Cadernos **BAD**

Mapping of ETDs in ProQuest dissertations and theses (PQDT) global database (2014-2018)

Manika Lamba

Department of Library and Information Science, University of Delhi
lambamanika07@gmail.com

Margam Madhusudhan

Department of Library and Information Science, University of Delhi
madhumargam@gmail.com

Abstract

The information explosion in the form of ETDs poses the challenge of management and extraction of appropriate knowledge for decision making by information practitioners. This study presents a solution to the problem by applying topic mining and prediction modeling to 441 *full-text* ETDs extracted from the PQDT Global database during 2014-2018 in the field of library science using the RapidMiner platform. This study was divided into three phases. In the first phase, metadata analysis of the ETDs retrieved from the database was performed to identify the association of various entities such as universities, departments, types of degrees, and geographical areas with the ETDs. In the second phase, 8 core topics namely *children literature; academic library; information retrieval; archival science; user study; digital library; library leadership; and digital communication* were determined using latent dirichlet allocation (LDA) and each ETD was then annotated with the modeled topic. Lastly, a prediction model using the Support Vector Machine (SVM) classifier was created to classify the untagged ETDs going to be submitted in the database under the 8 modeled topics (*a* to *h*).

Keywords: Latent Dirichlet Allocation, Machine Learning, Text Analytics, Topic Modeling, Prediction Modeling

Introduction

Electronic Theses and Dissertations (ETDs) are the most frequent types of educational resources that are being consulted by the scientific community from time to time. «The submission of theses and dissertations in electronic format has opened the door for the user community to have an entrance to the knowledge implanted in these works through different national and international ETDs and databases» (Haneefa and Divya, 2018). As the number of textual data including ETDs are increasing exponentially every day over the Web, the issue of organizing, managing and disseminating information has attracted attention and led to many efforts, including the knowledge management, content analysis, text analysis, text classification, text categorization, search strategy, linked data, semantic web, etc. to enhance information retrieval systems and their performance for decision making. Thus, the present study (i) conducts metadata analysis of the ETDs, (ii) discovers the hidden topical pattern from the corpus of ETDs, (iii) annotates the ETDs according to the discovered topics to organize, search and summarize text, and (iv) presents a best fitted predictive model for Library and Information Science (LIS) ETDs present at PQDT Global database from 2014 to 2018. This study will help to identify the highly-researched areas during the studied period in the PQDT Global database for the identified geographical region; improve the organization and management of the PQDT Global website, and provide better search experience to the users of the database. This work will have a broad application to those who want to know the research trends in the field of LIS at an international level.

Review of Related Literature

Topic modeling is an emerging technology in the field of LIS. Though there are many studies where it has been applied in the context of journals such as by Kurata et al. (2018) where they «examined 1,648 full-text articles in LIS using LDA method from five representative journals and 30 topics were identified»; Joo and Cahill (2018) who «used text mining to explore 20 research topics from the two leading research LIS journals in school librarianship using LDA»; and Authors (2019) «analyzed 928 full-text research articles retrieved from DESIDOC Journal of Library and Information Technology for the period of 1981-2018 using LDA and identified 50 core topics». But there are meager studies where topic modeling has been applied to ETDs. The classical paper which first applied topic mining to ETDs in LIS is by Sugimoto et al. (2011) who «identified the changes in dominant topics in LIS by analyzing 3,121 doctoral dissertations completed during 1930-2009 at North American Library and

Information Science programs and utilized LDA to identify latent topics diachronically». A study similar to the present study was also conducted by authors where they «applied topic mining and prediction modeling on ETDs found in Shodhganga database during 2013-2017» (Authors, 2018).

Methodology

A total of 442 ETDs were identified for the query "*library science*" but only 441 ETDs were accessed. The scope of the study was restricted to *full-text* ETDs submitted to the PQDT Global database for the epoch of 2014-2018 in the field of LIS in the English language. An excel file was also prepared simultaneously for each year with each download which contained metadata information for the ETDs i.e. title, author, advisor, university/institute name, university/institute location, abstract, and type of degree.

The present study was divided into three phases. In the first phase, metadata analysis of the ETDs was conducted to determine the prominent universities and departments; different levels and types of degrees; and the geographical area for the studied period. Secondly, the identified ETDs were downloaded from the database one-by-one and were then converted into a text format followed by the analysis of the corpus using LDA with the help of the RapidMiner platform. The implementation of LDA in RapidMiner used the ParallelTopicModel (Newman et al., 2009) of the Mallet library with SparseLDA (Yao et al., 2009) sampling scheme and data structure. Further, LDA in RapidMiner used Gibbs Sampling for the application of the model and exposed additional parameters in the application. Finally, prediction analysis was performed using the RapidMiner platform (2019). The prediction analysis process included the following steps:

(i) Pre-processing of the documents (i.e. tokenization, stemming, filtering stop-words, transforming the cases, and generating n-grams per terms);

(ii) Splitting the data into two subsets;

(iii) Training and testing of the data using split validation;

(iv) Application of the appropriate classifier such as Naive Bayes, Support Vector Machine (SVM), etc. to build the predictive model; and

(v) Performance evaluation of the model.

