International Journal of Information Science and Management
Vol. 24, No. 1, 2026, 29-64
DOI: https://doi.org/10.22034/ijism.2026.2022500.1393 / DOR: https://dor.isc.ac/dor/20.1001.1.20088302.2026.24.1.2.3

Original Research

In Search of Theories in Library and Information Science: A Topic Modeling Approach

Bijan Kumar Roy
Professor, Department of Library and Information Science, University of Calcutta, West Bengal, India
Corresponding Author: bijanlis@caluniv.ac.in / bkroylis@gmail.com
ORCID iD: https://orcid.org/0000-0001-9735-9586

Parthasarathi Mukhopadhyay
Professor, Department of Library and Information Science, University of Kalyani, West Bengal, India
psm@klyuniv.ac.in / psmukhopadhyay@gmail.com
ORCID iD: https://orcid.org/0000-0003-0717-9413

Received: 12 February 2024
Reviewed: 12 May 2024
Accepted: 17 March 2025

Abstract

Finding relevant topics or extracting useful information from large corpora of data has been challenging for academics, and topic modeling, a machine learning algorithm, has emerged as an alternative technique for discovering the underlying semantic structure of large, unstructured collections of documents. Our objectives are to identify the topics covered in the corpus data, group them by topic, show the development of research across different aspects of LIS, and demonstrate the application and use of theories from other domains in the LIS domain. We use several open-source tools for topic modeling, such as LDA (Latent Dirichlet Allocation), Gensim, Jupyter Notebook, ASReview, and OpenRefine, to extract key topics from titles and abstracts. The results of this study are summarized into three main sets: identification of specific topics, word clouds, trends in subjects, and the use and applications of theories in this domain. The model may help policymakers, funding agencies, and the government understand the current and future state of research and take corrective actions to address gaps in the literature on expert systems and applications. It also helps library professionals, classificationists, and researchers identify relevant topics in unstructured long texts and reduce information overload by removing unnecessary research documents.

Keywords: Text Mining, Machine Learning, Topic Discovery, Natural Language Processing, NLP, Library and Information Science, Theory, LDA.

Introduction

Due to advances in information and communication technology (ICT) and the open knowledge movement (OKM), large amounts of academic literature/data, both textual and non-textual, are being generated and made available on various open platforms. It has made information retrieval difficult, and LIS (Library and Information Science) professionals are facing challenges in analyzing, organizing, and classifying unstructured text data, as well as in accessing the relevant topics they seek. Data is often unstructured, and manually searching for and uncovering hidden insights in such data is highly tedious and time-consuming, and an excellent challenge for librarians/LIS professionals. This retrieval problem forces librarians to find new ways to organize, search, and understand patterns in vast amounts of information.

Generating an exact solution to extract helpful information/topic from an enormous volume of literature is often computationally intractable and poses a significant challenge for librarians. Topic modeling, an unsupervised text mining approach, is a popular analytical tool for automatically classifying documents and for evaluating and revealing the underlying semantic structure of extensive document collections. It has emerged as a powerful tool for extracting meaningful information and topics from many unstructured texts and for identifying semantically related documents based on the issues they address (Hong & Davison, 2010).

Benefits of the topic model

Topic modeling has been an active area of research. It has been successfully used over the last 40 years to facilitate information retrieval, classify documents, and support exploratory analysis of large corpora of texts. However, it began in the 1980s as a means for more accurate information retrieval programs to predict hidden topics from a text corpus. Initially, it was developed as an alternative to keyword search to enhance the exploration of text data collection (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). Now, it has become a valuable tool for information retrieval, aiding query expansion and document smoothing (Yi & Allan, 2008, 2009). It can extract useful information from unstructured texts and find hidden topics within a collection of text documents. They are probabilistic models for discovering the hidden structure of a corpus of documents based on a Bayesian analysis of the papers (Rosen-Zvi, Griffiths, Steyvers & Smyth, 2004). Topic modeling is a text mining and concept extraction method that extracts topics (i.e., coherent word clusters) from large corpora of textual documents to discover hidden semantic structures in the text (Miner, Elder, Fast, Hill, Nisbet & Delen, 2012). The advantages of topic modeling over other techniques are that it helps analyze long texts (Treude & Wagner, 2019; Miner et al., 2012), creates clusters as 'topics' (rather than individual words), and is unsupervised (Miner et al., 2012). Yau, Porter, Newman, and Suominen (2014) rightly argued that a topic model could successfully group a broad collection of scientific articles by discipline. Suominen and Toivanen (2016) opined that topic models help classify significant texts.

Topic modeling has emerged as a research direction that enhances information retrieval by considering semantic relationships between words, enabling similar content-related papers to be automatically categorized into subjects. This approach typically uncovers latent text topics by modeling word associations, enabling matching of queries and documents based on their topics (Manning, Raghavan & Schütze, 2009). In natural language processing (NLP) tasks, these topic modeling methods have become increasingly popular due to their ability to represent each dimension as a topic rather than a term or set of words, and to investigate long-term trends in research topics. It is one of the most popular NLP techniques for structuring unstructured data, regardless of domain or discipline. Researchers commonly use it as a text-mining tool to discover hidden semantic structures in textual contexts. It is considered a well-known topic mining/modeling procedure for extracting topics from a specified corpus, regardless of size or type. A topic model is a statistical model used in ML and NLP to identify and extract hidden topical patterns within a corpus of texts (Guo, Barnes, & Jia, 2017). Topic Modeling, or topic mining, has been a solution for handling unstructured text and extracting features from large corpora. It is crucial because it aids decision-making by extracting key insights from both short and large volumes of documents, providing an automatic means to organize, understand, and summarize extensive collections of textual information. The topic model can also be used as an

inductive tool to identify categories that have been largely undiscovered before (Nelson, 2020). Finally, we must say that topic modeling helps communication scholars extract specific meanings from the text data (DiMaggio, Nag & Blei, 2013; Grimmer & Stewart, 2013).

In the paper, the authors neither aim to compare different topic modeling techniques nor to provide a comprehensive overview of theories used in LIS research. This paper takes a holistic approach to present bigrams, unigrams, a word cloud, etc., including the core topics that have developed over time and the theories used in LIS research derived from the corpus data. Topic modeling has passed many milestones to reach the level it is at now. Table 1 gives a snapshot of significant events in the timeline using topic modeling results.

Table 1
Important Events in the History of Topic Modeling

Year	Events
1980	Latent Semantic Analysis (LSA) was developed in the late 1980s, which uses matrix factorization to identify latent topics.
1990	Application of Singular Value Decomposition (SVD) to the problem of automatic indexing and information retrieval by Deerwester et al (1990). They call their method Latent Semantic Indexing (LSI) or Latent Semantic Analysis (LSA).
1998	Probabilistic Latent Semantic Analysis (usage of a probabilistic model)
1999	Statistical analysis of LSA, Probabilistic Latent Semantic Analysis (PLSA) by Thomas Hofmann.
2003	Latent Dirichlet Allocation (LDA) at the NIPS 2001 conference by David Blei, Andrew Ng, and Michael I. Jordan
2005	Introduction of 'Correlated Topic Model' (CTM) by Blei and Lafferty
2007	Model 'Author-Recipient-Topic' (ART), based on LDA by the University of Massachusetts. Correlated topic model (CTM) proposed by Blei and Lafferty
2010	Introduction of Pointwise Mutual Information (PMI) by Newman et al.
2012	Introduction of Non-negative Matrix Factorization (NMF) by Arora et al. Chuang et al. presented 'Termite.' Introduction of 'word cloud' by Chaney and Blei
2013	Integrates document clustering and text modeling into a single unified framework (Multi-Grain Clustering Topic Model (MGCTM) by Xie and Xing
2014	GSDMM is introduced in the topic model
2015	BPTMs (Bayesian Probabilistic Topic Models) became the dominant topic model until 2015
2016	Word2vec, lda2vec proposed by Moody; GPUDMM sampling scheme
2018	Introduction of hierarchical Stochastic Block Model (hSBM) by Gerlach et al.
2019	Introduction of the Embedded topic model
2020	Introduction of BERT in the topic model by Thompson and Mimno

Literature review

The rise of big data due to the rapid growth of Internet technology in the twenty-first century has prompted a demand for advanced text analytic techniques (such as artificial intelligence (AI), machine learning, NLP, and TM) to uncover patterns and relations embedded in the data, reduce the dimensionality of data, and forecast future outcomes more effectively and efficiently (Elragal & Klischewski, 2017). To effectively extract features from a large

corpus of text data, numerous text mining approaches in the form of topic modeling have already been introduced by researchers in different areas of study (Li, Zhang, Hu, & Hu, 2019), as it serves as the most frequently adopted technique for the purpose (Hong & Davison, 2010). In particular, the use of topic modeling in the social sciences, employing different methods/techniques (e.g., conventional models such as latent Dirichlet allocation (LDA), a generative probabilistic model, and non-negative matrix factorization (NMF)), has soared in popularity across various domains in recent years. Some topic modeling or text mining approaches have already been introduced by scholars in different domains (Li et al., 2019), among which topic modeling using LDA, LSA (Latent Semantic Analysis), probabilistic LSA, etc., served as the most frequently adopted technique for large-scale data analysis (Hong & Davison, 2010).

Topic modeling or topic mining (both probabilistic and non-probabilistic models) using different models or techniques (such as Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization, Top2Vec, BERTopic, etc.) has already been applied in diverse fields, domains covering different subjects like sociology, digital humanities, political science, literary studies, and academic discourse to understand different types of texts (Murakami, Thompson, Hunston & Vajn, 2017). It can also be categorized into different types, such as *algebraic*, *fuzzy*, *probabilistic*, and *neural* models, each with its own characteristics and limitations. Since its emergence in the 1990s, several methods, models, and tools (e.g., algorithms) have been developed in this domain for topic modeling, and one of the most common is LDA. Topic modeling using LDA can be a useful tool for the statistical analysis of document collections and other discrete data, regardless of their form or format, for their accuracy. The LDA model, introduced by Blei and Jordan (2003) and Blei, Ng and Jordan (2003), uses a probability distribution model to generate topics from the hidden text (Blei et al., 2003; Koltcov & Ignatenko, 2020). However, Griffiths and Steyvers (2004) first reported a study on topic modeling of scientific papers using LDA (also known as the generative aspect model). However, Blei & Lafferty (2006) were the first to apply this topic modeling approach in their study based on 30,000 articles from the journal '*Science*'. LDA offers several advantages for text classification, including the ability to handle large, heterogeneous document collections without relying on predefined vocabulary or structure. This LDA model can describe the arrangement of words that are repeated together, occur frequently, and resemble one another.

It can also discover latent, meaningful topics that capture the main themes and variations across documents, as well as the relationships between words and topics. Additionally, it can provide interpretable and explainable results, showing the topic proportions for each document and the word probabilities for each topic. Furthermore, LDA can be extended and modified to incorporate various features and constraints, such as metadata, supervision, temporal dynamics, and more. For these reasons, topic models using LDA have become popular, and several scholars have already applied LDA for text classification in many domains. Since the first paper on topic modeling by Jordan Boyd-Graber and David M. Blei in 2007, it has spread its wings across multiple domains and disciplines. For example, to analyze and classify genomic sequences (Wu Ding, Wang & Xu, 2010); to classify images based on visual words (Rasiwasia & Vasconcelos, 2013); on transportation research (Das, Dixon, Sun, Dutta & Zupancich, 2017); library science journals (Lamba & Madhusdhuan, 2018, 2019; Lu & Wolfram, 2012; Kurata, Miyata, Ishita, Yamamoto, Yang & Iwase, 2018); library science dissertations (Sugimoto, Li, Russell, Finlay & Ding, 2011); library science/library & information science (Yan, 2014;

