IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 October 2023, accepted 5 November 2023, date of publication 20 November 2023,
date of current version 29 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3334619

TOPICAL REVIEW

Advancements of Artificial Intelligence Techniques in the Realm About Library and Information Subject–A Case Survey of Latent Dirichlet Allocation Method

XINGZHOU PAN AND YU XUE
Library of Shandong University, Jinan, Shandong 250100, China

Corresponding author: Yu Xue (LoanLoynd@gmx.us)

This work was supported by the "Case Study on the Application of Artificial Intelligence in Intelligent Library Service," Ministry of Education Industry-University Cooperative Education Project, under Project 220607027162401.

ABSTRACT To investigate the advancements of artificial intelligence techniques in the realm of library and information subject, we have chosen the Latent Dirichlet Allocation method as a case study to explore its current study status and implementations. Traditional theme mining analyses utilize methods such as word frequency statistics, co-occurrence analysis, community detection, and citation analysis to capture external quantitative features of words or documents. In contrast, the Latent Dirichlet Allocation theme modelling method employs a three-layer Bayesian structure of document-topic-word to describe the themes of documents and the semantic relationships among words, enabling a better exploration of latent semantic information in text. This method plays a pivotal role in fine-grained knowledge extraction and analysis. We systematically review more than a decade of relevant literature in the realm about library and information subject. Through content analysis, we construct an analytical architecture for the implementation of the Latent Dirichlet Allocation method. This architecture, viewed from the perspective of the implementation process of Latent Dirichlet Allocation, comprehensively summarizes the core stages and technical challenges, including text pre-processing, model construction (i.e., theme model selection and optimal theme number determination), and model solving. Additionally, we provide a comprehensive overview of the current study status of the Latent Dirichlet Allocation method across various implementation domains, such as theme exploration, knowledge organization, academic evaluation, sentiment analysis, and recommendation study. Our findings indicate that the Latent Dirichlet Allocation method has formed a mature analytical process in the realm of library and information subject, with ongoing growth in study interest.

INDEX TERMS Library and information subject, artificial intelligence, latent Dirichlet allocation method.

I. INTRODUCTION

Latent Dirichlet Allocation, abbreviated as LDA, is a three-layer Bayesian generative probabilistic model used for addressing the problem of topic modelling in text data. Originally proposed by Blei, Ng, and Jordan [1], the LDA method is an unsupervised machine learning algorithm primarily employed in the domains of textual mining and natural language processing (NLP). The fundamental concept behind the LDA method is to view text data as a mixture of multiple topics, with each topic composed of various words. The model's objective is to infer the document's topic distribution and the word distribution of each topic by analyzing the distribution of words within the document. The LDA method assumes that each document has a certain probability distribution over topics, and each topic has a certain probability distribution over words. Through iterative algorithms, these probability distributions can be gradually adjusted to

The associate editor coordinating the review of this manuscript and approving it for publication was Ali Kashif Bashir.

obtain the topic distribution of each document and the word distribution of each topic. Due to its strong dimensionality reduction capabilities and model scalability, the LDA method finds extensive applications in the domain of textual mining, including tasks such as text topic analysis, text classification, and information retrieval. Through the LDA method, researchers can unearth the latent topic structure hidden within large-scale text data, aiding in the comprehension of the content and underlying relationships within the textual information [2].

Currently, research in the domain of computer science mainly revolves around the classification and algorithmic improvements of the extended LDA method [3], particularly delving into discussions regarding deep learning-based topic models [4]. Relevant studies emphasize the significant role of topic models in textual mining and NLP, focusing on the principles, summarizing parameter estimation and training methods, and emphasizing the performance comparison among different topic modelling techniques, of the LDA topic model [5]. A minority of studies have summarized the applications of topic models in various disciplines such as linguistics, politics, biomedical sciences, and geography, pointing out the challenges and issues present in LDA topic models within multimedia information processing and other textual mining tasks [2]. Scholars in the domain of library and information science have predominantly conducted analytical research tailored to specific textual mining tasks [6]. Some existing reviews have summarized applications within certain scenarios [7] or specific categories of extended models [8]. Overall, existing research lacks a comprehensive overview of the complete application process of the LDA method. Inspired by this, our contributions are summarized as follows:

1) We conduct in-depth reviews of representative literature and establish an analytical architecture for the application research of the LDA method, based on key stages of the application process. Proposed architecture elucidates critical stages in the application process, guiding researchers in executing tasks such as text pre-processing, model construction, and solving, while underscoring the significance of topic model selection and determination of the optimal number of topics. This architecture provides systematic guidance for the application of artificial intelligence technology in the domain of library and information science, enabling a more profound exploration of its practical impact and potential.