PQDT Global Database

«ProQuest has partnered with academic institutions around the world to archive and disseminate a comprehensive collection of dissertations and theses. The program started in 1939 with a goal to create a U.S. national repository of graduate works. ProQuest now has partnerships with most of the doctoral institutions in the U.S. and Canada and with a significant and growing list of international universities. Each year, ProQuest adds more than 130,000 new dissertations and theses to its largest dissertation database, ProQuest Dissertations & Theses (PQDT)

Global. It offers abstracts and indexing for approximately 4 million dissertations and theses. Coverage for the database begins in 1637, and full-text coverage is primarily from 1997 forward. It enhances discovery within the author's discipline and also enables the authors' works to be showcased for use in other academic journals, working papers, reports, and studies. ProQuest continues to develop and expand partnerships in order to raise the visibility of dissertations and theses through high-quality, relevant indexes and databases» (*PQDT*).

PQDT has two publishing options: (i) traditional and (ii) open. PQDT with the traditional publishing of ETDs is not accessible through search engines and only the institutes who have subscribed to the database can access the ETDs whereas PQDT with open publishing provides the full text of open access dissertations and theses free of charge. It constitutes full-text (PDFs) for more than 2 million ETDs but the present study is confined to the ETDs submitted to PQDT Global i.e. with the traditional publishing in the field of LIS.

Latent Dirichlet Allocation (LDA)

LDA is a probabilistic model that was first introduced by Blei et al. in 2003. In this study, each ETD was represented as a pattern of LDA topics by making every ETD appeared. LDA automatically inferred the topic discussed in the given collection and these topics were then used to summarize and organize the collection. Figure I «demonstrate the functioning of LDA where the outer box represents the documents and the inner box represents the repeated choice of topics and words within a document. The variables shown in the figure are defined as follows» (Authors, 2019):

α—parameter of Dirichlet prior on the per-document topic distribution

β—parameter of Dirichlet prior on per-topic word distribution

θ—topic distribution for the document, *d*

z—topic for the nth word in the document, *d*

w—is the specific word

N—total number of words

M—total number of documents in the corpus

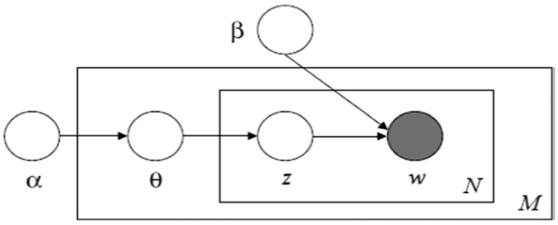


Figure I: Graphical model representation of Latent Dirichlet Allocation
(Source: Blei et al., 2003)

RapidMiner

RapidMiner platform (2019) is an easy-to-use visual environment.

«It uses a client/server model with the server offered as either on-premise or in public or private cloud infrastructures. It provides 99% of an advanced analytical solution through template-based frameworks that speed delivery and reduce errors by nearly eliminating the need to write code. It provides data mining and machine learning procedures including data loading and transformation (Extract, transform, load (ETL)), data pre-processing and visualization, predictive analytics and statistical modeling, evaluation, and deployment. It is written in the Java programming language. It provides a GUI to design and execute analytical workflows. Those workflows are called *Processes* in RapidMiner and they consist of multiple *Operators*. Each operator performs a single task within the process, and the output of each operator forms the input of the next one. Alternatively, the engine can be called from other programs or used as an API. Individual functions can be called from the command line. It provides learning schemes, models, and algorithms and can be extended using R and Python scripts. Its functionality can be extended with additional plugins that are made available via RapidMiner's Marketplace. RapidMiner's Marketplace provides a platform for developers to create data analysis algorithms and publish them to the community» (Hofmann & Klinkenberg, 2013).

Prediction Modeling

Prediction modeling is the process of creating, testing, and validating a model to predict the probability of an outcome to its best.

«Prediction analytics uses a number of modeling methods from machine learning, artificial intelligence, and statistics on the basis of testing, validation, and evaluation using 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» (Predictive Analytics Today, 2019)

