classifications: multi-variant Bernoulli model and multinomial model. Both models aim at finding the posterior probability of a class based on the distribution of the words in the document. The difference between these two models is one model takes into account the frequency of the words, whereas the other does not (Allahyari, Pouriyeh, Assefi, *et al.*, 2017).

Methodology

A total of 387 articles were downloaded from DJLIT journal for the period 2008–17. The data were analysed according to the LDA probabilistic topic modeling method. For the 10 years period, five topics were identified. Each topic contained a probability value. These topics were ranked according to their probability values and the top five topics were selected as the most representative one. Top five words were chosen as most representative of the topic according to their probability. Two text mining tools were used, one for probabilistic topic modeling and the other for prediction analysis. This paper is restricted to articles published in DJLIT journal from 2008 to 2017 in the English language excluding the guest editorials and special editions.

Latent Dirichlet Allocation

There are many algorithms that can be used to obtain topic models, such as latent semantic analysis (LSA), probabilistic latent semantic analysis (PLSA), latent Dirichlet allocation (LDA), and correlated topic model (CTM). This paper focuses on the use of LDA, which is based on the Dirichlet distribution, a family of

continuous multivariate probability distribution parameterized by a vector alpha of positive reals (Blei, Ng, and Jordan, 2003). In LDA, each document is viewed as a mixture of topics present in the corpus (KDnuggets, 2017). Each document gets represented as a pattern of LDA topics, making every document appear. LDA automatically infers the topics discussed in a collection of documents and these topics can be used to summarize and organize documents. It is based on probabilistic modeling and observed variables are the bags of words per document, whereas hidden random variables are the topic distributions per document. The main goal of LDA is to compute the posterior of the hidden variables given the value of the observed variables (Allahyari, Pouriyeh, Assefi, et al., 2017). The assumptions of LDA are as follows: (i) documents similar topics will use similar groups of words, (ii) documents are a probability distribution over latent topics, and (iii) topics are probability distributions over words (Figure 1).

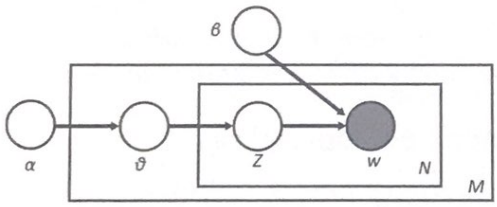

Figure 1: Graphical model representation of latent Dirichlet allocation
Source: Blei, Ng, and Jordon (2003)

The outer box in Figure 1 represents documents, while the inner box represents the repeated choice of topics and words within a document. In Figure 1, α is the parameter of Dirichlet prior on the per-document topic distribution, β is the parameter of Dirichlet prior on per-topic word distribution, θ is the topic distribution for the document, z is the topic for the nth word in the document, w is the specific word, N is the total number of words, and M is the total number of documents in the corpus (Blei, Ng, and Jordon, 2003).

Prediction Modeling

Prediction modeling is the process of creating, testing, and validating a model to predict the probability of an outcome to it best. Predictive analytics uses a number of modeling methods from machine learning, artificial intelligence, and statistics on the basis of testing, validation, and evaluation using the hit and trial method. Models can use one or more classifiers in order to determine the probability of the data. Every model has its own strengths and weaknesses and is best suited for particular types of problems. There are mainly three types of model categories: (i) *Predictive models*, which analyse the past performance for future predictions, (ii) *descriptive models*, which quantify the relationships in data in a way that is often used to classify data into set groups, and (iii) *decision models*, which describe the relationship between all the elements of a decision in order to predict the results of decisions involving many variables (Predictive Analytics Today, 2017).

The basic steps of the prediction modeling include the following: (i) *Creating the model* – create a model to run one or more algorithms on the data set, (ii) *testing the model* – test the model on the data set where, in some scenario, testing is done on past data to see how best the model predicts, (iii) *validating the model* – visualization tools are used to validate the model run results, and (iv) *evaluating the model* – evaluate the best fit model from the models used and choose the model right fitted for the data. It is an iterative process that often trains the model using multiple models on the same data set and chooses the best fit model (Predictive Analytics Today, 2017).