2) We focus on the overall research status of the LDA method in the library and information subject and meticulously analyze its application process and domains, aiming to offer references for theoretical research and practical application.

3) We review existing issues and innovations in the application of the LDA method, aiding in better addressing complex textual processing tasks in multi-dimensional scenarios. This, in turn, enhances the generalization ability of the LDA model and improves the accuracy, interpretability, and precision of modeling results, facilitating more accurate topic mining and identification.

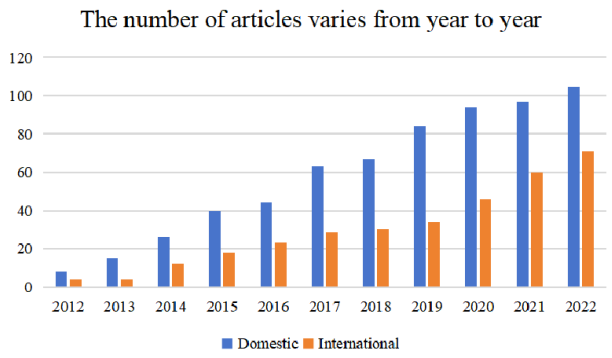

FIGURE 1. Statistics of library and information research.

4) We effectively address the current knowledge gap in comprehending the complete application process of the LDA method. This study not only showcases the potential of artificial intelligence technologies like LDA but also paves the path for future research, innovation, and practical application in the context of libraries and information science. It provides practical guidance for researchers aspiring to apply artificial intelligence technology in this domain.

II. OVERVIEW

This study conducted literature retrieval using Web of Science Core Collection, Library and Information Science Abstracts (LISA), and Google Scholar as English data sources, along with CNKI, Wanfang, and VIP databases as Chinese data sources. The English literature search was performed using the search query Latent Dirichlet Allocation OR Topic Model*, while the Chinese literature search employed the query LDA OR Latent Dirichlet Allocation (In Chinese) OR Probabilistic Topic Model (In Chinese). Both Chinese and English data sources were limited to journals in the domain of library and information science (with Chinese sources further limited to CSSCI-indexed journals). The search timeframe was set to the about past decade (from January 1, 2012, to December 31, 2022). Based on titles, keywords, abstracts, and content analysis, relevant literature aligned with the research topic was selected. This process resulted in the identification of 369 English articles and 426 Chinese articles. The distribution of literature is illustrated in Figure 1, depicting a consistent growth trend in domestic and international publications over the past decade, with domestic research significantly surpassing international output. It is evident that LDA-related research has garnered considerable attention from scholars in the domain of library and information science, yielding prolific research outcomes.

Furthermore, this study conducted an in-depth examination of representative literature and constructed an analysis architecture for LDA method applications based on key stages in the application process, as shown in Figure 2.

Specifically, the application process of the LDA method encompasses several stages:

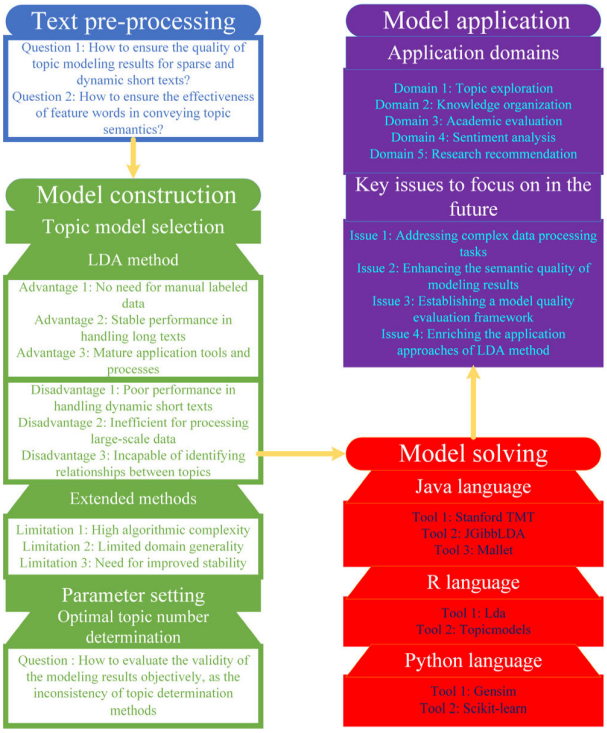

FIGURE 2. **The constructed analysis architecture about applications of LDA.**

1) In the text pre-processing stage, data sources relevant to topic modelling are processed to obtain formatted data required by the model.