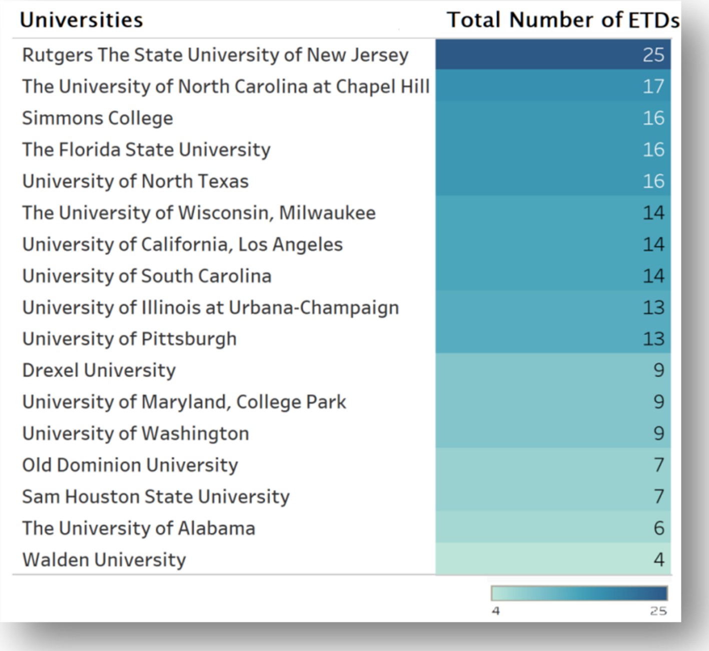
«The basic steps of the prediction modeling include: (i) Creating the model - create a model to run one or more algorithm on the data set; (ii) Testing the model - test the model on the data set, wherein some scenario testing is done on past data to see how best the model predicts; (iii) Validating the 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 choosing 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. As additional data become available in the future, the predicted statistical analysis model can be validated or revised accordingly» (Predictive Analytics Today, 2019)

Results

Metadata Analysis

In metadata analysis, the bibliographical data which was recorded for the studied period were analyzed to determine the (i) prominent universities; (ii) prominent departments; (iii) types and levels of degrees; and (iv) prominent geographical location associated with the ETDs found in the PQDT Global database for LIS discipline.

17 prominent universities were identified which submitted the highest number of ETDs during the studied period in PQDT Global (Figure II). The top five universities which submitted their ETDs were: (i) Rutgers, The State University of New Jersey (25); (ii) The University of North Carolina at Chapel Hill (17); and (iii) Simmons College, The Florida State University and University of North Texas (16 each).

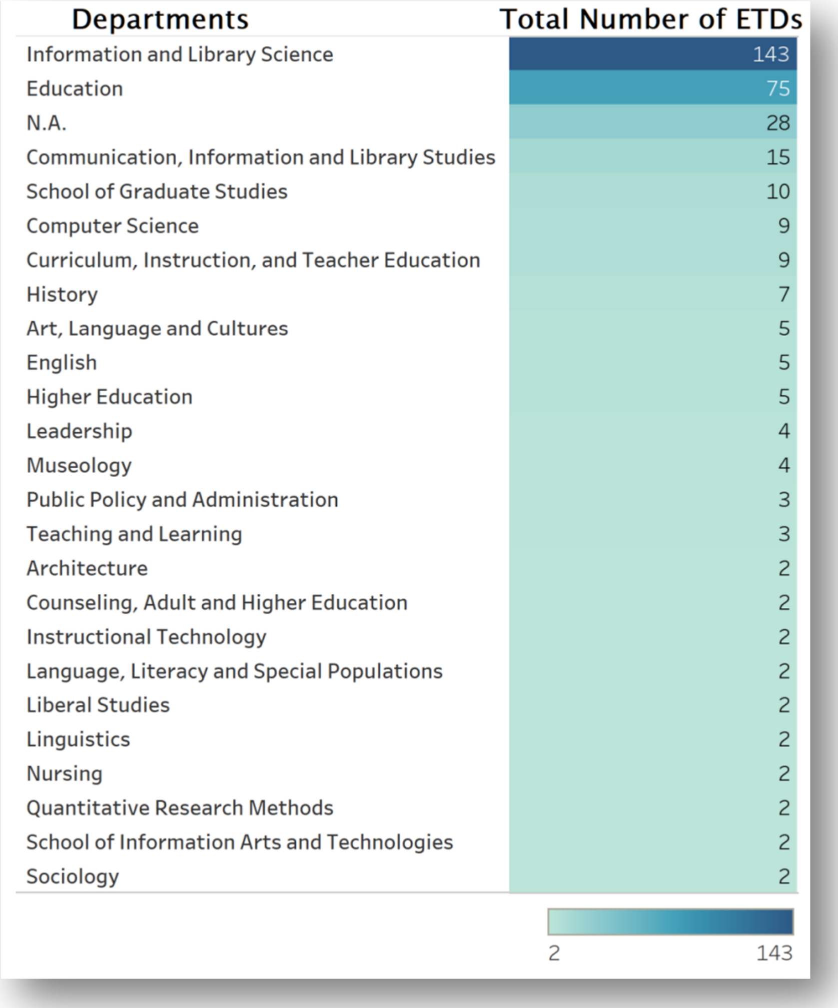


Universities	Total Number of ETDs
Rutgers The State University of New Jersey	25
The University of North Carolina at Chapel Hill	17
Simmons College	16
The Florida State University	16
University of North Texas	16
The University of Wisconsin, Milwaukee	14
University of California, Los Angeles	14
University of South Carolina	14
University of Illinois at Urbana-Champaign	13
University of Pittsburgh	13
Drexel University	9
University of Maryland, College Park	9
University of Washington	9
Old Dominion University	7
Sam Houston State University	7
The University of Alabama	6
Walden University	4

Figure II: Identification of Prominent Universities in PQDT Global (2014-2018)

25 prominent departments were identified which submitted ETDs on library science subject. Figure III indicates the strong interdisciplinary nature of the LIS domain. With no surprise, information and library science department (143) was at the top of the list followed by education department (75).