The present study employs SVM and Naive Bayes predictors using RapidMiner toolbox. As additional data become available in future, the predicted statistical analysis model can be validated or revised accordingly

Accuracy, Precision, and Recall

Performance of prediction models is judged using measures such as accuracy, precision, and recall. *Accuracy* is defined as the proportion of correct classification (true positive, true negatives) from the overall number of cases, whereas *precision* is defined as the proportion of correct positive classification (true positives) from cases that are predicted as positives. In other words,

$$\frac{\text{Number of relevant scientific articles retrieved}}{\text{Total number of scientific articles retrieved}} \times 100$$

The recall is the proportion of correct positive classification (true positives) from cases that are actually positives. In other words,

$$\frac{\text{Number of relevant scientific articles retrieved}}{\text{Total number of relevant scientific articles in a collection}} \times 100$$

Topic-Modeling-Toolkit

Topic-Modeling-Toolkit (TMT) (Google Code Archive, 2017) is powered by Java, a graphical interface tool for LDA topic modeling. It is a simple GUI-based application for topic modeling that uses the popular MALLET toolkit for the back-end (Abinaya and Winster, 2014). Topic models provide a simple way to analyse large volumes of unlabeled text. A 'topic' consists of a cluster of words that frequently occur together. Using contextual clues, topic models can connect words with similar meanings and distinguish between uses of words with multiple meanings. The GUI has two main windows: basic and advanced (Google Code Archive, 2017). All the 387 articles were converted into text format and then processed using Topic-Modeling-Toolkit. A total of 10 h was spent for converting the aforesaid process. In the toolkit, following parameters were being fixed for the study: (i) number of topics: 5, (ii) number of iterations: 200, (iii) number of topic words printed: 5, and (iv) topic proportion threshold: 0.05.

RapidMiner

'RapidMiner' is an easy-to-use visual environment for predictive analytics (RapidMiner, 2019). The research articles downloaded from DJLIT were divided into two sets: training and test randomly. These articles were divided in the ratio of 60% to 40% and then fed into the process. The process included the following steps:

1. Pre-processing of the documents (i.e., tokenization, stemming, filtering stop-words, transforming the cases, and generating *n*-grams per terms)
2. Splitting of the data into two subsets
3. Training and testing using split validation
4. Applying SVM and Naive Bayes classifiers
5. Measuring the performance of SVM and Naive Bayes classifiers

Results

This section analyses and interprets the various output files generated by Topic-Modeling-Toolkit and the performance analysis of the prediction model by RapidMiner toolbox.

Modeled Topics/Tags

Table 1 summarizes the LDA results generated by the TMT. The topics a–e for the 10-year period are organized in the descending order according to their probability values ('a' having the highest probability value). Evidence from the high-loading keywords and most representative articles (Table 1) reveals that Topic-a is about digital libraries. Topic-b is about information literacy. High-loading articles in Topic-c show a focus on bibliometrics, with an emphasis on scientometrics particular to India. Topic-d displays a focus on open access, with a focus on books, websites, and universities. The representative articles from Topic-e are on library resources with a focus on students and universities.

Table 1: Latent Dirichlet allocation results for the period 2008–17 (387 articles)

Topic-a	Topic-b	Topic-c	Topic-d	Topic-e
Library	Information	Research	Access	Library
Data	Knowledge	Publications	HTTP	Information
Software	Learning	Papers	Books	Resources
System	Education	Journals	Open	University
Web	Online	University	Web	Students

Representative Articles

Table 2 summarizes the top five most representative titles under each modeled topic generated by the toolkit for the period 2008–17. The titles are organized in the descending order according to their probability values, that is, the first title under a particular topic is having the highest probability value followed by the rest, thus ranked according to their topic proportion (Appendix 1).