Miyata, Ishita & Yang, 2020; Yan, 2015; Figuerola, Marco & Pinto, 2017; Saha & Ghosh, 2023) and its sub-domains such as information retrieval (Chen, Tsutsui, Ding & Ma, 2017), knowledge organisation (Joo, Choi & Choi, 2018), and electronic health records (Chen, Wei, Guo, Tang & Sun, 2017); banking (Hristova, 2021; Pronoza, Pronoza & Yagunova, 2018); management research (Hannigan et al., 2019); drug safety (Zou, 2018); biomedical literature (Kavvadias, Drosatos & Kaldoudi, 2020); medical sciences (Zhang et al., 2017; Jiang et al., 2012; Paul & Dredze, 2011; Wu et al., 2012); image classification (Wang & Mori, 2011; Cristani et al., 2008); emotion classification (Roberts et al., 2012; Rao, 2016; Rao et al., 2014); medical/biomedical (Liu et al., 2010; Huang, Lu & Duan, 2013; Xiao et al., 2017); healthcare (Wang, Huang & Gan, 2016); technology (Bulygin, Musabirov, Suvorova, Konstantinova and Okopnyi (2018); education (Afacan-Adanir, 2019; Willis et al., 2017); to identify research trends of IoT (Inaam ul Haq, Li & Hou, 2021); big data trends in the marketing field (Amado, Cortez, Rita & Moro, 2018); clinical psychology (Liu, Zhang & Kishimoto, 2021); bio-informatics (Liu et al, 2016); computer vision (Feng & Lapata, 2010); source code analysis (Linstead et al, 2007; Gethers & Poshyvanyk, 2010; Tian, Revelle & Poshyvanyk, 2009; Lukins, Kraft & Etzkorn, 2010; Linstead, Lopes & Baldi, 2008; Chen et al., 2012; Savage et al., 2010); opinion and aspect mining (Chen et al., 2010; Zheng et al., 2014; Cheng et al., 2014; Zhai et al., 2011; Bagheri, Saraee & De Jong., 2014; Wang et al., 2014; Xianghua et al., 2013; Jo & Oh, 2011; Paul & Girju, 2010; Titov & McDonald, 2008); financial markets (Nassirtoussi et al., 2014); event detection (Qian et al., 2016; Hu et al., 2012; Weng & Lee, 2011; Lin et al., 2010); political events or political science (Lozano, Schreiber & Brynielsson, 2017; Chen et al., 2010; Cohen & Ruths, 2013; Greene & Cross, 2015; Preotiuc-Pietro et al., 2017); improvement of recommendation algorithms (Rosa et al, 2018; Xiao et al., 2019); recommendation system (Zoghbi, Vulic & Moens, 2016; Cheng & Shen, 2016; Zhao et al., 2016; Lu & Lee, 2015; Wang et al, 2014; Yang & Rim, 2014; Kim & Shim, 2014); retail (Ibrahim & Wang, 2019); location-based sociological analysis (El-Diraby, Shalaby & Hosseini, 2019.); social trends and viral topics (Li, Wu & Feng, 2018.); social network (McCallum, Corrada-Emmanuel & Wang, 2005; Wang, Burke & Kraut, 2013; Henderson & Eliassi-Rad, 2009; Yu, He & Liu, 2015; Liu et al., 2016); software engineering (Sun et al., 2016; Chen, Thomas & Hassan, 2016; Linstead et al., 2007; Gethers & Poshyvanyk, 2010; Asuncion et al., 2010; Thomas, 2011; Thomas et al., 2011; Linstead et al., 2007; Gethers & Poshyvanyk, 2010; Linstead, Lopes & Baldi, 2008; Chen et al., 2012; Dam & Ghose, 2016); geography (Cristani et al., 2008; Eisenstein et al., 2010; Tang et al., 2013; Yin et al., 2011; Sizov, 2010); crime science (Chen et al., 2015; Gerber, 2014; Wang, Gerber & Brown, 2012); linguistic science (Bauer, Noulas, Séaghdha, Clark & Mascolo, 2012; McFarland et al., 2013; Eidelman, Boyd-Graber & Resnik. 2012; Wilson & Chew, 2010; Vulic, De Smet & Moens, 2011); computer linguistics (Hall et al., 2008); statistics (De Battisti, Ferrara & Salini, 2015); international speech communication (Liu et al., 2015); corporate governance (Kushkowski et al., 2020).

In most cases, LDA, filtered LDA, extended LDA, or similar techniques were used to identify relevant topics from unsupervised data and several trends in the respective fields, including the rise and fall of a subject or an idea. Most researchers have supported this view and opined that LDA, a three-level hierarchical Bayesian model, is the most common probabilistic technique used in topic modeling work (Schofield et al., 2017; Blei et al., 2003; Chen et al., 2017; Ding, 2011; Suominen & Toivanen, 2016; Kherwa & Bansal, 2019). Despite its popularity (e.g., LDA) as an algorithm for topic modeling (TM) in the social sciences, its

efficacy in analyzing corpus data and presenting results has been widely criticized, raising several questions among researchers. It has many limitations as identified in the literature, and is based on multiple assumptions, and the results depend on several conditions. Results may vary and depend on the size and types of corpora, the techniques and tools used, how data is processed for the purpose, etc. For example, the topic is defined by the words that frequently appear together; not all topics in the corpus appear equally often, and that they are unrelated; the number of topics is predefined; word order in a corpus is not essential; the words of each document arise from a mixture of topics, each of which is a distribution over the vocabulary; inability to model topic correlation.

However, the common limitations of LDA have been highlighted in the literature (Zhang, Xu & Qiao, 2014; Guo & Diab, 2011; Vayansky & Kumar, 2020; Dohaiha et al., 2018). Such standard probabilistic topic modeling techniques are only effective in some cases for deriving relevant topics from a large dataset. It cannot distinguish between old and new topics, and sometimes derived or inferred topics lack conceptual coherence. However, Agrawal, Fu, and Menzies (2018) warned about systematic errors in the analysis of LDA topic models that limit the validity of topics. Lin et al. (2014) also advised that classical topic models usually generate sub-optimal topics when applied "as is," in too small amounts, or in short text documents. Zhang et al. (2016) criticised the LDA model for failing to provide an optimal estimate of the number of topics to select. They opined that the LDA model's efficiency and effectiveness in topic classification are closely related to the selection of topics. For example, if the number of topics is too small (short-form text), then the meaning of each topic will be broad. Conversely, if the number of topics is large, it will give us useless topics, meaning topics with too much similarity (Hajjem & Latiri, 2017). Agrawal et al. (2018) criticized the LDA model for systematic errors in data analysis. Lin et al. (2014) also criticized classical topic models for usually generating sub-optimal topics. Tang et al. (2014) reported that incorrectly selecting the number of topics can result in poor performance.

To address the limitations of LDA, many researchers have advocated for newly developed non-probabilistic topic models such as latent semantic analysis (LSA) (Deerwester et al., 1990; Landauer et al., 1998; Lochbaum & Streeter, 1989) or the non-negative matrix factorization (NMF) (Lee & Seung, 1999; Wei et al., 2003). Unlike LDA, LSA focuses on dimensionality reduction and capturing semantic relationships between words, while LDA focuses on topic modeling and understanding document generation. LSA can also improve information retrieval, content understanding, and conceptual search, and allows users to perform searches based on concepts or topics rather than solely on specific keywords. Some other popular techniques, such as Structural Topic Modelling (STA) (Lindstedt, 2019); Correlation Explanation (CorEX) (Gallagher et al., 2017); Top2Vec (Angelov, 2020); probabilistic LSA (Albalawi, Yeap & Benyoucef, 2020; Hofmann, 1999); Relational Topic Modelling (Chang & Blei, 2009) were also suggested by experts to avoid the limitations of LDA. Yu and Qiu (2018) proposed a hybrid model that extends the user-LDA topic model.

Furthermore, the authors reported numerous other developments in the field of topic discovery. Chen et al. (2017) proposed a hierarchical approach to topic detection in which words are treated as binary variables and can appear in only one branch of the hierarchy. Similarly, a Gaussian Mixture Model (Jiang et al., 2018), the Gaussian LDA model based on the word-embedding technique (Das, Zaheer & Dyer, 2015), or the Biterm Topic Model (Pan

et al., 2014) could be used for topic modeling of sparse matrices, such as short texts. Blei et al. (2003) proposed a perplexity model to determine the quality of topic models with different numbers of topics. Finally, Dang, Gao, and Zhou (2016) proposed a dynamic Bayesian network approach to detect emerging topics in micro-blogging communities. Venkatesaramani et al. (2019) suggested clustering documents based on their similarity and then identifying topics for each cluster. Chang et al. (2009) discussed the challenges of evaluating the human interpretability of topic models and proposed a quantitative method for measuring the semantic meaning of inferred topics. Deng, Smith, and Quintin (2020) have proposed and developed an innovative semi-supervised learning approach by using deep learning and topic modeling to better understand the customer's voice from textual reviews. Newman et al. (2010) provided an analysis of how topics can be flawed and, more importantly, an automated evaluation metric for identifying such topics without human annotators or reference collections outside the training data. Wu et al. (2010) compared three topic modeling methods using China's web-based news portal. Hall, Jurafsky, and Manning (2008) discussed the trend visualization for LDA topic models.

Many authors have also compared topic modeling techniques across domains and tools to assess the effectiveness of these applications. Bianchi, Terragni, and Hovy (2020) proposed a comparative study of topic modeling methods applied to multiple English-language datasets using two evaluation metrics: Topic Coherence and Topic Diversity. Two Indian authors, Garbhapu and Bodapati (2020), also compared two popular topic modeling methods, viz. Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) were used on ‘Bible’ data, and it was concluded that LDA achieved superior performance when compared to LSA. In another study, Mifrah and Benlahmar (2020) compared LDA and NMF and concluded that LDA is more relevant than NMF for large corpora. Kherwa and Bansal (2020) compared three topic modeling techniques: latent semantic analysis, latent Dirichlet allocation, and correlated topic modeling. Another researcher (Yi et al., 2009) conducted a comparative study of various topic modeling methods, including LDA, PLSA, and LSI, and reported that LDA generally performed better than other models. In the same vein, Lu et al. (2011) compared two models, viz. LDA and PLSA were evaluated in an empirical study, and the results also favoured LDA for most tasks. Bozdogan and Kara (2024) compared traditional and modern topic modeling algorithms for topic identification in official documents.

Research Questions

Summarising the above, the authors have set the following research questions to fulfil the objectives of this study.

1. What are the essential terms or topics in the corpus data? Or which topics are most discussed? Or
   what were the research topics studied in the LIS area during the period?
2. How did specific terms or topics develop or change during the period? Or what is the evolution of different topics throughout the study?
3. What are the theories applied in LIS? What are the most commonly used theories in this domain?

To answer these research questions, authors have adopted a novel dynamic topic discovery product that offers multiple modules, including data crawling, data cleaning, topic discovery, most important keywords, word clouds, bigrams, etc., and results visualisation.

The paper is structured as follows: Section 2 describes related work, highlighting specific applications. Section 3 illustrates the experimental methodology used to fulfill the research questions. It involves several steps, such as data collection, extraction, and filtering. Section 4 shows that the results derived from our framework are discussed and presented from different perspectives. Section 5 provides an overall assessment of the model. Section 6 discusses the key topic (decade-wise). Section 7 covers the identification of key issues developed over time and the application of theories in LIS research. Moreover, Section 8 presents the conclusion of this research, with a detailed discussion of the challenges of topic modeling and avenues for future work, which will provide researchers with insight into good research in this domain.

Materials and Methods

To conduct a topic model in any domain or discipline, researchers need to preprocess text data and transform it into a matrix (e.g., a document-term matrix) for the model's input. Before topic modeling, the corpus data requires preparation, which usually includes *tokenization,* meaning splitting the text into tokens, usually words, numbers, and punctuation; *lemmatisation*, meaning assigning the base form to each token; and *part-of-speech tagging*, meaning assigning the part of speech (e.g., verb) to each token. These crucial text-cleaning techniques remove noise from input data, thereby improving its quality. The procedure enables us to perform topic modeling at the single-word level.

Several steps and sub-steps were performed sequentially in this topic modeling workflow. Data was extracted only from LIS journals against three search terms: *library, librarian, and theory*. It was analysed using various open-source tools at different stages of this topic modeling experiment. The whole topic modeling process involves five crucial steps: selecting the corpus; pre-processing the corpus; preparing the corpus for *input into the software; building the topic model; and analyzing the topic model against different parameters*. After collecting data from the Scopus database (https://www.elsevier.com/), the data were analysed, evaluated, and prepared (training corpus) for topic modelling. Topics were analysed using *LDA and Gensim*. The titles and abstracts of the papers were used for topic modeling, and the publication dates were used to show trends in LIS research. The authors' keywords were not used, as they were manually added and might not capture all the topics discussed in the papers. The paper's title was not used either, as it would only reiterate the topics already included in the abstract. The pre-processing methodology for the corpus has been implemented across several stages and levels and is discussed in detail below (Sections 4.1–Section 4.6).

Selection of data (crawling phase)

Fourteen thousand two hundred ninety-four (14294) raw articles (from 1970–2021) were downloaded or extracted from the Scopus database (https://www.scopus.com/sources.uri) against the three search terms, viz., *library*, *librarian*, and *theory*, appearing in the '*title*' and '*abstract*' of the paper. The corpus data covers the last 50 years, i.e., from 1970 to 2021, as a key sample. The authors have removed 61 articles from three unrelated journals. In addition, the authors excluded 368 titles and documents that lacked an abstract. Finally, a total of 13,865 articles has been considered as the corpus for this paper. All relevant bibliographic data, such as titles, abstracts, authors, journal names, etc., are recorded in a single Excel file. The sample articles, or corpus, from this study are 640 (actual sample data/corpus) research articles related directly to LIS (Library and Information Science) theory that met the objectives of our research.

The authors have included only those papers in which theories from LIS or other disciplines have been applied in the LIS domain. Only full-length articles written in English are collected where the specific application of theories was made. Sometimes, scholars used keywords such as ‘framework’, ‘model’, ‘pattern’, ‘paradigm, and 'method’ interchangeably and did not use them to search for information. Editorials, book reviews, letters, interviews, commentaries, and news items were also excluded from the analysis.