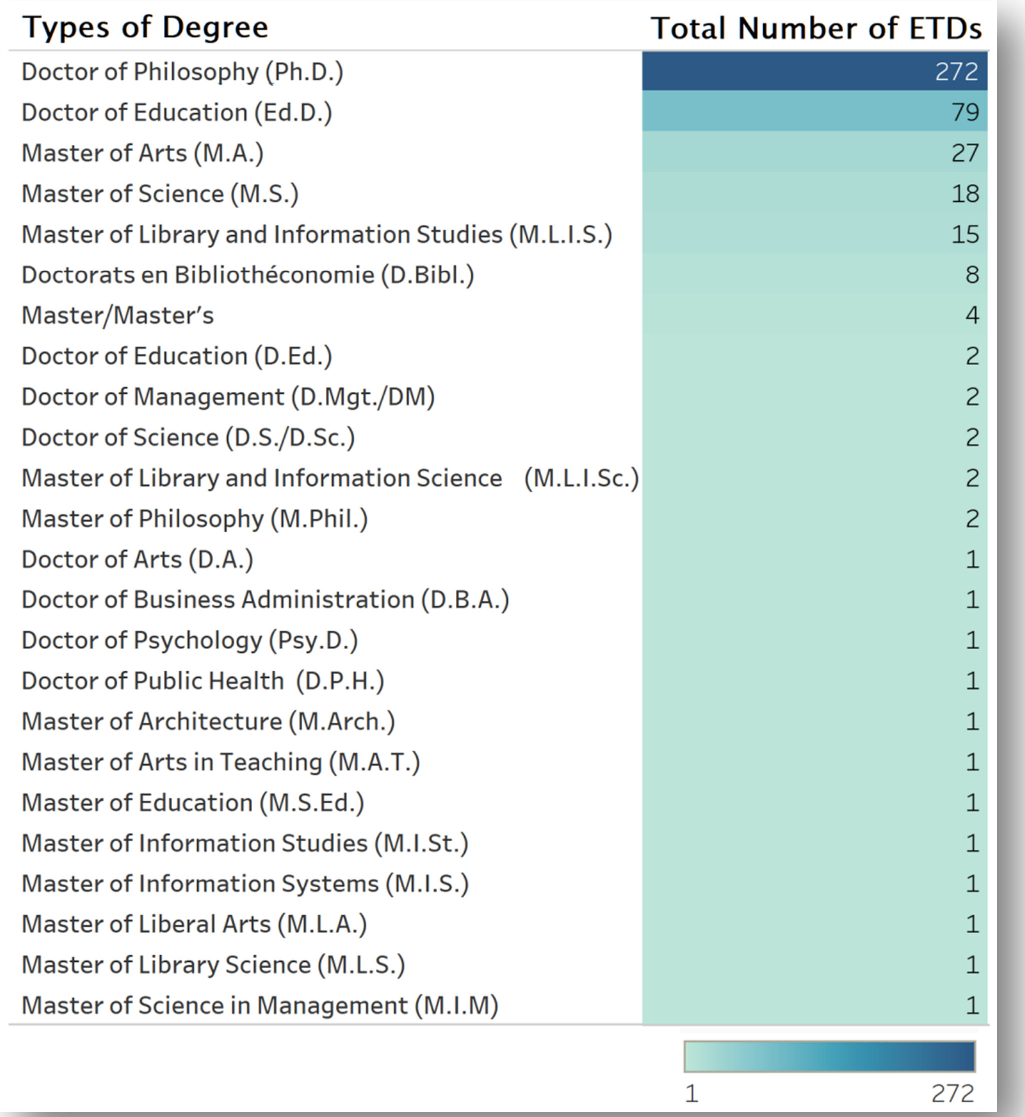
Interestingly, it was noticed that 28 ETDs had missing information for department names in the database (Figure III) whereas 17 ETDs were identified which were not having information regarding the advisor's name. This incomplete nature of metadata information of ETDs in the PQDT Global database may restrict users to search ETDs of their interest in LIS domains. Further, it was observed that most of the ETDs which were extracted from an external database to PQDT Global were having incomplete metadata information for the advisor's name, page number, and department's name.