Table 2: Titles corresponding to the representative articles for 2008–17 (387 articles)

	Representative Title 1	Representative Title 2	Representative Title 3	Representative Title 4	Representative Title 5
Topic-a	Application Domain and Functional Classification of Recommender Systems—A Survey	Content-Based Document Recommender System for Aerospace Grey Literature: System Design	Web Interface in Library Management Software Systems	From Clay Tablets to Web: Journey of Library Catalogue	Nanotechnology Ontology: Semantic Access to Information in the Nano World
Topic-b	An Overview of Online Exhibitions	Animated and Hypertext User Interfaces: A Comparative Study	Training Needs of School Librarians in India	ADDIE: Designing Web-enabled Information Literacy Instructional Modules	Information Literacy in India and Germany: University Libraries as Activators of Life-long Learning
Topic-c	Indian Computer Science Research Output during 1999-2008: Qualitative Analysis	Research Trends in Nanoscience and Nanotechnology in India	Bangladesh: A Scientometric Analysis of National Publications Output in S&T, 2001-10	Scientometric Dimensions of Neutron Scattering Research in India	Research Activities in Biochemistry, Genetics and Molecular Biology during 1998-2007 in India: A Scientometric Analysis

Table 2: Contd....

	Representative Title 1	Representative Title 2	Representative Title 3	Representative Title 4	Representative Title 5
Topic-d	Use of Social Media by Online Newspapers in Saudi Arabia	Open Access to Electronic Theses and Dissertations	Mapping of E-books in Science & Technology: An Analytical Study of Directory of Open Access Books	Availability of Open Access Books in DOAB: An Analytical Study	INDEST-AICTE Consortium: Present Services and Future Endeavours
Topic-e	Web as a Learning Resource at the Medical College Libraries in Coastal Karnataka: Perception of Faculty and Students	LSQA Scale: A Tool for Measuring Users' Perceptions of Service Quality in Libraries	Awareness and Use of E-resources: A Case Study of Mohinder Singh Randhawa Punjab Agricultural University Library, Ludhiana	Information Management Skills Required by the Minority Libraries in Kolkata and Hooghly Districts, West Bengal	Usage of Electronic Resources at Dr T.P.M. Library, Madurai Kamaraj University: A Case Study

Topic Proportion

Appendix 1 gives the percentage composition of all the 387 articles mined from the DJLIT journal under the modeled topics. It shows that an article can be composed of single topic or a mixture of topics. But the core topic is decided on the basis of the highest value of topic proportion of the modeled topic. Users can study different articles at a glance according to different modeled topics and choose the desired article for reading related to his/her research interest instead of looking through all the articles one by one in DJLIT for the studied period. The articles on the basis of topic proportion can be searched in two ways: (i) based on the core topic, with this approach all the similar articles on a particular topic are placed together and can be easily retrieved in spite of ambiguous title names and (ii) based on the research article published in DJLIT during the period 2008–17. For instance, one can either see articles on scientometrics by directly scanning through the Topic-c column in Appendix 1 with a value of ≥0.5 or can directly see the topic composition of an article of interest. Meta-tagging of the articles using modeled topics not only saves the time of users but also helps in organizing and managing the e-resources.

Prediction Modeling

Two prediction models using SVM and Naive Bayes classifiers were created and tested (Figures 2 and 3). Each model was created using 387 articles with the modeled topics, where 60% (231) of the data was allocated to the training set and 40% (156) was allocated to test set randomly using the split validation technique. Once the parameters of the models were finalized, the testing set was run through the model. The actual test class was compared to the predicted class to determine the accuracy, precision, and recall values for each model. Figures 2 and 3 show the results of the tested data set against the trained data set from the predictive models with 87.18% accuracy for the SVM model and 82.69% for the Naïve Bayes model.