2) In the model construction stage, an appropriate topic model is selected based on the research context, followed by determining the optimal number of topics through relevant model evaluation methods.

3) In the model solving stage, suitable topic modelling tools are employed to complete the model solving.

4) Finally, based on the actual research context, relevant methods and tools are combined to address questions in the specific application domain [7].

Currently, the application scope of the LDA method spans various domains, including topic exploration, knowledge organization, scholarly evaluation, sentiment analysis, and recommendation research. This paper summarizes the existing state of application research based on the aforementioned process.

III. LDA TOPIC MODELLING APPLICATION PROCESS

A. TEXT PRE-PROCESSING

The process of text pre-processing involves various techniques such as tokenization, stop word removal, and feature selection, aiming to obtain the formatted data required by the model. This step facilitates preliminary dimension reduction of document content, reduces model inference time, and forms the foundation of LDA topic modelling. The features obtained from the text that express the semantic essence of topics have a significant impact on the interpretability of topic modelling results [8].

Regarding the data sources for topic modelling, recent years have witnessed valuable research in the domain of library and information science focusing on topic mining from short text data on social media platforms like Weibo and Twitter. However, due to the limited length, high sparsity, rapid updates, and large scale of short texts, the LDA method's performance is relatively poor in handling short texts [9]. Additionally, these types of data often contain slang, abbreviations, and emoticons, resulting in relatively vague text semantics and weak logical relationships. As a result, pre-processed texts may express fewer topic-related terms, making the extraction of meaningful topics more complex [10]. The challenge lies in ensuring the quality of topic modelling results for sparse and dynamic short texts. Scholars address this by constructing pseudo-documents for information integration to increase text length or utilizing improved extended models that adjust model assumptions and enhance topic generation processes [11]. Examples include the Dirichlet Multinomial Mixture (DMM) model, which caters to the characteristic of shorter text terms compared to longer text documents by strictly limiting the number of topics in the model assumptions; and the Biterm Topic Model (BTM), which mines local linguistic information from word pairs to enhance the comprehensiveness and accuracy of short text topic mining.

After determining the data source for topic modelling, specific text pre-processing techniques are employed to obtain the formatted data required by the model. Thanks to mature tokenization tools [7], [12] and stop word lists, Chinese text data has established an integrated application process for tokenization and stop word removal. In contrast, English text data typically undergoes unigram tokenization using spaces as delimiters. However, this practice may result in limited semantic content for individual words and subsequently poor interpretability of modelling results. Some research suggests phrase-based tokenization [13] or the introduction of phrase generation algorithms [11] to improve the topic representation capacity of feature words.

Feature selection further reduces the dimensionality of modelling materials after tokenization and stop word removal. Common methods include evaluation functions [14] (such as term frequency, information gain, etc.), domain ontologies [15], part-of-speech filtering [16], and regular expressions [17]. Different feature selection methods exhibit varying performance in feature word dimension reduction, which also affects the interpretability of modelling results. For instance, domain ontology-based vocabulary filtering method can effectively enhance the domain relevance of feature words and improve the interpretability of modelling results within professional contexts [15]. However, this approach relies heavily on domain knowledge and involves high manual effort. Part-of-speech filtering [16] can extract nouns and verbs that contribute significantly to topics. While it enables more convenient and efficient text dimension reduction compared to ontology-based methods, its relevance to specific domains is relatively

TABLE 1. Representative LDA methods for different library and information science domains.