Departments	Total Number of ETDs
Information and Library Science	143
Education	75
N.A.	28
Communication, Information and Library Studies	15
School of Graduate Studies	10
Computer Science	9
Curriculum, Instruction, and Teacher Education	9
History	7
Art, Language and Cultures	5
English	5
Higher Education	5
Leadership	4
Museology	4
Public Policy and Administration	3
Teaching and Learning	3
Architecture	2
Counseling, Adult and Higher Education	2
Instructional Technology	2
Language, Literacy and Special Populations	2
Liberal Studies	2
Linguistics	2
Nursing	2
Quantitative Research Methods	2
School of Information Arts and Technologies	2
Sociology	2

Figure III: Identification of Prominent Departments in PQDT Global (2014-2018)

24 different types of degrees where identified at masters and doctorate level in the database for the studied period. Out of 24 types of degrees, Ph.D. (273) was at the top of the list followed by Ed.D. (79), M.A. (27), M.S. (18), and M.L.I.S. (15) (Figure IV). Furthermore, for LIS at the master's level, 5 different kinds of degrees were identified in the database with different nomenclature viz. M.L.I.S.; M.L.I.Sc.; M.I.S.; M.L.S; and M.I.St.



Types of Degree	Total Number of ETDs
Doctor of Philosophy (Ph.D.)	272
Doctor of Education (Ed.D.)	79
Master of Arts (M.A.)	27
Master of Science (M.S.)	18
Master of Library and Information Studies (M.L.I.S.)	15
Doctorats en Bibliothéconomie (D.Bibl.)	8
Master/Master's	4
Doctor of Education (D.Ed.)	2
Doctor of Management (D.Mgt./DM)	2
Doctor of Science (D.S./D.Sc.)	2
Master of Library and Information Science (M.L.I.Sc.)	2
Master of Philosophy (M.Phil.)	2
Doctor of Arts (D.A.)	1
Doctor of Business Administration (D.B.A.)	1
Doctor of Psychology (Psy.D.)	1
Doctor of Public Health (D.P.H.)	1
Master of Architecture (M.Arch.)	1
Master of Arts in Teaching (M.A.T.)	1
Master of Education (M.S.Ed.)	1
Master of Information Studies (M.I.St.)	1
Master of Information Systems (M.I.S.)	1
Master of Liberal Arts (M.L.A.)	1
Master of Library Science (M.L.S.)	1
Master of Science in Management (M.I.M)	1

Figure IV: Different Types of Degree Associated with ETDs in PQDT Global (2014-2018)

7 countries were identified that submitted in the PQDT Global database for the studied period (Figure Va). These countries were United States (427); United Kingdom (8); Canada (7); India (2); Nigeria (1); Australia (1); and Belgium (1). As the majority of the ETDs were submitted from the United States, a State-level map was also constructed to determine the contribution from the various US states in the database (Figure-Vb). As it can be observed from Figure V (b), the majority of the ETDs had been submitted from Pennsylvania (37) state followed by Texas (36), California (34), New Jersey (33), North Carolina (25), Massachusetts (23), Illinois (23), and Florida (21).