Pre-processing of corpus data/ data preparation (filtering phase)

The second phase is filtering (of corpus data), which focuses on filtering the corpus data to extract only the text's keywords. Pre-processing raw data is the most crucial step, as it can influence topic modeling results and improve output quality (Aggarwal & Zhai, 2012; Bi, Liang, Tang & Yang, 2018). It provides sound dimensionality reduction and removes unnecessary words from the unstructured textual data. It plays a significant role in transferring text from human language to machine-readable format (Rajasundari, Subathra, & Kumar, 2017). This process takes the longest to process raw data. This pre-processing step is essential for correctly analysing the corpus and understanding the relationships among its elements, thereby gaining valuable insights. Several steps are involved in creating the 'bag of words matrix'. All keywords were pre-processed and prepared for the topic modeling analysis in this step. Pre-processing mainly includes removing or cleaning punctuation, numbers, and special characters; converting to lowercase; performing phrase modeling; removing non-ASCII characters such as À µ ∅ © $ @; stemming; tokenization; and lemmatization. For example, lemmatization means transforming words to their most basic form or removing inflectional endings such as "seeing > saw" or "see." In the same way, stemming includes chopping off the ends of words or the removal of derivational affixes such as *"studying" > "study, "studies" >studi); t*okenization (character sequence and a defined document unit). Tokenization splits a text into words, phrases, or other meaningful parts, namely tokens. In addition, stop words were excluded to improve topic quality and reduce noise. Sometimes, authors use phrases in their articles. All were removed to avoid loss of context and meaning from the topic results.

Data curation (filtering phase)

The data were curated to develop a classification corpus. Initially, all these titles were ranked for readability using ASReview (https://asreview.nl/), an open-source, AI-compliant systematic review tool based on a machine learning framework. All titles, including abstracts, were examined one by one in the light of LIS theory to ensure that the content was relevant to the development of LIS theory. Moreover, 640 (4.61%) articles were finally considered for further research after removing all unrelated text unrelated to LIS theory.

Data formatting

This step is necessary to convert text data to an appropriate format for automated processing. In this phase, the extracted corpus data in Excel format is split into CSV (Comma Separated Values) files containing only titles and abstracts. Additionally, *ASReview* (https://asreview.nl/) and *OpenRefine* (https://openrefine.org/) have been used for further deep faceting operations on corpus data.

Data analysis

It includes analyzing corpus data with various tools to identify key topics and reveal relationships using different matrices. Finally, in the last phase, topic modeling was applied to the corpus to identify key issues, which were visualized as word clouds. It also shows the development of the subject (decade-wise) and the use and application of theories in the LIS domain. Let us examine the different processes and significant steps involved in our topic model, from selecting the corpus to initializing the model. The whole process is depicted in Figure 1.

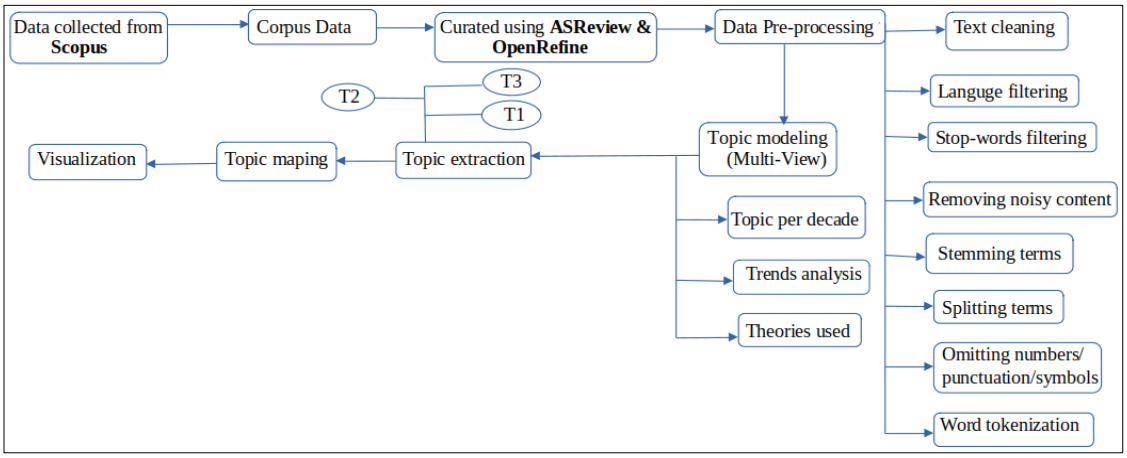

Figure 1: Data processing for topic modeling

Toolkits used

Many open-source tools and toolkits are available and employed to develop topic models in many applications. Some toolkits are discussed and mainly utilized in NLP (natural language processing). Many open-source tools meant for topic modeling have been used in different stages and layers of their implementation. Multiple algorithms are used to create this topic model, including Latent Dirichlet Allocation (LDA) and Gensim (an open-source Python library for automatically extracting semantic topics from documents) (https://radimrehurek.com/gensim/), which consider 'words to topics' and 'topics to documents'. Apart from these two tools, viz., the *Google Topic Model* (https://code.google.com/archive/p/topic-modeling-tool/) and *NLTK* (https://www.nltk.org/

nltk_data/(developed by Stanford University, USA) has also been consulted for selecting the best tool for our purposes. *Spacy* (https://spacy.io/) was used for data processing, and the *NLTK dictionary* was used for lemmatization (WordNetLemmatizer from NLTK), which removes unnecessary words or noisy text, such as adjectives, verbs, adverbs, symbols, and phrases. *Jupyter Notebook* (https://jupyter.org/) has also been used for scientific computing and for automatically making data readable. Authors have used *NTLK's library's 'English' stop-word list here.* The authors have also created a custom stop-word list for pre-processing corpus data effectively and efficiently.

Finally, '*Distant Reader'* (https://distantreader.org/), an '*off-the-shelf'* open-source software developed by Eric Lease Morgan at the University of Notre Dame, has been selected for our purpose due to its readability and web interface. Another necessary logic behind choosing this tool is that it is almost a finished product, and all the software required for topic modeling is already compatible with it. However, Distant Reader does not offer online or website support; it can be downloaded and installed locally.

Results

In essence, although topic models bring statistical analysis and can advance social science research, each algorithm has its own unique properties and relies on different assumptions. This section (Sub-section 5.1 to Sub-section 5.6) and subsequent sections (Section 6 & Section 7 consist of two parts, viz. *Identification of Core Topics and Identification of Core Theories* show eight types of results related to the research questions described above.

Readability measurement

On a scale from 0 to 100, where achieving zero is very difficult and achieving 100 is very easy. Our model has achieved an average readability score of 28. The following chart illustrates the overall readability of this model (Figure 2).

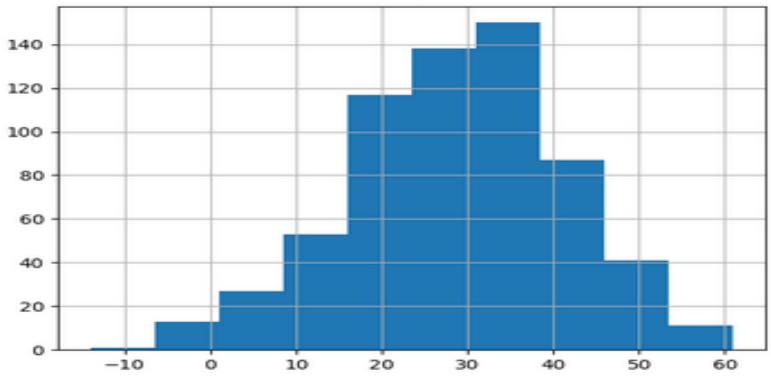

Figure 2: Readability measurement of the model

Bigrams and unigrams

In this topic modeling framework, we have created word clusters (a data visualization technique) rather than text clusters to show the most essential words in the corpus data based on their frequency of occurrence. Some of the more frequent words include the following and are also depicted in Fig. 3. Bigrams and Unigrams are presented to show word occurrences in corpus data. A unigram has been prepared, considering a word's occurrence without looking at the previous word. Here, it is assumed that all words in a document are generated from a distribution and that this distribution draws a topic for each document (Figure 3). The figure shows the word cloud for each topic.

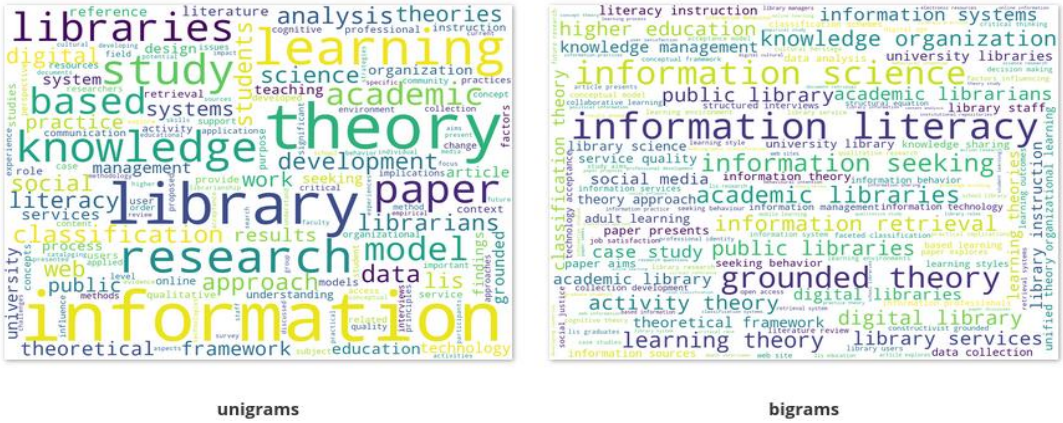

Figure 3: Bigram and unigram

On the other hand, Bigrams are prepared to consider only the previous word when predicting the current word; all word pairs are collocations. This Bigrams occurrence considers a relationship or edge between two words only if they appear in sequence. It consists of phrases containing two adjacent words. Figure 4 shows the most frequent (frequent appearance) two-word phrases i.e., Bigrams, and their occurrence in corpus data.

bigram

This is a table of type bigram and their frequencies. Use it to search & browse the list to learn more about your study carrel.

Show 10 entries Search:

bigram	frequency
information literacy	233
information science	180
grounded theory	152
information seeking	97
academic libraries	85
learning theory	73
knowledge organization	70
information retrieval	69
public libraries	64
activity theory	62

Figure 4: Frequency of occurrence (bigram)

Data visualization/word cloud

A word cloud visualizes terms by their frequency in a given topic. It shows the most frequent words and the co-occurrence frequencies of words in a text, where the relative size of a word reflects its frequency and indicates that it is often used together (Barua, Thomas & Hassan, 2014; Treude & Wagner, 2019). It can be insightful in topic modeling, showing which keywords occur most frequently in the given text. Visualizing topic terms with word clouds provides an overview of the probabilistic presence of terms within topics. This "word cloud" indicates the frequency and weight of words in a corpus or document. This model can display associated words, the most relevant documents matching the topic, and a list of related topics in the corpus. Here, words with a higher frequency are indicated by their larger size in the visualized graph (Figure 5).

Figure 5: Generation of a word cloud of the most frequent words

Topic modeling

As stated, the topic model is a statistical language model used to extract and uncover hidden structures in a collection of texts. If the corpus data is summed up in a single word, then that word might be '*information*,' '*learning*,' and '*theory*' ('topics'). Again, if corpus data is summed up into five topics and each topic is denoted with three words, then those topics and their most significantly associated words would be '*information*, *theory*, *knowledge*'; '*learning*, *theory*, *information*'; '*learning, theory*, *students*'; '*libraries*, *library*, *theory*'; '*library*, *theory*, *development*' (Table 2). Table 2 shows the topics and word occurrences in the corpus.

Table 2
Topic modeling

Three words: 'topics.'	
Words	**Frequency of Occurrence**
information	436
learning	60
theory	507
Five words: 'topics.'	
information, theory, knowledge	436
learning, theory, information	504
learning, theory, students	350
libraries, library, theory	470
library, theory, development	434

The study shows the top five topics, each denoted by three keywords (Figure 6).

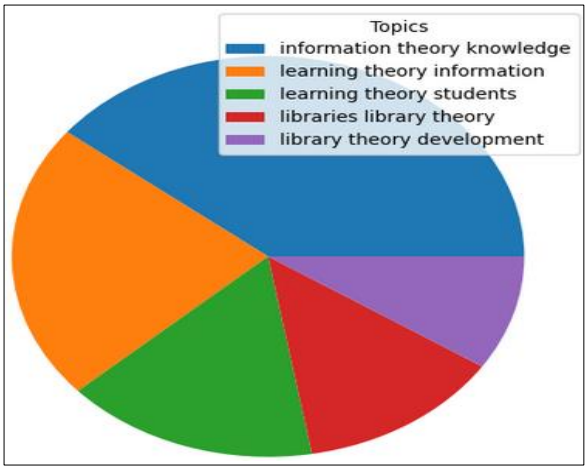

Figure 6: Topic modeling (five topics)

Not all topic modeling techniques apply to every setting or type of text. It depends on the kind of document or the nature of the text. Each topic modeling technique has limitations and constraints (Chen et al., 2016). However, Distant Reader has its limitations. For example, it cannot generate the best topic or model and does not provide perplexity or topic coherence. However, calculating these values is essential for assessing a topic model's performance. The authors used Python's Gensim library in a *Jupyter Notebook* to calculate these two values.