Model type	Model name	Proposed time	Model description	Application example
Traditional topic model	LDA[24]	2003	The first fully meaningful generative probabilistic topic model, suitable for handling static long text data	Core topic analysis of research literature related to open government data and information freedom
Non-parametric topic model	HDP (Hierarchical Dirichlet Process) [25]	2006	Automatically determines the number of topics, overcoming the subjectivity and randomness of manually determining the number of topics	Detection of sub-events in Twitter during the Ferguson riots
Dynamic topic model	DTM (Dynamic Topic Model) [26]	2006	Introduces time information to dynamically track the evolution of topics over time	Analysis of Reddit users' smartwatch-related interests and evolution trends
Correlated topic model	CTM (Correlated Topic Model) [27]	2007	Describes the correlation between topics through covariance matrix, addressing the issue of unrelated results in LDA model	Identification and correlation analysis of environmental science literature topics
Supervised topic model	Labeled-LDA [28]	2009	Introduces label information to control the number of topics and improve the authenticity and effectiveness of modeling results	Sentiment analysis of product attributes based on online comments
Structural topic model	STM (Structural Topic Model) [29]	2014	Adds more covariates (such as document-level metadata) when computing topic-word conditional distributions to enhance model inference ability	Identification of topics in information management-related research literature
Author topic model	ATM (Author Topic Model) [30]	2004	Introduces author information to identify authors' research topics and mine content features from multiple angles	Collaborative research across topics among genomics researchers
Sentimental topic model	JST (Joint Sentiment Topic Model) [31]	2009	Constructs an additional sentiment layer between documents and topics to simultaneously detect topics and sentiment-related information	Detection of tablet computer online reviews' topic and sentiment-related information
Short text topic model	BTM(Bi-Term Topic Model) [32]	2013	Models topic modeling for co-occurring word pairs in the corpus to address document sparsity	Identification of topics in tweets related to natural disasters, i.e., Typhoon Haiyan
Word vector topic model	TWE (Topical Word Embeddings) [33]	2015	Constructs topic word embeddings based on LDA modeling, better revealing latent associations between topics	Analysis of topic evolution in medical science reports
Neural network topic model	NTM (Neural Topic Model) [34]	2015	Describes document-topic and topic-word distributions from the perspective of a feed-forward neural network, learning model parameters through back-propagation, with a simple structure	Construction of human-like dialogue systems
Joint training topic model	LDA2vec [35]	2016	Utilizes Word2Vec model to introduce contextual	Personalized news recommendation research

TABLE 1. *(Continued.)* **Representative LDA methods for different library and information science domains.**

			relationships among words for modeling, enhancing the identification of text's latent topic semantics	
Multilingual author topic model	JointAT (Joint Author Topic Model) [36]	2020	Simultaneously introduces author and multilingual information to enhance accuracy in modeling authors' research interests in multilingual contexts	Identification of research interests among scientists in the field of information science using a multilingual dataset

lower, resulting in weaker interpretability within specialized contexts.

In summary, tokenization, stop word removal, and feature selection have matured into well-established processes. Current application research often adheres to existing text pre-processing tools and semantic resources, focusing on simple combinations of single methods or a few methods. It's noteworthy that the choice of tokenization algorithm [18], construction of domain-specific terms [19] and stop word lists [20], as well as the effectiveness of different feature selection methods in expressing topic semantics, all have varying degrees of impact. Therefore, in-depth exploration is required for different application scenarios to enhance the quality of text pre-processing.

B. TOPIC MODEL CONSTRUCTION

After obtaining the formatted data required by the topic model through text pre-processing, the process enters the phase of model construction and solving. The first step is to select or construct an appropriate topic model based on the data characteristics and research context. Subsequently, the optimal number of topics is determined through model evaluation methods. Finally, an appropriate topic modelling tool is chosen or developed to perform automatic parameter estimation and complete model solving.

1) TOPIC MODEL SELECTION

The initial step in topic model construction is selecting an appropriate topic model. The LDA model is a widely used three-layer Bayesian probabilistic topic model based on the bag-of-words model. It does not require manually annotated data during training, making it effective for discovering latent semantics in longer texts like scientific literature [21]. LDA method has a mature toolset and workflow, and is commonly used for topic modelling. However, as the objects of data processing and the tasks of text analysis become increasingly complex and diverse, the naive LDA method has limitations. For instance, it is not very effective in handling dynamic short texts and can be slow in training on large datasets [22]. Furthermore, it struggles to identify relationships between topics [23]. The model's generalization ability, accuracy, and interpretability of results are challenged [4]. In light of these limitations, scholars have proposed various extension models to enhance the effectiveness of topic modelling based on text characteristics and task contexts [3]. Table 1 summarizes representative studies of various LDA method extensions in the domain of library and information science.