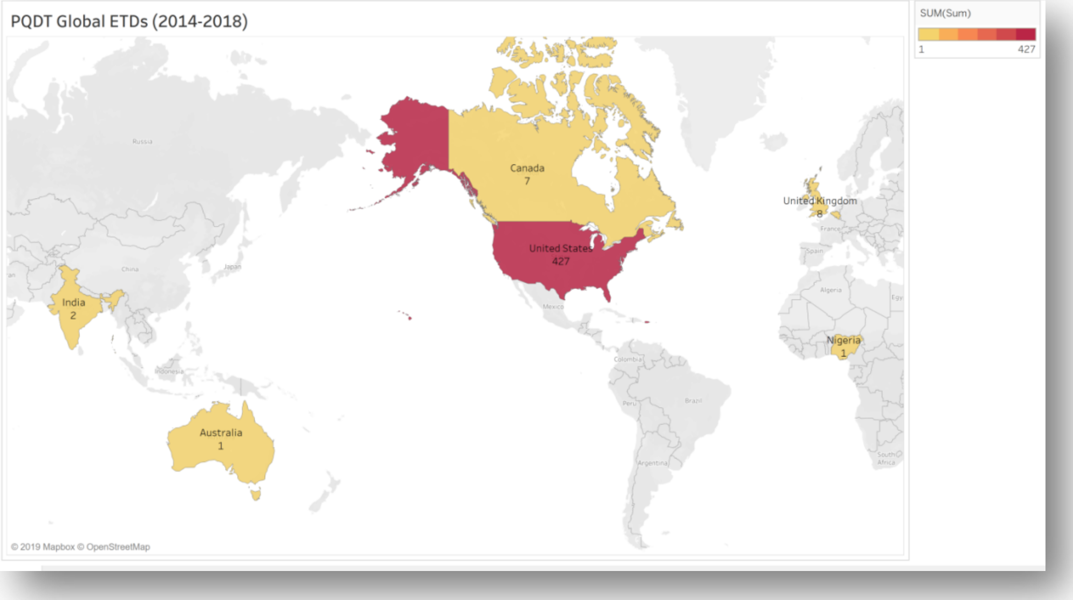
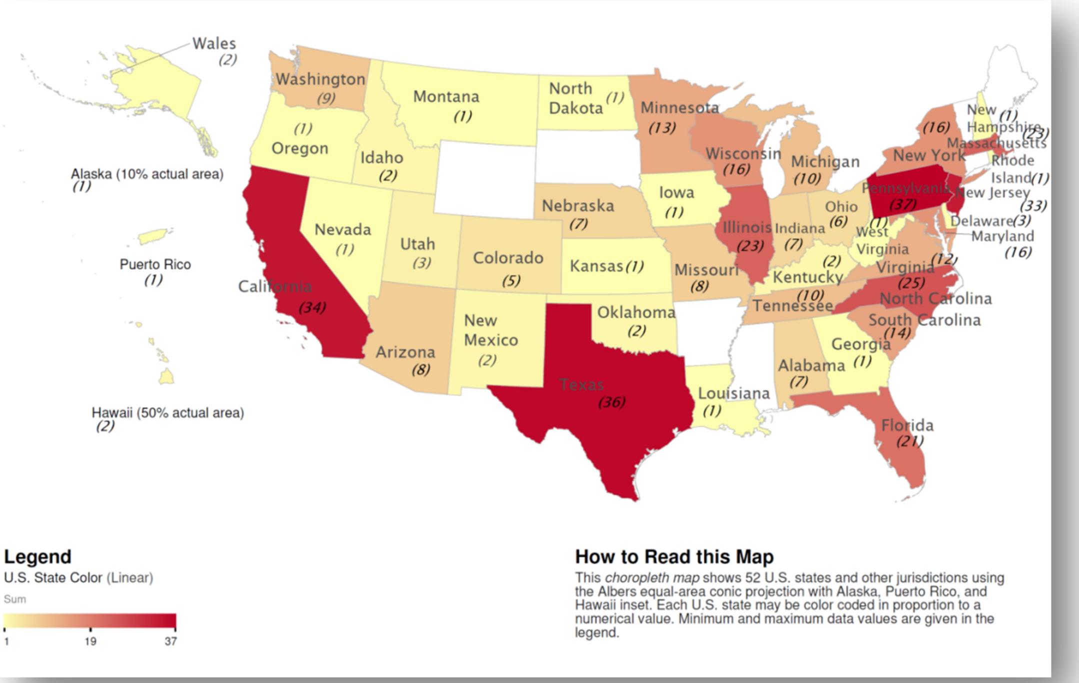


Figure V (a): Choropleth Map of PQDT Global ETDs (2014-2018) at World Level



This *choropleth map* shows 52 U.S. states and other jurisdictions using the Albers equal-area conic projection with Alaska, Puerto Rico, and Hawaii inset. Each U.S. state may be color coded in proportion to a numerical value. Minimum and maximum data values are given in the legend.

Figure V (b): Choropleth Map of PQDT Global ETDs (2014-2018) for US States

Topic Analysis

«Topic analysis is a process of assigning topics to a group of high-frequency words arranged in descending order and analyzing the results generated from an automated tool for the purpose of management and organization of the text documents» (Authors, 2019). After topic modeling was conducted to the corpus of ETDs extracted from PQDT Global using the RapidMiner platform, the topical analysis was performed to generate knowledge and to assign appropriate topics to the high probability co-occurring words. Table I summarizes the LDA results for the study which shows the labelling of the topics (*a* through *h*) in descending order according to their probability values where *Topic a* had the highest probability value. To determine the topics, both co-occurrence words and representative ETDs were consulted. These representative ETDs were the top five ETDs which were ranked on the basis of the highest topic proportion percentage for the given modelled topic (Appendix-I).

For the studied period (2014-2018), 8 topics were identified where Number of articles=441; Number of Words=5; AlphaSum=1.874; Beta= 0.06. The evidence from high-loading keywords and most representative ETDs showed that *Topic a* was about *children literature* with a focus on reading habits. *Topic b* was about *academic library* with a focus on school library, research, information, librarians, and students whereas *Topic c* was about *information retrieval* with a focus on searching, user, system and task. *Topic d* displayed a focus on *archival science* with an emphasis on collection, records, and history in contrast to *Topic e* which was focused on *user study* with a focus on information, participants, research and society. *Topic f* indicated a focus on *digital library* with a focus on data, research, journal, and access in comparison to *Topic g* which was on *library leadership* with a focus on staff, librarian, and service in libraries. Lastly, *Topic h* was on *digital communication* with a focus on archives, media, and technology.