Model Comparison

Using Support Vector Machine (linear regression) and Naive Bayes as the training classifiers with features such as document

accuracy: 87.18%						
	true a	true b	true c	true d	true e	class precision
pred. a	24	3	0	2	0	82.76%
pred. b	0	22	0	0	1	95.65%
pred. c	0	1	28	0	1	93.33%
pred. d	2	1	1	16	1	76.19%
pred. e	0	3	0	4	46	86.79%
class recall	92.31%	73.33%	96.55%	72.73%	93.88%	

Figure 2: Performance output by SVM prediction model

accuracy: 82.69%						
	true a	true b	true c	true d	true e	class precision
pred. a	21	4	1	2	1	72.41%
pred. b	2	23	0	0	2	85.19%
pred. c	0	1	27	1	2	87.10%
pred. d	1	0	1	17	3	77.27%
pred. e	2	2	0	2	41	87.23%
class recall	80.77%	76.67%	93.10%	77.27%	83.67%	

Figure 3: Performance output by Naive Bayes prediction model

frequencies and term frequencies, a correlation between predicted and actual performances was obtained in the conducted experiment. With SVM model, our approach yielded decent prediction quality. The resultant prediction quality of SVM was better than that of Naive Bayes predictor. The resultant SVM retrieval effectiveness of the corpus on which search was performed was attested. Therefore, a clear advantage of our approach is that it will help in providing a successful future automated classification of the unseen incoming articles in DJLIT according to the five modeled topics.

Discussion

Topics evolve over time. New topics emerge and old ones become obsolete. Topic modeling helps researchers not only to determine the trending themes with respect to their field of interest but also to identify new concepts or fields. For instance, computer science emerged as a discipline during the 1950s and there was no evidence of computer science as an academic stream before 1950s; therefore, the emergence of new fields can be identified faster using topic modeling. Topic mining can be described as a precursor for many subsequent tasks, such as information retrieval, classification, sentiment analysis, and opinion mining. Moreover, as demonstrated by Kanojia, Joshi, Bhattacharyya, *et al.* (2015), topic modeling can also be applied in creating coarse bilingual dictionaries and specific field ontologies.

Platforms such as social networks, websites, blogs, and research journals generate an enormous amount of data in the form of unstructured text and it is essential to analyse, synthesize, and process such data for knowledge discovery and decision-making. In libraries, topic

mining can be extensively used for searching, exploring, and recommending articles. Librarians can search similar readings based on the modeled topic to give suggestion to users to indicate what they could be interested to read next. Topic modeling not only helps the librarians in finding a similar type of journal articles, newspapers, websites, blogs, and other e-resources in text format and prioritizing them based on the modeled topics but also helps to provide reference service, current awareness service, and selective dissemination of information service efficiently in a very short period of time. Thus, librarians can create a list of interested readings using the topic modeling tool based on the queries submitted by users. However, this study is restricted to just articles and it can be extended to topic mining of data retrieved from library blogs, news, literature, tweets, social platforms, sentiment analysis or opinion analysis. This method not only is a quick information retrieval tool for researchers but also fulfils the fourth law of library science, that is, save the time of the reader (Ranganathan, 1957).

This study was divided into two phases. The first phase was to determine the major themes of the DJLIT articles under which the articles were placed. On the basis of the five topics and the research articles, which were representative of each modeled topic in the DJLIT journal, the core topics (tags) for the period 2008–17 were determined. Furthermore, each research article published in DJLIT during the period 2008–17 was broken down into various topic proportions of percentage probabilities. Thus, all the 387 articles for the studied period were segregated on the basis of the five modeled topics (Appendix 1). These findings can be mentioned on the DJLIT website, which can ultimately help the user in faster information retrieval.

The second phase of the study was prediction modeling. Predictive modeling is nearly absent in many scientific fields as a tool for developing theory. One of the reasons could be that prediction is often considered unscientific like in social sciences (Shmueli, 2010). Therefore, this paper attempts to apply the prediction modeling to scientific articles in the field of social sciences, particularly LIS, and tries to accurately predict the placement of future research articles going to be published in DJLIT under the five modeled topics ('a' to 'e') on the basis of the performances of the models studied. Hence, the comparison between the two prediction models (Figures 2 and 3) showed that the SVM performed better than the Naive Bayes model and can be used to provide a successful future automated classification of DJLIT under the five modeled topics ('a' to 'e'). One problem with prediction modeling was that the data set was not truly a representative of DJLIT. The training of the model to learn and fit the parameters could be done perfectly if more data are taken into account. Thus, this study can be extended to perform probabilistic topic modeling of DJLIT from 1981 till present so that the result output of the prediction model can be relied upon with total confidence.