Topic model investigated

Naturally, the question arises: How helpful is the model in identifying topics from the corpus? The program or software authors have used does not give us the meanings of the words, but rather groups words based on co-occurrence patterns. It does not convey the text's complete sense, but it provides a good overview of the themes. Here, five topics have been selected, as experts (Mathew, Agrawal & Menzies, 2017) opined that selecting 5–10 topics is an excellent heuristic for the model. Here, topics are labelled from '0' to '4,' and the model is built on five topics (Figure 7), where each topic is a combination of keywords, and each keyword contributes a certain weight to the topic. For each topic, there is a list of words most relevant to the particular topic. All five topics are distinguishable, and almost all the words are appropriate. The top 5 topics consisted of 10 keywords, and their probabilities are displayed. Here, five topics have been presented, focusing on the quality of corpus data. These five topics cover most of the angles of the sample data or corpus. Including more topics increases the risk of overfitting, and the model may not form proper clusters. So, let us see what a topic looks like in the proposed model. Figure 7 shows the top 10 keywords that contribute to this topic (topic 0) are '*learning*', '*library*', '*theory*', and so on, and the weight of the word/topic '*learning*' on topic '0' is 0.028. These weights reflect how important a keyword is to that particular topic. We can say that the proposed model has successfully identified the five topics from the dataset.

```
[(0,
  '0.028*"learning" + 0.018*"library" + 0.018*"theory" + 0.016*"student" + '
  '0.015*"librarian" + 0.011*"academic" + 0.007*"research" + 0.007*"study" + '
  '0.007*"literacy" + 0.007*"practice"'),
 (1,
  '0.053*"information" + 0.041*"theory" + 0.027*"knowledge" + 0.015*"paper" + '
  '0.013*"concept" + 0.009*"research" + 0.009*"organization" + '
  '0.008*"activity" + 0.007*"science" + 0.007*"document"'),
 (2,
  '0.032*"library" + 0.021*"theory" + 0.013*"research" + 0.013*"information" + '
  '0.011*"study" + 0.009*"paper" + 0.009*"cataloging" + 0.007*"model" + '
  '0.007*"management" + 0.007*"science"'),
 (3,
  '0.036*"information" + 0.020*"study" + 0.016*"model" + 0.016*"user" + '
  '0.014*"seeking" + 0.010*"research" + 0.008*"using" + 0.008*"behavior" + '
  '0.008*"used" + 0.008*"theory"'),
 (4,
  '0.033*"classification" + 0.014*"system" + 0.013*"model" + 0.013*"based" + '
  '0.012*"metadata" + 0.011*"retrieval" + 0.011*"indexing" + 0.008*"theory" + '
  '0.007*"document" + 0.006*"approach"')]
```

Figure 7: Topic modeling results with topic term weights

Topic coherence and model perplexity

As with other models, the quality of this topic model can be assessed using perplexity and coherence values. The coherence method was proposed by Röder, Both, and Hinneburg (2015) to evaluate a topic's classification effectiveness by measuring the degree of semantic co-occurrence between high-probability words in the topic. It is one of the most popular metrics for evaluating topic models, measuring the degree of semantic similarity among words within a topic. The Coherence Score is an evaluation method used to measure topic consistency resulting from the application of topic model algorithms. Assessing the coherence of a topic is essential and is considered the best way to provide insight into its interpretability for humans. It measures how similar the words are to each other in the corpus data or how often a topic is encountered (i.e., topic strength). The coherence score indicates the quality of the learned topics via a single quantitative, scalar value and also measures the relative distance between words within a topic. So, the higher the score, the better the topics are extracted.

Apart from topic modeling, the authors have used the perplexity and coherence scores in this experiment to assess how interpretable the topics are to humans, how well the topics are extracted, and how informative the topics are derived from the proposed model. The perplexity metric measures how well a topic model predicts the issues of new data. It considers and measures the predictive ability of a statistical model for documents. Perplexity is inversely proportional to the likelihood of the data given in the model. The smaller or lower the perplexity value, the better the model with that number of topics (Griffiths & Steyvers, 2004). For a good model, the perplexity score should be low, and the coherence score should be high. Here, the perplexity score is negative (-7.494) (Figure 8), which indicates the model's better performance or prediction ability. The coherence score of our model is 0.343 (Figure 8). By using this value, authors have tried to measure the degree of semantic similarity between high-scoring words in the derived topics. The objective is to help the reader distinguish between semantically interpretable issues and those that are artifacts of statistical inference.

```
Perplexity:   -7.494536512558358
Coherence Score:   0.34311529303084953
```

Figure 8: Perplexity value and topic coherence value

Here, model number 15 (Figure 9) is the most helpful in representing the topics covered in

the corpus data. As a result, the model with 15 topics was chosen as the number of topics (Figure 9). It indicates the highest coherence score, which implies better interpretability. The model with the best coherence score has 15 topics. It shows that most topics exhibit a relatively high level of (subjective) coherence. The coherence score also increases with the number of topics. Here, the coherence value has been calculated for each number of topics to determine the optimal number. Here, the coherence value peaks at 15, so we will use 15 topics; the higher the topic coherence, the more human-interpretable the topic.

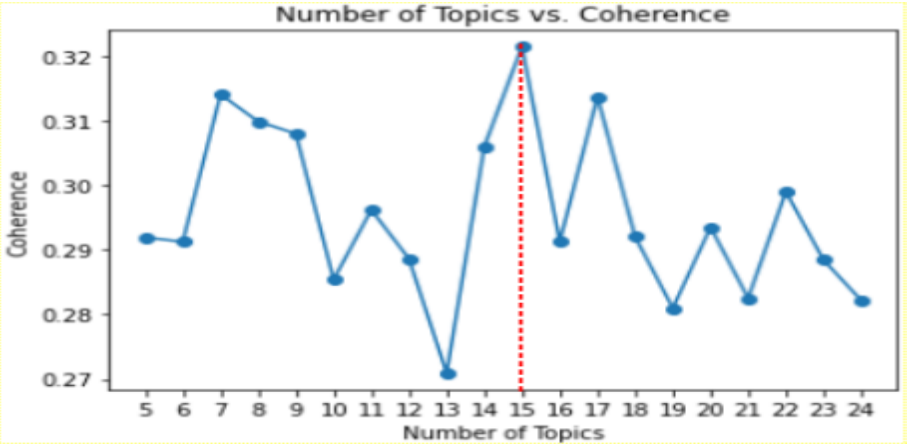

Figure 9: Best topic model

Topic identification (decade-wise)

This section aims to identify the most prevailing research trends (e.g., Identification of Core Topics) in LIS from 1970 to 2021. By grouping topics into broad thematic areas, we can trace the evolution of LIS research during this period. Even from this model, we can track how topics evolved, i.e., "topics over time" (TOT). We have grouped the entire period into five blocks or clusters (Table 3), with each block spanning 10 years (except for cluster V), to understand the chronological growth of subjects or the transition of the knowledge structure in the LIS domain. A total of 5 topics, each with the top 10 keywords, are labeled based on their frequency of occurrence in the corpus. In doing so, in some cases, similar topics or keywords overlap and are thus avoided.

Table 3
Results of the topic model (block/cluster-wise)

Period (Ten Years)	No. of Articles	Topics (5) & Keywords (10) (block/cluster-wise)	Bigram (10 words)
1970-1979 (Block/Cluster 1)	12	Topics [information, retrieval, document], [theory, indexing, systems], [model, theory, information], [librarianship, online, information], [information, theory, retrieval] Keywords information, theory, retrieval,	information retrieval (search patterns, fuzzy set, Boolean retrieval), information theory, information service, set theory, online bibliographic, contingency theory, serials cataloging, bibliographic searching, Bayesian model

Period (Ten Years)	No. of Articles	Topics (5) & Keywords (10) (block/cluster-wise)	Bigram (10 words)
		indexing, librarianship, fuzzy, boolean, reference, shannon, library, boolean, bayesian	
1980-1989 (Block/Cluster II)	28	Topics [information, theory, indexing], [theory, library, interactional], [information, theory, techniques], [theory, information, job], [analysis, information, research] Keywords information, theory, library, Ranganathan, indexing, reference, organization, learning, Goffman, education, classification, Boolean, management	information retrieval, information theory, classification theory, information processing, job satisfaction, information seeking, collection management, grounded theory, indexing theory, organization theory, knowledge organization, reference policy/services
1990-1999 (Block/Cluster III)	92	Topics information, theory, library], [information, theory, library], [information, library, theory], [information, theory, library], [theory, information, reference] Keywords information, library, theory, knowledge, reference, communication, classification, ontology, OAI, indexing, cataloging	information seeking, knowledge organization, information retrieval, information literacy, learning theory, collection development, grounded theory, classification schemes, job satisfaction, style theory, communication theory, unified theory, bibliographic control, information theory, classification theory
2000-2009 (Block/Cluster IV)	271	Topics [learning, theory, library], [information, web, theory], [information, library, theory], [information, theory, paper], [information, library, theory] Keywords information, theory, web, library, knowledge, learning, classification, theory, literacy, activity, cultural, grounded, metadata, digital, reference, Ranganathan, catalog	information literacy, grounded theory, digital library, activity theory, public library, information seeking, learning theory, classification theory, academic library, knowledge organization, seeking behaviour or information behaviour, institutional repositories, open access, digital culture, adult learning
		Topics library, use, theory],	information literacy, grounded theory, academic libraries,

Period (Ten Years)	No. of Articles	Topics (5) & Keywords (10) (block/cluster-wise)	Bigram (10 words)
2010-2021 (Block/Cluster V)	237	[information, theory, learning], [information, theory, knowledge], [research, library, information], [learning, information, study] Keywords information, library, classification, theory, learning, digital, Ranganathan, knowledge, ontology, Moodle	public libraries, information seeking, social media, learning theory, library services, activity theory, knowledge management, digital library, unified theory, learning theories,

Discussion

Here, the discussion is presented in two parts/sections: (A) *identification of core topics* and (B) *identification of unique theories* used in LIS research during the period, as shown in Table 3. The objective is to identify the most popular research topics and the most prominent theories used in LIS research through topic modeling. Here, the discussion is organized into five broad clusters/blocks within two broad parts/sections, as stated above.

Identification of core topics (decade-wise)

This particular section addresses the following questions:

Q.1. Which topics were most discussed? Or

Q.2. What were the research topics studied in the LIS area during the period?

Identifying the specific focus of the topics under each decade or cluster is impossible for several reasons. It is because one topic has been discussed for many decades. For example, information retrieval has been addressed in all the clusters from different viewpoints. Before 1990 or the introduction of computers in information retrieval, pre-coordinate/post-coordinate indexing were used, and Uniterm indexing came to the market after 1985. However, after 1990, different retrieval models (such as the vector space model, probabilistic model, and cognitive model) were introduced. Thus, different search techniques (such as Boolean logic) were used by retrieval systems. Here, the authors have tried to present the decade-wise trends for the subject LIS. The topics could be grouped into the following six broad categories to better understand the subject and our purposes in preparing the paper: The main issues may be like the following, and all these key terms have been incorporated into these six (06) main groups:

- *Information retrieval;*
- *Information seeking (information seeking behaviour/information needs);*
- *Knowledge organisation (information/knowledge management);*
- *User study (information literacy, user education);*
- *Metric study (such as bibliometrics); and*
- *Digital library (institutional repository, open access/open knowledge).*

A.1. Cluster I (1970–1979): The primary focus was on probabilistic information retrieval methods. However, it originated in the late 1950s. Various search techniques, including fuzzy logic, set theory, and Boolean operators, were used. All these techniques were best suited for text retrieval in bibliographic databases. During this period, various models for document retrieval were developed. Apart from that, the decade was found to be the decade of

bibliographic classification and indexing. There was a strong emphasis on classification and indexing, and theories related to these areas were also areas of research. General systems theory (GST), as a meta-theory proposed by Ludwig von Bertalanffy, was outlined. Again, critical theory, which began in Germany in the 1920s, entered the LIS field. Moreover, knowledge management appeared as a discipline before 1980.

A.2. Cluster II (1980–1989): Each cluster saw some progress towards information retrieval. However, developing user-centric information retrieval systems was dominant in cluster II (1980–1989). Information processing, or information management, was the main focus, and a range of search strategies was used in this cluster. By the 1980s, full-text search had become commercially established in online retrieval systems. Here, searching was restricted to document searching, and many retrieval models used ranking techniques to improve search effectiveness. However, Boolean retrieval dominated the commercial world. Apart from that, theories of classification and indexing were introduced to improve the organisation and representation of knowledge. Grounded theory, initially developed in sociology in 1960, was introduced in information science research as a qualitative data analysis method. However, information scientists like Shannon (1948) and Shannon and Weaver (1949) had already developed and used statistical communication theory (also known as the classical theory of communication or information theory) in information science.