Topic a *Children Literature*	Topic b Academic Library	Topic c *Information Retrieval*	Topic d *Archival Science*	Topic e *User Study*	Topic f *Digital Library*	Topic g *Library Leadership*	Topic h *Digital Communication*
book	student	inform	archiv	inform	data	librari	digit
read	school	search	collect	particip	research	servic	commun
librari	librarian	user	histori	research	journal	librarian	archiv
children	inform	system	book	studi	scienc	leadership	media
school	research	task	record	social	access	staff	technolog

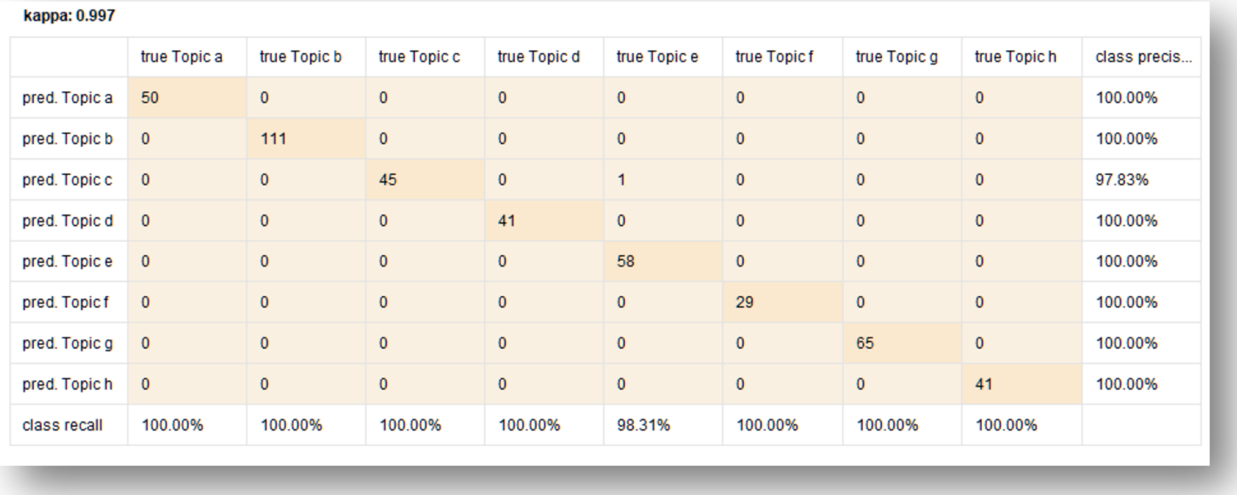
Table I: Latent Dirichlet Allocation Topic and Word Result for PQDT Global ETDs during 2014-2018 (n=441)

Further, Appendix-I summarizes the results for the topic proportion generated by the RapidMiner platform where each ETD retrieved from the database for the studied period was tagged for the modelled topic. This percentage composition segregates the ETDs on the basis of their similarity and thus, would help a user to search the database on the basis of core topics associated with the ETDs. Furthermore, the creators of the PQDT Global database can provide recommendation service to their users on the basis of their search or reading habits by embedding the topic proportion for the tagged ETDs.

Moreover, as it can be observed from Figures V (a and b) that most of the ETDs had been submitted from the United States, therefore, the modelled topics resulted from topic modeling were predominantly associated with LIS research conducted in the United States in comparison to the contribution from other countries which were very less in number. Thus, the resulted modelled topics could be considered as the highly-researched areas for LIS in the United States for the studied period.

Prediction Analysis

A prediction model using the Support Vector Machine (SVM) classifier was created and tested (Figure VI). The model was created using 441 ETDs under the 8 modelled topics, where 70% (308) of the ETDs were allocated to the training set and 30% (133) was allocated to test set randomly using the split validation technique. Once the parameter of the model was finalized, the testing set was run through the model. The true class was compared to the predicted class to determine the kappa, precision, and recall values. Figure VI shows the result for the evaluation of the prediction model with 0.997 kappa value.



kappa: 0.997

	true Topic a	true Topic b	true Topic c	true Topic d	true Topic e	true Topic f	true Topic g	true Topic h	class precis...
pred. Topic a	50	0	0	0	0	0	0	0	100.00%
pred. Topic b	0	111	0	0	0	0	0	0	100.00%
pred. Topic c	0	0	45	0	1	0	0	0	97.83%
pred. Topic d	0	0	0	41	0	0	0	0	100.00%
pred. Topic e	0	0	0	0	58	0	0	0	100.00%
pred. Topic f	0	0	0	0	0	29	0	0	100.00%
pred. Topic g	0	0	0	0	0	0	65	0	100.00%
pred. Topic h	0	0	0	0	0	0	0	41	100.00%
class recall	100.00%	100.00%	100.00%	100.00%	98.31%	100.00%	100.00%	100.00%	

Figure VI: Confusion Matrix for Support Vector Machine (SVM) based Prediction Model

Discussion

This study first analyzed the metadata of the LIS ETDs submitted to the PQDT Global database for the period 2014-2018 to understand the association of the various entities such as universities, departments, types of degree, and geographical area with respect to ETDs followed by tagging of the corpus using topic modeling. 8 core topics (tags) namely *children literature; academic library; information retrieval; archival science; user study; digital library; library leadership; and digital communication* were identified which fitted the corpus meticulously. Furthermore, each ETD was broken down into various topic proportions of percentage probabilities to segregate on the basis of the modelled topics. These findings can be mentioned on the PQDT Global website which can ultimately help the user in faster information retrieval and could also be used to provide recommendation service on the basis of their search or reading habits. The major limitations of topic modeling in the study include identification of an appropriate number of topics for the data in advance before performing the LDA; the inherent incompetence of Dirichlet to correlate among topics; and lastly, the manual interpretation and labelling of *topics*.