As shown in Table 1, each type of LDA method extension has its unique characteristics. Based on the characteristics of model improvements and existing LDA extensions [3], [7], these extension models can be broadly classified into six categories: 1) Bayesian non-parametric model, i.e., HDP; 2) time-based extension model, i.e., DTM; 3) parameter-based extension model, i.e., CTM; 4) supervised model, i.e., Labeled-LDA; 5) extension model STM based on document metadata; and 6) task-specific extension models such as ATM, JST, BTM, TWE, NTM, LDA2vec, and JointAT. Different extension models cater to diverse application scenarios, meeting the varying modelling needs of researchers. For static longer texts, the traditional LDA method is suitable. However, for data with distinct dynamic, subjective, or sparse features, dynamic topic models or sentiment topic models specifically tailored to certain tasks can enhance modelling accuracy. In the realm of LDA extension model application research, early extensions based on non-parametric and correlation-based models are still widely used due to their good performance.

With the emergence of novel application scenarios, the diversification of topic mining tasks has made task-specific extension models a prominent trend. These models include those based on word embeddings, multilingual author information, mixed contexts, and more [37]. However, these extension models still have limitations, involving numerous latent variables and additional information, resulting in higher algorithmic complexity. They are also influenced by training corpus and task context, which affects their domain generality and model stability.

It's important to note that topic models based on deep learning principles and methods have become a significant branch of LDA extension model research. Compared to other extension models, deep learning topic models combine techniques such as word embeddings and neural networks to fully explore contextual information and relationships between words, resulting in a stronger ability to understand topic semantics and higher interpretability of modelling results [2]. Currently, deep learning topic models fall into three categories: word embedding-assisted probabilistic topic models, neural network-based topic models, and jointly trained topic

models [4]. Word embedding-based models characterize word semantic similarity by training low-dimensional dense word embeddings, which enhances semantic consistency of topic words when applied to short texts and domain-specific texts, such as the Gaussian LDA method based on Gaussian distribution [38]. Neural network-based topic models often utilize the bag-of-words as model inputs and add additional layers to capture semantic relationships between words. They also incorporate sparse constraints to address the sparsity of the topic-word distribution and enhance the quality of topic model generation. Jointly trained topic models combine the strengths of probabilistic topic models and neural language models. They not only discover global semantic relationships between documents, topics, and words, but also identify dependencies between word sequences at the sentence level, overcoming the limitations of the bag-of-words assumption. These deep learning extension models generally outperform traditional models in short texts and domain-specific texts, offering richer functionality. However, they usually require support from large-scale corpora and involve more complex training processes, often requiring parameter tuning. Aside from word embedding-based models, further exploration is needed for the other two categories of deep learning extension models.

2) OPTIMAL TOPICS NUMBER SELECTION

Once the most suitable topic model has been chosen based on the research context, parameter estimation and setting must be conducted. Parameter estimation is used to infer the distributions of document-topic and topic-word pairs. Currently, there are multiple approximate inference algorithms available for parameter estimation [7]. Parameter setting is closely related to model performance and can imbue the model with specific attributes [39], involving Dirichlet prior α, β parameters, and the number of topics. α and β are typically set based on empirical values.

The selection of the optimal number of topics is based on quality evaluation methods of topic models, and it's a long-standing challenge. Too many topics can lead to small topic generalization scope and minor semantic content differences, making topic division difficult. Conversely, too few topics can result in overly broad semantic content generalization, neglecting smaller topics. The selection of the optimal number of topics directly affects the accuracy and interpretability of LDA topic modelling results. Currently, researchers often estimate the initial number of topics contained in documents based on prior knowledge and then select the optimal number of topics using quality evaluation methods like perplexity, coherence, and topic similarity. This paper provides a summarized comparison of the core ideas and advantages and disadvantages of several typical methods for determining the optimal number of topics, as shown in Table 2.

Table 2 reveals a rich variety of methods for determining the number of topics in LDA methods. The evaluation perspectives vary significantly, resulting in different methodologies. As a result, a consistent set of evaluation criteria for topic modelling results has not yet been established, and the problem of objectively assessing the effectiveness of modelling results remains unresolved. Perplexity remains a common choice in practical applications. However, some research suggests that coherence is the most effective method for measuring topic quality [46], with increased usage of this metric in recent studies. Despite the guidance provided by the above model evaluation methods, issues such as mixed topics, illogical topics, and indistinguishable topics can still arise. To further ensure the effectiveness of modelling results, researchers are starting to improve traditional evaluation methods [42], proposing new metrics [40], prioritizing interpretability in model evaluation [47], and introducing expert opinion metrics such as homogeneity, completeness, and V-Measure [48] to guarantee the quality and reliability of topic generation. Some scholars also suggest the combined application of relevant methods and the establishment of evaluation mechanisms during model operation to dynamically adjust the optimal number of topics [10], thereby enhancing the flexibility of topic number selection. Additionally, due to LDA method's parameter estimation based on random sampling and its sensitivity to the characteristics of the modelling corpus, resulting in poor stability of modelling results, some studies have introduced new stability analysis algorithms [49] and model quality evaluation indicators such as robustness and descriptiveness [50] to determine the optimal number of topics and ensure the predictive ability and reliability of the topic model.