A.3. Cluster III (1990–1999): The arrival of the Internet in 1990 and the search engines changed the information search process for users. As a result, information needs and search behaviour have changed. There has been a dramatic shift from proprietary information systems and library resources to web-based systems and e-resources. Learning theory was introduced to understand how scholars receive, process, and retain knowledge in this new environment. In 1991, Keith Devlin (1991), developed an information semantic theory, and in 1994, Wolfgang Hofkirchner (2009) put forward a new concept of unified information theory.

A.4. Cluster IV (2000–2009): We witnessed a paradigm shift towards greater emphasis on scholarly communications and papers related to open access, digital libraries, and institutional repositories. This is due to the emergence of many open-access initiatives at the international level, such as *the BOAI, the Berlin & Bethesda statements, popularly known as the* 3Bs. As a result, resources are now available on open platforms. After 2000, we saw the emergence of concept search in information retrieval, where humans acted as intermediaries.

A.5. Cluster V (2010–2021): It is noticed that during this decade, the traditional topics have lost their value due to the advancement of ICT and the open-source software movement. New issues like integrated library systems (Moodle), machine learning, and big data were introduced. A decreasing trend was observed among traditional topics, and the advent of social media influenced LIS research. Moreover, significant changes occurred in natural language processing, information retrieval (image processing), and knowledge management. Various language models have been introduced, and data science has emerged as a cutting-edge field of study. Activity theory originated in psychology and was introduced to inform the design and development of information systems research. However, *papers on* '*information seeking*' and '*information retrieval*' appeared in almost all the clusters with a new look. Two well-known works, viz. '*Kuhlthau's Information Search Process' (ISP)* and '*Inversen's Cognitive Information Retrieval'* were also introduced in the LIS domain.

Identification of core theories (decade-wise)

The main objectives of this section are to address the following questions:

Q.1. What were the theories applied in LIS research (cluster-wise as stated above)? Or

Q.2. Which were the most critical theories discussed in the corpus data during the period?

Many LIS researchers have developed theories, especially in information science, including information-seeking behaviour. Initially, a theory was created for a particular discipline and has since been modified and utilised in other disciplines or for different sets of phenomena. The use and application of theories for conducting research are everyday in any academic discipline, and LIS is no exception. This section shows the use of theories in corpus data from 1970 to 2021.

B.1. Block/Cluster I (1970-1979): In this period, twelve (12) documents were included in the corpus; the cluster consists of topics (5) and keywords (10). Some popular theories, such as *organizational theory*, *fuzzy set theory*, *grounded theory*, *cultural-historical activity theory*, and *social reproduction theory*, were used during this period. However, *learning theory* was found to be the most dominant theory in researchers' papers.

B.2. Block/Cluster II (1980-1989): In this block, twenty-eight (28) documents were reviewed and highlighted, including some key topics (5) and corresponding keywords (10). Here, we have identified applications of popular theories, such as cognitive load theory, dual coding theory, adaptive resonance theory, field theory, the socio-cultural theories of Lev Vygotsky and Jean Lave, the theory of green library management, and the theory of management strategy, that were used during this period. However, the *grounded theory* and *learning theory* were in focus during this cluster.

B.3. Block/Cluster III (1990-1999): Ninety-two (92) articles were reported, and some key topics (5), including keywords (10), were extracted. Like cluster I and cluster II, articles have covered many popular theories, such as *Darwin's theory*, *Pask's conversation theory*, the *cognitive theory of Howard Gardner*, the *core competency theory of Selznick*, the *theory of symbolic interactionism*, *Critical Race theory*, the *educational theory of John Dewey*, *Howard Gardner's multiple intelligence theory*, the *Theory of Change*, the *Theory of Reasoned Action (TRA)*, *management theory of Henry Mintzberg*, *Mayer's theory of multimedia learning*, *Ikujiro Nonaka's theories*, *cognitive development theory* (Perry's theory), and so on. Unlike *learning or grounded theory, most articles were based on the Technology Acceptance Model (TAM)* or the *Unified Theory of Acceptance*.

B.4. Block/Cluster IV (2000-2009): Here, two hundred seventy-one (271) articles were reported, and some key topics (5), including keywords (10), were extracted. Some of the popular theories covered are *activity theory*, *change management theory*, the *critical theories of Stuart Selber and Gunther Kress*, *structuration theory*, *Ranganathan's classification theories*, *information literacy theories*, the *postmodern theory of knowledge organization*, *cataloguing and classification theory*, *sense-making theory*, the *theory of collection development*, and *information systems design theory*. Here, it is difficult to say which theories were mainly used. However, *grounded theory, learning theory, information theory, or information library theory* were in focus.

B.5. Block/Cluster V (2010-2021): In this period, a total of two hundreds thirty seven (237) articles were found covering different theories such as the *theory of indexing* (Jonker, Heilprin, Landry, and Salton), *cataloguing theory* (Anthony Panizzi and Seymour Lubetzky), *classification theory* or *facet analytical theory* (Ranganathan), *information retrieval theory*

(fuzzy set theory.), *5S Theory* (digital library), *theory of book selection* (Ranganathan), and so on. Although there is debate over whether these (e.g., cataloguing theory, classification theory, theory of book selection, theory of collection development) are theories of LIS, researchers differ in their opinions on this (Roy & Mukhopadhyay, 2023).

Based on the discussion, the theoretical foundation of the LIS domain is relatively weak, and the field relies on other disciplines for its theoretical underpinnings. The majority of theories were from social sciences, and research in LIS is confined to theories developed and used in complementary disciplines. A recent study (Roy & Mukhopadhyay, 2023) also shares this view and reports that the theoretical foundations of LIS are understudied. In addition to LIS theory, researchers in the LIS field have used theories from other disciplines to pursue their research. This study also shows that each cluster has its own research focus. An in-depth analysis of keywords, titles, and abstracts suggested commonalities among topics and theories used in LIS. However, many developments occurred due to technological advancements, especially the Internet and open-source software such as Koha (for library automation) - https://koha-community.org/, DSpace (for developing a digital repository) - https://dspace.org/. As a result, we noticed significant changes in topics but slight changes in categories.

Conclusion

Topic modeling is a technique used in information retrieval to automatically identify and organize the hidden topics in an extensive collection of textual documents. It provides a way to understand and summarize large amounts of text and is widely used in natural language processing tasks. Baghmohammad, Mansouri, and Cheashmehsohrabi (2021) rightly opined that it could help researchers identify the thrust areas of LIS. But over time, the intent of topic modeling has changed in many ways. New topics are emerging/coming, and old ones become obsolete over time in all domains. Initially, topic models were applied mainly in text mining, a sub-branch of data mining, and information retrieval, and are familiar with content analysis, or citation analysis. All these analyses are mainly for short texts and do not span related fields.

With advances in topic modeling, various techniques, including an extended version of LDA based on related theories, have emerged, but each algorithm has its strengths and weaknesses. Different models and model categories have distinct characteristics and thus coexist to serve in other contexts and on different corpora. Some of the techniques mentioned are very promising but need more study and development to become widely used. Specific questions, such as how to select the optimal number of topics, how to choose priors, and which inference technique is best suited to the purpose, remain unresolved. As stated, the topic model has several advantages and certain restrictions. LDA is a standard topic modeling method, but it has several drawbacks. One of the main disadvantages of LDA is that it can produce ambiguous or incoherent topics, especially if the data is noisy, sparse, or heterogeneous. It relies on the assumption that the words in each topic are related to one another and meaningful, but this may not always hold in practice. For example, some words may have multiple meanings, some topics may overlap or be too broad, and some documents may cover multiple or unrelated topics. Another disadvantage of LDA is that it can be computationally intensive, especially when the corpus is extensive, the number of topics is high, or the model is complex. LDA requires multiple iterations and optimization steps to estimate topic distributions, which can consume significant resources and memory. In addition, LDA has other limitations, such as it generates a limited number of topics; produces incoherent topics, especially with noisy or sparse

data; may not capture nuanced variations or specific patterns; can be computationally demanding, especially with large datasets or complex models; performs poorly with short documents; cannot model advanced data relationships; does not consider word order, which is a key or crucial characteristic of language; randomly assigns topics to documents, so the results are affected by document order; as is unsupervised, so it may not capture nuanced variations or specific patterns that supervised methods can handle; requires careful adjustment of hyperparameters, which can be time-consuming.

Despite several limitations, it has many advantages over other models, making it popular among researchers for automatic text analysis that extracts the main topics in a corpus. For these reasons, it has been a powerful tool, particularly in Social Sciences research. For example, it helps researchers gain a quick overview of the primary contents from a large volume of text data (DiMaggio, Nag & Blei, 2013; Nelson, 2020). As a dimension-reduction technique, topic modeling transforms a large sample of text into a much smaller set of topics. Another notable advantage is that it offers scholars a novel analytical perspective, enabling the identification of patterns that would likely remain undetected through manual coding alone, particularly when dealing with large volumes of textual data (DiMaggio, Nag & Blei, 2013).

This research certainly has its limitations. Not all these models are always suited for modeling more complex data relationships. This prototype, keeping in mind the limitations of the existing models used in different text-mining studies, has been designed to organise and analyse large corpora more effectively and efficiently than before, enabling better information retrieval. Apart from these, this model may be used for *automated classification and categorization of the subject, summarization of long texts for qualitative analysis, query expansion, ranking documents by relevance, personalizing search results, or making recommendations by mapping user preferences* to topics. It also makes it easier to find clusters of similar documents in the corpus and related secondary issues. From a librarian's point of view, this framework may help LIS professionals find similar texts or offer users suggestions for what to read next. The authors cannot claim that this topic model can select the optimal number of topics from any corpus. Even authors cannot say this model is foolproof or performs like a human, but the results closely represent the text. Besides, the success of any topic model depends on human judgement (Hannigan et al., 2019), the domain knowledge of the scholars (Egger & Yu, 2022), and how far it can accomplish the task(s) for which it was built (Carter, Brown, & Rahmani, 2016).

However, this robust model (from the same source) generates similar solutions with strong predictive ability and can help avoid duplicate topics. This solution has nicely fulfilled its promises (as stated in section 3) by discovering major topics and analysing which documents cover them. It can find patterns of word use and connect and correlate documents that share similar patterns. Even so, it can differentiate topics that are too similar. This framework may be used for any exploratory analysis, discovery, browsing, or finding similar or related issues in any domain. Although topic models have quantified short-text data, both the interpretation and justification of the results come at the expense of data accuracy. The results may vary and depend on the nature of the corpus, the techniques/tools used, the models applied, the data preprocessing, the corpus size, and many other factors. Thus, future research should continue to explore the effectiveness of topic modeling algorithms across different platforms. However, there is room for further improvement, allowing the researchers to cover a broader range of LIS journals and access more full-text data to gain a more in-depth understanding of the knowledge

structure of the LIS domain.

References

Afacan Adanir, G. (2019). Detecting topics of chat discussions in a computer-supported collaborative learning (CSCL) environment. *Turkish Online Journal of Distance Education,* 20(1), 96 -114. https://doi.org/10.17718/tojde.522398

Aggarwal, C. C. & Zhai, C. (2012). *Mining text data*. Springer: New York. Agrawal, A., Fu, W. & Menzies, T. (2018). What is wrong with topic modeling? And how to fix it using search-based software engineering. *Information and Software Technology,* 98(4), 74-88. https://doi.org/10.1016/j.infsof.2018.02.005

Albalawi, R., Yeap, T. H. & Benyoucef, M. (2020). Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence*, 3(42). https://doi.org/10.3389/frai. 2020.00042

Amado, A., Cortez, P., Rita, P. & Moro, S. (2018). Research trends on big data in marketing: A text mining and topic modeling-based literature analysis. *European Research on Management and Business Economics*, 24(1), 1-7. https://doi.org/10.1016/j.iedeen.2017.06.002

Angelov, D. (2020). *Top2Vec: Distributed Representations of Topics*. Retrieved from https://arxiv.org/pdf/2008.09470

Asuncion, H. U., Asuncion, A. U. & Taylor, R. N. (2010). Software traceability with topic modeling. *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering* (Volume 1) (pp. 95-104). Cape Town, South Africa: ACM. https://doi.org/10.1145/1806799.1806817

Bagheri, A., Saraee, M. & De Jong, F. (2014). ADM-LDA: An aspect detection model based on topic modeling using the structure of review sentences. *Journal of Information Science*, 40(5), 621-636. https://doi.org/10.1177/0165551514538744

Baghmohammad, M., Mansouri, A. & Cheashmehsohrabi, M. (2021). Identification of the topic, the development process of the knowledge and information science field based on the topic modelling (LDA). *Iranian Journal of Information Processing and Management*, 36(2), 297-328. https://doi.org/10.35050/JIPM010.2020.001 [in Persian]