The present study further applied prediction modeling to the ETDs and tried to accurately predict the classification of the future ETDs going to be submitted to the PQDT Global database under the 8 modelled topics (*a* to *h*) on the basis of performance of the predictive model. Hence, the results from the prediction analysis based on the Support Vector Machine (SVM) classifier could be used to provide a successful future automated classification of LIS ETDs. The limitation of prediction modeling for the study was that the dataset was not truly representative of LIS ETDs of the PQDT Global database. The training of the model to learn and fit the parameters could be done perfectly if more data is taken into account. Thus, this study can be extended to perform probabilistic topic modeling of all the ETDs submitted to PQDT Global so far, so that the resulting output of the prediction model can be relied upon with total confidence.

Conclusion

Topic modeling is a text-mining tool that helps to process, organize, manage and extract knowledge from a corpus of text documents whereas prediction modeling is a machine learning approach that helps to predict the future classification of text documents either in a supervised or unsupervised process. 8 core topics namely *children literature; academic library; information retrieval; archival science; user study; digital library; library leadership; and digital communication* were determined using latent Dirichlet allocation (LDA) followed by tagging of each ETD with the modelled topic and a prediction model using Support Vector Machine (SVM) classifier was created to classify the untagged ETDs going to be submitted in the database under the 8 modelled topics (*a* to *h*).

References

BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I. - Latent dirichlet allocation. **Journal of machine Learning research.** 3:1 (2003) 993-1022.

HANEEFA K., Mohamed; DIVYA, P. (2018) - Electronic Theses and Dissertations (ETDs) in India. In **ICT Application in Academic Library Management.** New Delhi: Ess Ess Publications, [s.d.] [Consult. 28 Sep. 2019]. Available at Internet:<URL:https://www.researchgate.net/publication/322599283_Electronic_Theses_and_Dissertations_ETDs_in_India>.

HOFMANN, Markus; KLINKENBERG, Ralf (EDS.) - **RapidMiner: Data Mining Use Cases and Business Analytics Applications.** 1 edition ed. Boca Raton : Chapman and Hall/CRC, 2013. ISBN 978-1-4822-0549-7.

JOO, Soohyung; CHOI, Inkyung; CHOI, And Namjoo - Topic Analysis of the Research Domain in Knowledge Organization: A Latent Dirichlet Allocation Approach. **KNOWLEDGE ORGANIZATION.** . ISSN 0943-7444. 45:2 (2018) 170-183. doi: 10.5771/0943-7444-2018-2-170.

KURATA, Keiko *et al.* - Analyzing library and information science full-text articles using a topic modeling approach: Analyzing Library and Information Science Full-Text Articles Using a Topic Modeling Approach. **Proceedings of the Association for Information Science and Technology.** . ISSN 23739231. 55:1 (2018) 847-848. doi: 10.1002/pra2.2018.14505501143.

LAMBA, Manika; MADHUSUDHAN, Margam - Mapping of topics in DESIDOC Journal of Library and Information Technology, India: a study. Scientometrics. 120:2 (2019) 477-505.

LAMBA, Manika; MADHUSUDHAN, Margam - Metadata Tagging of Library and Information Science Theses: Shodhganga (2013-2017). In ETD2018 Taiwan Beyond the Boundaries of Rims and Oceans: Globalizing Knowledge with ETDs. Taipei, Taiwan, 2018. Available on the Internet: <URL: https://etd2018.ncl.edu.tw/images/phocadownload/3-2_Manika_Lamba_Extended_Abstract_ETD_2018.pdf>.

NEWMAN, David *et al.* - Distributed Algorithms for Topic Models. **Journal of Machine Learning Research.** 10 (2009) 1801-1828.

PREDICTIVE ANALYTICS TODAY [Consult. Sep. 28, 2019] Available on the Internet: <URL: https://www.predictiveanalyticstoday.com/>.

PQDT [Consult. Sep. 28, 2019] Available on the Internet: <URL: <https://www.proquest.com/products-services/dissertations/ProQuest-Dissertations-FAQ.html>.

RAPIDMINER [Consult. Sep. 28, 2019] Available on the Internet: <URL:<https://rapidminer.com/>.

SUGIMOTO, Cassidy R. *et al.* – 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.** . ISSN 15322882. 62:1 (2011) 185-204. doi: 10.1002/asi.21435.

YAO, Limin; MIMNO, David; MCCALLUM, Andrew – Efficient methods for topic model inference on streaming document collections. Em **Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09** [Em linha]. Paris, France: ACM Press, 2009 [Consult. 3 set. 2019]. Available on the Internet:<URL:http://portal.acm.org/citation.cfm?doid=1557019.1557121>. ISBN 978-1-60558-495-9