The studies conducted by Sugimoto, Li, Russell, *et al.* (2011) and Figuerola, Marco, and Pinto (2017) are notable studies in LIS similar to this study with major modifications, where Sugimoto, Li, Russell, *et al.* (2011) identified the changes in dominant topics in LIS over time by analysing 3121 doctoral dissertations completed between 1930 and 2009 at North American Library and Information Science programs. The authors utilized LDA to identify latent topics diachronically and representative dissertations of those topics. Figuerola, Marco, and Pinto (2017) provided an overview of the bibliometric study in the domain of LIS, with the aim of providing a multidisciplinary perspective of the topical boundaries and the main areas and research tendencies. Based on a retrospective and selective search, they obtained the bibliographical references (title and abstract) of academic production on LIS in the database LISA for the period 1978–2014, which ran to 92,705 documents using LDA. Further the authors (Lamba and Madhusudhan, 2018) have performed a study similar to the present paper,

but it was on electronic theses and dissertations (ETDs) rather than on research articles.

Conclusion

Topic modeling acts as a text mining tool to process, organize, manage, and extract knowledge from a large amount of textual data present in various databases. It is based on probabilistic modeling and used to discover hidden structures in large archives of documents on the basis of similar patterns of word usage in each document. It is typically used to determine the underlying topics in text documents. Topic modeling is based on the approaches that make use of the latent topic layer between words and documents to capture the text semantics-based relationships between entities (Benton, Paul, Hancock, *et al.*, 2016). It has multidisciplinary applications and can be applied to other fields to provide a large-scale overview of topical changes over time.

The major limitations of the study will include the identification of an appropriate number of topics for the articles before performing the latent Dirichlet allocation, the incompetence of the Dirichlet topic distribution to correlate among topics, and manual interpretation and labeling of 'topics'. The core topics for DJLIT for the studied period were found to be *digital libraries, information literacy, scientometrics, open access,* and *library resources.* The findings from the present study can be mentioned on the DJLIT website, which can ultimately help the user in faster information retrieval. DJLIT is offering a service where it has a link to 'related article', but not all the research papers have that link. Also, those related articles are retrieved on the basis of similar keywords in the title but not on the basis of the main concept/theme behind those articles. Thus, this study tries to fill that gap and provides a solution to suggest topics (tags) on the basis of full text analysis of the articles. Furthermore, the performance of the prediction model is analysed on the basis of statistical evaluation measures, namely, accuracy, precision, and recall, and this prediction modeling of DJLIT helps in targeting the classification of future scientific articles and acts as a decision support in the organization of the scientific articles.

References

Abinaya, G. and Winster, S. G. 2014. Event identification in social media through latent dirichlet allocation and named entity recognition. In *IEEE International Conference on Computer Communication and Systems ICCCS14,* Chennai, India. Details available at http://doi.org/10.1109/ICCCS.2014.7068182, last accessed on 6 June 2019

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. 2017. A brief survey of text mining: classification, clustering and extraction techniques. Details available at http://arxiv.org/abs/1707.02919, last accessed on 27 March 2019

Benton, A., Paul, M. J., Hancock, B., and Dredze, M. 2016. Collective supervision of topic models for predicting surveys with social media. In *Thirtieth AAAI Conference on Artificial Intelligence (AAAI),* pp. 2892–2898

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* **3**(1): 993–1022

Cronen-Townsend, S., Zhou, Y., and Croft, W. B. 2002. Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* New York, pp. 299–306

DJLIT. 2019. DESIDOC Journal of Library & Information Technology: About the Journal. Details available at https://publications.drdo.gov.in/ojs/index.php/djlit/about, last accessed on 27 March 2019

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

Google Code Archive. 2017. Long-term storage for Google Code Project Hosting. Details available at https://code.google.com/archive/p/topic-modeling-tool/, last accessed on 27 November 2017