Barua, A., Thomas, S. W. & Hassan, A. E. (2014). What are developers talking about? An analysis of topics and trends in Stack Overflow. *Empirical Software Engineering,* 19(3), 619-654. https://doi.org/10.1007/s10664-012-9231-y

Bauer, S., Noulas, A., Séaghdha, D. O., Clark, S. & Mascolo, C. (2012, September). Talking places: Modelling and analysing linguistic content in Foursquare. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing* (pp. 348-357). IEEE. https://doi.org/10.1109/SocialCom-PASSAT.2012.107

Bi, T., Liang, P., Tang, A. & Yang, C (2018). A systematic mapping study on text analysis techniques in software architecture. *Journal of Systems and Software,* 144(10), 533-558. https://doi.org/10.1016/j.jss.2018.07.055

Bianchi, F., Terragni, S. & Hovy, D. (2020). Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers)*. *Association for Computational Linguistics.* https://doi.org/10.18653/v1/2021.acl-short.96

Blei, D. M. & Jordan, M. I. (2003). Modeling annotated data. S*IGIR '03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 127–134). Toronto, Canada. https://doi.org/10.1145/860435.860460

Blei, D. M. & Lafferty, J. D. (2006). Dynamic topic models. *In Proceedings of the 23rd International Conference on Machine Learning* (ICML '06) (pp. 113-120). Pittsburgh: ACM. https://doi.org/10.1145/1143844.1143859

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research,* 3, 993–1022. https://dx.doi.org/http://dx.doi.org/10.1162/jmlr.2003.3.4-_5.993

Bozdogan, Z. & Kara, R. (2024). Comparison of traditional and modern topic-modeling algorithms for topic determination in official documents. *Journal of Engineering Research and Applied Science*, 13(1), 2490-2499. Retrieved from https://www.journaleras.com/index.php/jeras/article/view/328

Bulygin, D., Musabirov, I., Suvorova, A., Konstantinova, K. & Okopnyi, P. (2018). Between an arena and a sports bar: Online chats of esports spectators. Retrieved from https://arxiv.org/pdf/1801.02862.pdf

Carter, D. J., Brown, J. & Rahmani, A. (2016). Reading the high court at a distance: Topic modeling the legal subject matter and judicial activity of the high court of Australia, 1903-2015. *University of New South Wales Law Journal*, 39(4), 1300-1354. Retrieved from https://www.unswlawjournal.unsw.edu.au/wp-content/uploads/2017/09/39-4-16.pdf

Chang, J. & Blei, D. M. (2009). Relational topic models for document networks. *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics* (AISTATS) (Vol. 9) (pp. 81–88). Clearwater Beach, Florida, USA. Retrieved from https://proceedings.mlr.press/v5/chang09a/chang09a.pdf

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, (pp.288–296). USA: Curran Associates Inc. https://doi.org/10.5555/2984093.2984126

Chen, B., Tsutsui, S., Ding, Y. & Ma, F. (2017). Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval. *Journal of Informetrics*, 11(4), 1175-1189. https://doi.org/10.1016/j.joi.2017.10.003

Chen, B., Zhu, L., Kifer, D., & Lee, D. (2010, July). What is an opinion about? Exploring political standpoints using an opinion scoring model. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 24, No. 1, pp. 1007-1012). https://doi.org/10.1609/aaai.v24i1.7717

Chen, H., Zhang, G., Zhu, D. & Lu, J. (2017). Topic-based technological forecasting based on patent data: A case study of Australian patents from 2000 to 2014. *Technological Forecasting and Social Change*, 119(C), 39-52. https://doi.org/10.1016/j.techfore.2017.03.009

Chen, J., Wei, W., Guo, C., Tang, L., & Sun, L. (2017). Textual analysis and visualization of research trends in data mining for electronic health records. *Health Policy and Technology*, 6(4), 389-400. https://doi.org/10.1016/j.hlpt.2017.10.003

Chen, P., Zhang, N. L., Liu, T., Poon, L. K., Chen, Z., & Khawar, F. (2017). Latent tree models for hierarchical topic detection. *Artificial Intelligence*, 250, 105-124. https://doi.org/10.1016/j.artint.2017.06.004

Chen, S. H., Santoso, A., Lee, Y. S., & Wang, J. C. (2015, September). Latent Dirichlet Allocation-Based Blog Analysis for a Criminal Intention Detection System. In *2015, the International Carnahan Conference on Security Technology (ICCST)* (pp. 73-76). https://doi.org/10.1109/CCST.2015.7389660

Chen, T. H., Thomas, S. W. & Hassan, A. E. (2016). A survey on the use of topic models when mining software repositories. *Empirical Software Engineering,* 21(5), 1843–1919. https://doi.org/10.1007/s10664-015-9402-8

Chen, T. H., Thomas, S. W., Nagappan, M., & Hassan, A. E. (2012, June). Explaining software defects using topic models. In *201, the 29th IEEE Working Conference on Mining Software Repositories (MSR*) (pp. 189-198). IEEE.

Cheng, V. C., Leung, C. H., Liu, J. & Milani, A. (2013). Probabilistic aspect mining model for drug reviews. *IEEE Transactions on Knowledge and Data Engineering*, 26(8), 2002-2013. https://doi.org/10.1109/TKDE.2013.175

Cheng, Z. & Shen, J. (2016). On effective location-aware music recommendation. *ACM Transactions on Information Systems*, 34(2), 1-32. https://doi.org/10.1145/2846092

Cohen, R. & Ruths, D. (2013). Classifying political orientation on Twitter: It's not easy!. In *Proceedings of the International AAAI Conference on Web and Social* Media (Vol. 7, No. 1, pp. 91-99). https://doi.org/10.1609/icwsm.v7i1.14434

Cristani, M., Perina, A., Castellani, U. & Murino, V. (2008, June). Geo-located image analysis using latent representations. In *2008 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1-8). IEEE.

Dam, H. K. & Ghose, A. (2016, August). Analyzing topics and trends in the PRIMA literature. In *International Conference on Principles and Practice of Multi-Agent Systems* (pp. 216-229). Cham: Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-44832-9_13

Dang, Q., Gao, F. & Zhou, Y. (2016). Early-detection method for emerging topics based on dynamic Bayesian networks in microblogging networks. *Expert Systems with Applications*, 57(3), 285-295. https://doi.org/10.1016/j.eswa.2016.03.050

Das, R., Zaheer, M. & Dyer, C. (2015, July). Gaussian LDA for topic models with word embeddings. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 795-804). https://doi.org/10.3115/v1/P15-1077

Das, S., Dixon, K., Sun, X., Dutta, A. & Zupancich, M. (2017). Trends in transportation research: Exploring content analysis in topics. *Transportation Research Record*, 2614(1), 27-38. https://doi.org/10.3141/2614-04

De Battisti, F., Ferrara, A. & Salini, S. (2015). A decade of research in statistics: A topic model approach. *Scientometrics*, 103(2), 413–433. https://doi.org/10.1007/s11192-015-1554-1

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, Th. K. & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science,* 416), 391-407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6%3C391::AID-ASI1%3E3.0.CO;2-9

Deng, X., Smith, R. & Quintin, G. (2020). Semi-supervised learning approach to discover enterprise user insights from feedback and support. https://doi.org/10.48550/arXiv.2007.09303

Devlin, K. (1991). *Logic and information*. New York: Cambridge University Press.

DiMaggio, P., Nag, M. & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of the U.S. Government arts funding. *Poetics,* 41(6), 570–606. https://doi.org/10.1016/j.poetic.2013.08.004

Ding, Y. (2011). Topic-based PageRank on author co-citation networks. *Journal of the American Society for Information Science and Technology*, 62(3), 449–466. https://doi.org/10.1002/asi.21467

Do, H. H., Prasad, P. W., Maag, A. & Alsadoon, A. (2019). Deep learning for aspect-based sentiment analysis: a comparative review. *Expert Systems with Applications*, 118, 272-299. https://doi.org/10.1016/j.eswa.2018.10.003

Egger, R., Yu, J. (2022). Epistemological Challenges. In: Egger, R. (eds) Applied Data Science in Tourism. Tourism on the Verge. Springer, Cham. https://doi.org/10.1007/978-3-030-88389-8_2

Eidelman, V., Boyd-Graber, J. & Resnik, P. (2012). Topic models for dynamic translation model adaptation. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (8-14 July 2012, Jeju, Republic of Korea) (pp. 115–119). Retrieved from https://aclanthology.org/P12-2023.pdf

Eisenstein, J., O'Connor, B., Smith, N. A. & Xing, E. (2010, October). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 1277-1287). https://dl.acm.org/doi/pdf/10.5555/1870658.1870782

El-Diraby, T., Shalaby, A. & Hosseini, M. (2019). Linking social, semantic, and sentiment analyses to support modeling transit customers' satisfaction: Towards a formal study of opinion dynamics. *Sustainable Cities and Society*, 49, 101578. https://doi.org/10.1016/j.scs.2019.101578

Elragal, A., & Klischewski, R. (2017). Theory-driven or process-driven prediction? Epistemological challenges of big data analytics. *Journal of Big Data,* 4(2), 19. https://doi.org/10.1186/s40537-017-0079-2

Feng, Y. & Lapata, M. (2010). Topic models for image annotation and text illustration. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL (pp. 831-839). Association for Computational Linguistics. Retrieved from https://aclanthology.org/N10-1125.pdf

Figuerola, C. G., Marco, F. J. G. & Pinto, M. (2017). Mapping the evolution of library and information science (1978–2014) using topic modeling on LISA. *Scientometrics,* 112(3), 1507-1535. https://doi.org/10.1007/s11192-017-2432-9

Gallagher, R. J., Reing, K., Kale, D. & Ver Steeg, G. (2017). Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics,* 5, 529–542. https://doi.org/10.1162/tacl_a_00078

Garbhapu, V. K. & Bodapati, P. (2020). A comparative analysis of Latent Semantic Analysis and Latent Dirichlet Allocation topic modeling methods using Bible data. *Indian Journal of Science and Technology,*13(44), 4474-4482. https://doi.org/10.17485/IJST/v13i44.1479

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, 61, 115-125. https://doi.org/10.1016/j.dss.2014.02.003

Gethers, M. & Poshyvanyk, D. (2010, September). Using relational topic models to capture coupling among classes in object-oriented software systems. In *2010, IEEE International Conference on Software Maintenance* (pp. 1-10). https://doi.org/10.1109/ICSM.2010.5609687

Greene, D. & Cross, J. P. (2015, June). Unveiling the political agenda of the European Parliament plenary: A topical analysis. In *Proceedings of the ACM web science conference* (pp. 1-10). https://doi.org/10.1145/2786451.2786464

Griffiths, T. L. & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101 Suppl 1(Suppl 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Grimmer, J. & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3), 267–297. https://doi.org/10.1093/pan/mps028

Guo, W. & Diab, M. (2011, July). Semantic topic models: Combining word distributional statistics and dictionary definitions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 552-561). USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D11-1051.pdf

Guo Y., Barnes S. J. & Jia Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent Dirichlet allocation. *Tourism Management,* 59(7), 467-483. https://doi.org/10.1016/j.tourman.2016.09.009

Hajjem, M. & Latiri, C. (2017). Combining IR and LDA topic modeling for filtering microblogs. *Procedia Computer Science*, 112, 761-770. https://doi.org/10.1016/j.procs.2017.08.166

Hall, D., Jurafsky, D. & Manning, C. D. (2008). Studying the history of ideas using topic models. *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 363–371). Honolulu, Hawaii. Association for Computational Linguistics. Retrieved from https://aclanthology.org/D08-1038.pdf

Hannigan, T. R., Haans, R. F., Vakili, K., Tchalian, H., Glaser, V. L., Wang, M. S., Kaplan, S. & Jennings, P. D. (2019). Topic modeling in management research: Rendering new theory from textual data. *Academy of Management Annals,* 13(2), 586-632. https://doi.org/10.5465/annals.2017.0099

Henderson, K. & Eliassi-Rad, T. (2009, March). Applying latent Dirichlet allocation to group discovery in large graphs. In *Proceedings of the 2009 ACM symposium on Applied Computing* (pp. 1456-1461). https://doi.org/10.1145/1529282.1529607

Hofkirchner, W. (2009). How to achieve a unified theory of information. *TripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society*, 7(2), 357-368. Retrieved from file:///C:/Users/Reza/Downloads/jeverett,+Journal+manager,+114-389-1-LE.pdf

Hofmann, T. (1999, August). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 50-57). https://doi.org/10.1145/312624.312649

Hong, L. & Davison, B. D. (2010). Empirical study of topic modeling in Twitter. *Proceedings of the 1st Workshop on Social Media Analytics* (pp. 80–88). USA: Association for Computing Machinery. https://doi.org/10.1145/1964858.1964870

Hristova, G. (2021). Topic modeling of chat data: A case study in the banking domain. *AIP Conference Proceedings,* 2333(1), 150014. https://doi.org/10.1063/5.0044139

Hu, Y., John, A., Wang, F. & Kambhampati, S. (2021). ET-LDA: Joint topic modeling for aligning events and their Twitter Feedback. *Proceedings of the AAAI Conference on Artificial Intelligence,* 26(1), 59-65. https://doi.org/10.1609/aaai.v26i1.8106

Huang, Z., Lu, X. & Duan, H. (2013). Latent treatment pattern discovery for clinical processes. *Journal of Medical Systems*, 37(2), 9915. https://doi.org/10.1007/s10916-012-9915-2

Ibrahim, N. F. & Wang, X. (2019). A text analytics approach for online retailing service improvement: Evidence from Twitter. *Decision Support Systems*, 121, 37-50. https://doi.org/10.1016/j.dss.2019.03.002

Inaam ul Haq, M., Li, Q. & Hou, J. (2022). Analyzing the research trends of IoT using topic modeling. *The Computer Journal*, 65(10), 2589-2609. https://doi.org/10.1093/comjnl/bxab091

Jiang, B., Li, Z., Chen, H. & Cohn, A. G. (2018). Latent topic text representation learning on statistical manifolds. *IEEE Transactions on Neural Networks and Learning Systems*, 29(11), 5643-5654. https://doi.org/10.1109/TNNLS.2018.2808332

Jiang, Z., Zhou, X., Zhang, X. & Chen, S. (2012, October). Using the link-topic model to analyze regularities between traditional Chinese medicine clinical symptoms and herbs. In *2012, the IEEE 14th International Conference on e-Health Networking, Applications and Services (Healthcom)* (pp. 15-18). IEEE. https://doi.org/10.1109/HealthCom.2012.6380057

Jo, Y. & Oh, A. H. (2011, February). Aspect and sentiment unification model for online review analysis. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* (pp. 815-824). https://doi.org/10.1145/1935826.1935932

Joo, S., Choi, I. & Choi, N. (2018). Topic analysis of the research domain in knowledge organization: A latent Dirichlet allocation approach. *Knowledge Organization*, 45(2), 170–183. https://doi.org/10.5771/0943-7444-2018-2-170

Kavvadias, S., Drosatos, G. & Kaldoudi, E. (2020). Supporting topic modeling and trends analysis in biomedical literature. *Journal of Biomedical Informatics,* 110, 103574. https://doi.org/10.1016/j.jbi.2020.103574

Kherwa, P. & Bansal, P. (2019). Topic modeling: A comprehensive review. *EAI Endorsed Transactions on Scalable Information Systems*, 7(24), e2, 1-16. https://doi.org/10.4108/eai.13-7-2018.159623

Kherwa, P. & Bansal, P. (2021). A Comparative Empirical Evaluation of Topic Modeling Techniques. In: Gupta, D., Khanna, A., Bhattacharyya, S., Hassanien, A.E., Anand, S., Jaiswal, A. (eds) *International Conference on Innovative Computing and Communications. Advances in Intelligent Systems and Computing*, vol 1166. Springer, Singapore. https://doi.org/10.1007/978-981-15-5148-2_26

Kim, Y. & Shim, K. (2014). TWILITE: A recommendation system for Twitter using a probabilistic model based on latent Dirichlet allocation. *Information Systems*, 42, 59-77. https://doi.org/10.1016/j.is.2013.11.003

Koltcov, S. & Ignatenko, V. (2020). Renormalization analysis of topic models. *Entropy,* 22(5), 556. https://doi.org/10.3390/e22050556

Kurata, K., Miyata, Y., Ishita, E., Yamamoto, M., Yang, F. & Iwase, A. (2018). Analyzing library and information science full-text articles using a topic modeling approach. *Proceedings of the Association for Information Science and Technology*, 55(1), 847-848.Hoboken, NJ: Wiley. https://doi.org/10.1002/pra2.2018.14505501143

Kushkowski, J. D., Shrader, C. B., Anderson, M. H., & White, R. E. (2020). Information flows and topic modeling in corporate governance. *Journal of Documentation*, 76(6), 1313-1339. https://doi.org/10.1108/JD-10-2019-0207

Lamba, M. & Madhusdhuan, M. (2018). Application of topic mining and prediction modeling tools for library and information science journals. In M. R. Murali Prasad & Others (Eds.), *Library Practices in Digital Era: Festschrift in Honor of Prof. V Vishwa Mohan* (pp. 395-401). Hyderabad: B. S Publications. https://doi.org/10.5281/zenodo.1298739

Lamba, M. & Madhusudhan, M. (2019). Mapping of topics in DESIDOC Journal of Library and Information Technology, India: A study. *Scientometrics,* 120(2), 477-505. https://doi.org/10.1007/s11192-019-03137-5

Landauer, T. K., Foltz, P. W. & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes,* 25(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Li, Q., Li, S., Zhang, S., Hu, J. & Hu, J. (2019). A review of text corpus-based tourism big data mining. *Applied Sciences,* 9(16), 3300. https://doi.org/10.3390/app9163300

Lee, D. & Seung, H. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature,* 401, 788–791. https://doi.org/10.1038/44565

Li, J., Wu, N. & Feng, Z. (2018). Model-based non-Gaussian interest topic distribution for user retweeting in social networks. *Neurocomputing*, 278, 87-98. https://doi.org/10.1016/j.neucom.2017.04.078

Lin, C. X., Zhao, B., Mei, Q. & Han, J. (2010, July). Pet: A statistical model for popular events tracking in social communities. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 929-938). New York: Association for Computing Machinery. https://doi.org/10.1145/1835804.1835922

Lin, T., Tian, W., Mei, Q., & Cheng, H. (2014, April). The dual-sparse topic model: Mining focused topics and focused terms in short text. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 539-550). ACM, Seoul. https://doi.org/10.1145/2566486.2567980

Linstead, E., Lopes, C. & Baldi, P. (2008, December). An application of latent Dirichlet allocation to analyzing software evolution. *In 2008, the 7th International Conference on Machine Learning and Applications* (pp. 813-818). IEEE. USA: IEEE. https://doi.org/10.1109/ICMLA.2008.47

Linstead, E., Rigor, P., Bajracharya, S., Lopes, C. & Baldi, P. (2007, November). Mining concepts from code with probabilistic topic models. In *Proceedings of the 22nd IEEE/ACM International Conference on Automated Software Engineering* (pp. 461-464). New York: Association for Computing Machinery. https://doi.org/10.1145/1321631.1321709

Lindstedt, N. C. (2019). Structural topic modeling for social scientists: A brief case study with social movement studies literature, 2005–2017. *Social Currents,* 6(4), 307–318. https://doi.org/10.1177/2329496519846505

Liu, B., Liu, L., Tsykin, A., Goodall, G. J., Green, J. E., Zhu, M., Kim, C. H. & Li, J. (2010). Identifying functional miRNA-mRNA regulatory modules with correspondence latent Dirichlet allocation. *Bioinformatics (Oxford, England)*, 26(24), 3105–3111. https://doi.org/10.1093/bioinformatics/btq576

Liu, L., Tang, L., Dong, W., Yao, S. & Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5(1), 1608. https://doi.org/10.1186/s40064-016-3252-8

Liu, P., Jameel, S., Lam, W., Ma, B., & Meng, H. (2015, January). Topic modeling for conference analytics. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (Vol. 2015, pp. 707-711). https://doi.org/10.21437/Interspeech.2015-245

Liu, S., Zhang, R. Y. & Kishimoto, T. (2021). Analysis and prospect of clinical psychology based on topic models: Hot research topics and scientific trends in the latest decades. *Psychology, Health & Medicine*, 26(4), 395-407. https://doi.org/10.1080/13548506.2020.1738019

Liu, Y., Zhang, L., Nie, L., Yan, Y. & Rosenblum, D. (2016, February). Fortune teller: predicting your career path. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 30, No. 1). https://doi.org/10.1609/aaai.v30i1.9969

Lochbaum, K. E. & Streeter, L. A. (1989). Comparing and combining the effectiveness of latent semantic indexing and the ordinary vector space model for information retrieval. *Information Processing & Management*, 25(6), 665–676. https://doi.org/10.1016/0306-4573(89)90100-3

Lozano, M. G., Schreiber, J. & Brynielsson, J. (2017). Tracking geographical locations usinga geo-aware topic model for analyzing social media data. *Decision Support Systems*, 99, 18-29. https://doi.org/10.1016/j.dss.2017.05.006

Lu, H. M. & Lee, C. H. (2015). A Twitter hashtag recommendation model that accommodates temporal clustering effects. *IEEE Intelligent Systems*, 30(3), 18-25. http://dx.doi.org/10.1109/MIS.2015.20

Lu, K. & Wolfram, D. (2012). Measuring author research relatedness: A comparison of word-based, topic-based, and author co-citation approaches. J*ournal of the Association for Information Science and Technology,* 63(10), 1973-1986. https://doi.org/10.1002/asi.22628

Lu, Y., Mei, O. & Zhai, C. (2011). Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. *Information Retrieval Journal,* 14(2), 178-203. http://dx.doi.org/10.1007/s10791-010-9141-9

Lukins, S. K., Kraft, N. A. & Etzkorn, L. H. (2010). Bug localization using latent Dirichlet allocation. *Information and Software Technology*, 52(9), 972-990. https://doi.org/10.1016/j.infsof.2010.04.002

Manning, C. D., Raghavan, P. & Schütze, H. (2009). *An introduction to information retrieval*. USA: Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

Mathew, G., Agrawal, A. & Menzies, T. (2017). Trends in topics at SE conferences (1993-2013). *International Conference on Software Engineering Companion* (pp. 397-398). IEEE/ACM. https://doi.org/10.1109/ICSE-C.2017.52

McCallum, A., Corrada-Emmanuel, A. & Wang, X. (2005). Topic and role discovery in social networks. In *IJCAI'05: Proceedings of the 19th International Joint Conference on Artificial Intelligence*.

McFarland, D. A., Ramage, D., Chuang, J., Heer, J., Manning, C. D., & Jurafsky, D. (2013). Differentiating language usage through topic models. *Poetics*, 41(6), 607-625. http://dx.doi.org/10.1016/j.poetic.2013.06.004

Mifrah, S. & Benlahmar, E. H. (2020). Topic modeling coherence: A comparative study between LDA and NMF models using the COVID-19 corpus. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(4), 5756-5761. https://doi.org/10.30534/ijatcse/2020/231942020
doi: https://doi.org/10.30534/ijatcse/2020/231942020

Miner, G. D., Elder, J., Fast, A., Hill, T., Nisbet, R. & Delen, D. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press.

Miyata, Y., Ishita, E. & Yang, F. (2020). Knowledge structure transition in library and information science: Topic modeling and visualization. *Scientometrics,* 125(1), 665-687. https://doi.org/10.1007/s11192-020-03657-5

Murakami, A., Thompson, P., Hunston, S. & Vajn, D. (2017). What is this corpus about? Using topic modelling to explore a specialised corpus. *Corpora*, 12(2), 243-277. https://doi.org/10.3366/cor.2017.0118

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y. & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16), 7653-7670. https://doi.org/10.1016/j.eswa.2014.06.009

Nelson, L. K. (2020). Computational grounded theory: A methodological framework. *Sociological Methods and Research*, 49(1), 3-42. https://doi.org/10.1177/0049124117729703

Newman, D., et al. (2010). Automatic evaluation of topic coherence. *Human language technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (June, Los Angeles, California, 2010) (pp. 100-108). USA: Association for Computational Linguistics. https://dl.acm.org/doi/pdf/10.5555/1857999.1858011

Pan, Y., et al. (2014). A Biterm-based Dirichlet Process Topic Model for Short Texts. *Proceedings of the 3rd International Conference on Computer Science and Service System* (13–15 June 2014, Bangkok, Thailand). https://doi.org/10.2991/csss-14.2014.71

Paul, M. & Dredze, M. (2011). You are what you Tweet: Analyzing Twitter for public health. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1), 265-272. https://doi.org/10.1609/icwsm.v5i1.14137

Paul, M., & Girju, R. (2010). A two-dimensional topic-aspect model for discovering multi-faceted topics. *Proceedings of the AAAI Conference on Artificial Intelligence*, 24(1), 545-550. https://doi.org/10.1609/aaai.v24i1.7669

Preotiuc-Pietro, D., et al. (2017). Beyond binary labels: Political ideology prediction of twitter users. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 729–740). https://doi.org/10.18653/v1/P17-1068

Pronoza, E., Pronoza, A., & Yagunova, E. (2018). Extraction of typical client requests from bank chat logs. *17th Mexican International Conference on Artificial Intelligence* (Guadalajara, Mexico, 2018). https://doi.org/10.1007/978-3-030-04497-8_13

Qian, S., et al. (2016). Multi-modal event topic model for social event analysis. *IEEE Transactions on Multimedia*, 18(2), 233-246. https://doi.org/10.1109/TMM.2015.2510329

Rajasundari, T., Subathra, P., & Kumar P. N. (2017). Performance analysis of topic modeling algorithms for news articles. *Journal of Advanced Research in Dynamical and Control Systems*, 2017(11), 175-183.