Hotho, A., Nürnberger, A., and Paaß, G. 2005. A brief survey of text mining. *LDV Forum* **20**(1): 19–62

Hurtado, J., Agarwal, A., and Zhu, X. 2016. Topic discovery and future trend forecasting for texts. *Journal of Big Data* **3**(7): 1–21

Joachims, T. 1998. Text categorization with support vector machines: learning with many relevant features. Details available at https://link.springer.com/chapter/10.1007/BFb0026683, last accessed on 27 March 2019

Kanojia, D., Joshi, A., Bhattacharyya, P., and Carman, M. J. 2015. Using multilingual topic models for improved alignment in English-Hindi MT. Proceedings of the 12th International Conference on Natural Language Processing, pp. 304–311

Koltsova, O. and Koltcov, S. 2013. Mapping the public agenda with topic modeling: the case of the Russian LiveJournal. *Policy and Internet* **5**(2): 207–227

Lamba, M. and Madhusudhan, M. 2018. Metadata tagging of library and information science theses: Shodhganga (2013-2017). In ETD 2018 Taiwan. Beyond the Boundaries of Rims and Oceans: Globalizing Knowledge with ETDs. Details available at https://etd2018.ncl.edu.tw/images/phocadownload/3-2_Manika_Lamba_Extended_Abstract_ETD_2018.pdf, last accessed on 27 March 2019

Yezheng, L., Fei, D., Jianshan, S. and Yuanchun, J. 2019. iLDA: An interactive latent Dirichlet allocation model to improve topic quality. *Journal of Information Science*. Details available at https://doi.org/10.1177/0165551518822455, last accessed on 27 March 2019

Ma, T., Li, R., Ou, G., and Yue, M. 2018. Topic based research competitiveness evaluation. *Scientometrics* **117**(2): 789–803

Mehler, A. and Waltingerm, U. 2009. Enhancing document modeling by means of open topic models: crossing the frontier of classification schemes in digital libraries by example of the DDC. *Library Hi Tech* **27**(4): 520–539

Muresan, G. and Harper, D. J. 2004. Topic modeling for mediated access to very large document collections. *Journal of the American Society for Information Science and Technology* **55**(10): 892–910

Nikolenko, S., Koltcov, S., and Koltsova, O. 2017. Topic modelling for qualitative studies. *Journal of Information Science* **43**(1): 88–102

Predictive Analytics Today. 2017. B2B reviews, buying guides and best practices. Details available at https://www.predictiveanalyticstoday.com/, last accessed on 27 March 2019

Ranganathan, S. R. 1957. *The Five Laws of Library Science*, p. 336. Bombay: Asia Publishing House

RapidMiner. 2019. Lightning fast data science platform. Details available at https://my.rapidminer.com/nexus/account/index.html#downloads, last accessed on 3 June 2019

Özmutlu, S. and Çavdur, F. 2005. Neural network applications for automatic new topic identification. *Online Information Review* **29**(1): 34–53

Shmueli, G. 2010. To explain or to predict? *Statistical Science* **25**(3): 289–310

Sugimoto, C. R., Li, D., Russell, T. G., Finlay, S. C., and 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

KDnuggets. 2017. Text mining 101: Topic modeling. Details available at https://www.kdnuggets.com/2016/07/text-mining-101-topic-modeling.html, last accessed on 27 November 2017

Wu, Q., Kuang, Y., Hong, Q., and She, Y. 2019. Frontier knowledge discovery and visualization in cancer field based on KOS and LDA. *Scientometrics* **118**(3): 979–1010

Yau, C.-K., Porter, A., Newman, N., and Suominen, A. 2014. Clustering scientific documents with topic modeling. *Scientometrics* **100**(3): 767–786

Zhang, W., Sim, Y. C., Su, J., and Tan, C. L. 2011. Entity linking with effective acronym expansion, instance selection, and topic modeling, pp. 1909–1914. In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI), Barcelona, Catalonia, Spain, 16–22 July