Rao, Y. (2016). Contextual sentiment topic model for adaptive social emotion classification. *IEEE Intelligent Systems*, 31(1), 41-47. https://doi.org/10.1109/MIS.2015.91

Rao, Y., et al. (2014). Building an emotional dictionary for sentiment analysis of online news. *World Wide Web*, 17(4), 723-742. https://doi.org/10.1007/s11280-013-0221-9

Rasiwasia, N. & Vasconcelos, N. (2013). Latent Dirichlet allocation models for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11), 2665-2679. https://doi.org/10.1109/TPAMI.2013.69

Roberts, K., et al. (2012). EmpaTweet: Annotating and Detecting Emotions on Twitter. *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (May 23-25, 2012, Istanbul, Turkey) (pp. 3806–3813). Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/201_Paper.pdf

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (2–6 February, 2015, Shanghai, China) (pp. 399–408). Association for Computing Machinery: Shanghai, China. https://doi.org/10.1145/2684822.2685324

Rosa, R. L., et al. (2018). A Knowledge-Based Recommendation System that includes Sentiment Analysis and Deep Learning. *IEEE Transactions on Industrial Informatics*, 15(4), 2124-2135. http://dx.doi.org/10.1109/TII.2018.2867174

Rosen-Zvi, M., Griffiths, T., Steyvers, M. & Smyth, P. (2004). The author-topic mode for authors and documents. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (pp. 487–494). AUAI Press. https://dl.acm.org/doi/10.5555/1036843.1036902

Roy, B. K., & Mukhopadhyay, P. (2023). Theoretical backbone of library and information science: A quest. *Liber Quarterly: The Journal of the Association of European Research Libraries,* 33(1), 1-57. https://doi.org/10.53377/lq.13269

Saha, M. & Ghosh, S. (2023). Topic modelling in library and information science from the primary data: Swing in thrust areas. *International Journal of Information Science and Management,* 21(3), 19-34. https://doi.org/10.22034/ijism.2023.1977569.0 or https://dor.isc.ac/dor/20.1001.1.20088302.2023.21.3.2.6

Savage, T., et al. (2010). Topic XP: Exploring topics in source code using latent dirichlet allocation. *Proceedings of the International Conference on* Software Maintenance (September 12-18, 2010, Romania). USA: IEEE. https://www.cs.wm.edu/semeru/papers/ICSM2010_SavDit.pdf

Schofield, A., et al. (2017). Understanding text preprocessing for latent Dirichlet allocation. *Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics* (April 2017, Spain) *(*vol. 2) (pp. 432–436). USA: Association for Computational Linguistics. https://www.cs.cornell.edu/~xanda/winlp2017.pdf

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal,* 27(3), 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana, IL, USA: University of Illinois Press. Retrieved from https://pure.mpg.de/rest/items/item_2383164_3/component/file_2383163/content

Sizov, S. (2010). Geofolk: Latent spatial semantics in web 2.0 social media. *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (pp. 281-290). New York, USA. https://doi.org/10.1145/1718487.1718522

Sun X., et al. (2016). Exploring topic models in software engineering data analysis: A survey. *17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)* (May 30 -June 01, 2016) (pp. 357- 362). Shanghai: IEEE. https://doi.org/10.1109/SNPD.2016.7515925

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 Association for Information Science and Technology,* 62(1), 185-204. https://doi.org/10.1002/asi.21435

Suominen, A. & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of The Association for Information Science & Technology*, 67(10), 2464-2476. https://doi.org/10.1002/asi.23596

Tang, H., et al., (2013). A multiscale latent Dirichlet allocation model for object-oriented clustering of VHR panchromatic satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 51(3), 1680-1692. https://doi.org/10.1109/TGRS.2012.2205579

Tang, J., et al. (2014). Understanding the limiting factors of topic modeling via posterior contraction analysis. *Journal of Machine Learning Research,* 32(1), 190-198. Retrieved from https://proceedings.mlr.press/v32/tang14.html

Tian, K., Revelle, M., & Poshyvanyk, D. (2009). Using latent Dirichlet allocation for automatic categorization of software. *6th IEEE International Working Conference on Mining Software Repositories* (May 16-17, 2009, Vancouver, BC, Canada) (pp. 163-166). https://doi.org/10.1109/MSR.2009.5069496

Thomas, S. W. (2011). Mining software repositories using topic models. *Proceedings of the 33rd International Conference on Software Engineering* (pp. 1138–1139). https://doi.ieeecomputersociety.org/10.1145/1985793.1986020

Thomas, S. W., et al. (2011). Modeling the evolution of topics in source code histories. *Proceedings of the 8th Working Conference on Mining Software Repositories (*Waikiki, Honolulu, HI, USA, May 21-28, 2011) (pp. 173–182).

Titov, I., & McDonald, R. (2008). Modeling online reviews with multi-grain topic models. *Proceedings of the 17th International Conference on World Wide Web* (April 21–25, 2008, Beijing, China) (pp. 111–120). https://doi.org/10.1145/1367497.1367513

Treude, C. & Wagner, M. (2019). Predicting suitable configurations for GitHub and Stack Overflow topic models. *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR), Montreal, QC, Canada* (pp. 84-95), https://doi.org/10.1109/MSR.2019.00022

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems*, 94, 1-28. https://doi.org/10.1016/j.is.2020.101582

Venkatesaramani, R., et al. (2019). A semantic cover approach for topic modeling. *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*pp. 92–102). Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-1011

Vulic, I., De Smet, W., & Moens, M. F. (2011). Identifying word translations from comparable corpora using latent topic models. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (*short papers, Volume 2) (June 19-24, 2011, Portland, Oregon) (pp. 479–484). https://aclanthology.org/P11-2084.pdf

Wang, J., et al. (2014). Image tag refinement by regularized latent Dirichlet allocation. *Computer Vision and Image Understanding*, 124, 61-70. https://doi.org/10.1016/j.cviu.2014.02.011

Wang, T., Cai, Y., Leung, H. F., Lau, R. Y., Li, Q., & Min, H. (2014). Product aspect extraction supervised with online domain knowledge. *Knowledge-Based Systems*, 71, 86-100. https://doi.org/10.1016/j.knosys.2014.05.018

Wang, T., Huang, Z., & Gan, C. (2016). On mining latent topics from healthcare chat logs. *Journal of Biomedical Informatics*, 61, 247-259. https://doi.org/10.1016/j.jbi.2016.04.008

Wang, X., Gerber, M. S., & Brown, D. E. (2012). Automatic Crime Prediction Using Events Extracted from Twitter Posts. *Proceedings of the 5th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction* (April 2012) (pp. 231–238). https://doi.org/10.1007/978-3-642-29047-3_28

Wang, Y., & Mori, G. (2011). Max-margin latent Dirichlet allocation for image classification and annotation. *Proceedings of the British Machine Vision Conference* (September 2011) (pp. 112.1-112). http://dx.doi.org/10.5244/C.25.112

Wang, Y. C., Burke, M., & Kraut, R. E. (2013). Gender, topic, and audience response: an analysis of user-generated content on Facebook. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (April 27 – May 02, 2013, New York) (pp. 31–34). New York: ACM. https://doi.org/10.1145/2470654.2470659

Wei, X., Xin, L., & Yinhong, G. (2003). Document clustering based on non-negative matrix factorization. *Proceedings of the twenty-sixth annual international ACM SIGIR conference on Research and Development in Information Retrieval* (July 28 to August 1, 2003, Toronto, Canada) (pp. 267–273). USA: ACM Press. https://doi.org/10.1145/860435.860485

Weng, J., & Lee, B. S. (2011). Event detection in Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1), 401-408. https://doi.org/10.1609/icwsm.v5i1.14102

Willis, A., et al. (2017). Identifying domain reasoning to support computer monitoring in typed-chat problem-solving dialogues. *Journal of Computing Sciences in Colleges,* 33(2), 11-19. https://dl.acm.org/doi/10.5555/3144645.3144647

Wilson, A. T., & Chew, P. A. (2010). Term weighting schemes for latent Dirichlet allocation. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL* (June 2010, Los Angeles, California) (pp. 465–473). Retrieved from https://aclanthology.org/N10-1070.pdf

Wu, Y., Ding, Y., Wang, X. & Xu, J. (2010, July). A comparative study of topic models for topic clustering of Chinese web news. In the *3rd International Conference on Computer Science and Information Technology* (Vol. 5, pp. 236-240). IEEE. https://doi.org/10.1109/ICCSIT.2010.5564723

Wu, Y., Liu, M., Zheng, W. J., Zhao, Z., & Xu, H. (2012). Ranking gene-drug relationships in biomedical literature using latent dirichlet allocation. In *Biocomputing 2012* (pp. 422-433). https://doi.org/10.1142/9789814366496_0041

Xianghua, F., et al. (2013). Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon. *Knowledge-Based Systems*, 37, 186-195. https://doi.org/10.1016/j.knosys.2012.08.003

Xiao, C., et al. (2017). Adverse drug reaction prediction with symbolic latent dirichlet allocation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1). https://doi.org/10.1609/aaai.v31i1.10717

Xiao, Y., et al. (2019). Sense-based topic word embedding model for item recommendation. *IEEE Access,* 7, 44748-44760. https://doi.org/10.1109/ACCESS.2019.2909578

Yan, E. (2014). Research dynamics: Measuring the continuity and popularity of research topics. *Journal of Informetrics,* 8(1), 98-110. https://doi.org/10.1016/j.joi.2013.10.010

Yan, E. (2015). Research dynamics, impact, and dissemination: A topic-level analysis. *Journal of the Association for Information Science and Technology,* 66(11), 2357–2372. https://doi.org/10.1002/asi.23324

Yang, M. C., & Rim, H. C. (2014). Identifying engaging Twitter content using topical analysis. *Expert Systems with Applications*, 41(9), 4330-4336. https://doi.org/10.1016/j.eswa.2013.12.051

Yau, C. K., Porter, A., Newman, N. & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics,* 100(3), 767–786. https://doi.org/10.1007/s11192-014-1321-8

Yi, X. & Allan, J. (2008). Evaluating topic models for information retrieval. In *International Conference on Information and Knowledge Management (CIKM)* (pp. 1431-1432). https://doi.org/10.1145/1458082.1458317

Yi, X. & Allan, J. (2009). A comparative study of utilizing topic models for information retrieval. In Boughanem, M., Berrut, C., Mothe, J., & Soule-Dupuy, C. (Eds.), *Advances in Information Retrieval, Proceedings of the* 31st *European Conference on Information Retrieval Research* (pp. 29-41). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-00958-7_6

Yin, Z., et al. (2011). Geographical topic discovery and comparison. *Proceedings of the 20th International Conference on World Wide Web, WWW* (March 28 - April 1, 2011, Hyderabad, India) (pp. 247-256). https://doi.org/10.1145/1963405.1963443

Yu, J., & Qiu, L. (2018). ULW-DMM: An effective topic modeling method for microblog short texts. *IEEE Access*, 7, 884-893. https://doi.org/10.1109/ACCESS.2018.2885987

Yu, R., He, X., & Liu, Y. (2015). Glad: group anomaly detection in social media analysis. *ACM Transactions on Knowledge Discovery from Data,* 10(2), 1–22. https://doi.org/10.1145/2811268

Zhai, Z., et al. (2011). Constrained LDA for grouping product features in opinion mining. *Advances in knowledge discovery and data mining: 15th Pacific-Asia Conference, PAKDD* (May 24-27, 2011, Shenzhen, China) (pp. 448-459). https://doi.org/10.1007/978-3-642-20841-6_37

Zhang, C., et al. (2016). A hybrid term–term relations analysis approach for topic detection. *Knowledge-Based Systems*, 93, 109-120. https://doi.org/10.1016/j.knosys.2015.11.006

Zhang, H., Xu, S., & Qiao, X. D. (2014). Review on topic models integrating intra- and extra-features of scientific and technical literature. *Journal of the China Society for Scientific and Technical Information*, 10, 1108–1120.

Zhang, Y., et al. (2017). iDoctor: Personalized and professionalized medical recommendations based on hybrid matrix factorization. *Future Generation Computer Systems*, 66, 30-35. https://doi.org/10.1016/j.future.2015.12.001

Zhao, F., et al. (2016). A personalized hashtag recommendation approach using an LDA-based topic model in microblog environment. *Future Generation Computer Systems*, 65, 196-206. http://dx.doi.org/10.1016/j.future.2015.10.012

Zheng, X., et al. (2014). Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification. *Knowledge-Based Systems*, 61, 29-47. https://doi.org/10.1016/j.knosys.2014.02.003

Zoghbi, S., Vulic, I., & Moens, M. F. (2016). Latent Dirichlet allocation for linking user-generated content and e-commerce data. *Information Sciences*, 367, 573-599. http://dx.doi.org/10.1016/j.ins.2016.05.047

Zou, C. (2018). Analyzing research trends on drug safety using topic modeling. *Expert Opinion on Drug Safety,* 17(6), 629-636. https://doi.org/10.1080/14740338.2018.1458838