C. MODEL SOLVING

After selecting the optimal number of topics, the next step is to choose or develop a suitable topic modelling tool to complete model solving. Various open-source LDA modelling tools have been developed for automated parameter estimation, and seven of them are commonly used in the domain of library and information science. These tools include the Stanford TMT (Topic Modelling Toolbox) [51], JGibbLDA [52], Mallet [53], Lda library [54] and topicmodels library [55] based on Java and R languages respectively, as well as the Gensim library and Scikit-learn library based on Python. By employing these tools to perform model solving, parameters are estimated for document-topic and topic-word distributions [56]. Then, topic naming is carried out through topic word filtering to uncover latent topics within each document.

IV. APPLICATION DOMAINS OF LDA TOPIC METHOD

The LDA method effectively extracts latent semantic information from text, and it has been widely applied in various domains such as topic exploration, knowledge organization, academic evaluation, sentiment analysis, and recommendation research.

A. TOPIC EXPLORATION

Topic exploration research primarily includes topic discovery and evolution analysis [6]. The LDA method possesses

TABLE 2. Introduction of different topic number determination methods.

Method	Core idea	Proposed time	Advantages	Disadvantages
Perplexity [41]	Reciprocal of the geometric mean of the probabilities generated by each vocabulary word in a document. lower perplexity indicates better predictive ability on new data	2018	Measures model's predictive power on new data	Stability of topic number selection based on perplexity is poor and often biased towards larger values [40], leading to vague semantic interpretation of extracted topics
Coherence [29]	Higher coherence indicates tighter semantic relatedness among words in a topic, implying better model interpretability	2021	Measures interpretability of topics	Ineffective for low-frequency topic words and inability to differentiate high-frequency words from information words representing the topic [42]
Topic inter similarity [44]	Model is optimal when the average similarity between topics is minimized	2017	Measures stability of topic structure	Subjectivity in selecting and constructing similarity measurement indicators [43]
Empirical approach [45]	Refers to past literature or practical experience, iteratively experiment and observe the effects of topic clustering, and make judgments manually	2020	Simple and easy to use, higher controllability with human supervision	High subjectivity, high time and labor costs

strong dimensionality reduction capabilities, allowing it to extract latent semantics from large-scale text data through unsupervised means. This ensures the relative objectivity and efficiency of topic extraction, making it a popular tool in the domain of topic exploration. Based on content analysis of relevant literature, the current topic exploration in the domain about library and information science is mainly oriented towards scientific literature, user-generated content represented by online consumer platform reviews and online public opinion data, as well as mining and analysis of web information resources such as news reports and policy texts.

1) TOPIC EXPLORATION OF SCIENTIFIC LITERATURE

Scientific literature is a crucial carrier for the dissemination of scientific and technological information, including scientific journals, conference papers, patents, and technical reports [7]. Early topic discovery in scientific literature relied on traditional quantitative methods such as word frequency statistics, co-word analysis, and citation analysis on a document level, focusing on external quantitative features such as keywords or citations. However, the LDA method allows for topic modelling of text content, gradually becoming one of the mainstream tools for topic exploration in scientific literature. For instance, it has been applied to analyze topics in SIGIR (Special Interest Group on Information Retrieval) conference papers [57] and research hotspots in China's ICT (Information and communication Technology) industry [21]. Overall, the problem of excessive reliance on a single LDA method in the exploration of scientific literature topics exists, with only some scholars attempting to apply new methods to further enhance the understanding of textual semantics. For example, Hui et al. [58] used journal articles and patent documents as data sources and applied the LDA2vec model to identify machine learning research hotspots. This model, based on the global modelling of the LDA method, models the local contextual information of the corpus through Word2vec word vectors, thereby mining richer latent semantics. However, classical word vector models like Word2vec usually train only one vector representation for each vocabulary, making it difficult to discover different meanings of words in different contexts. Research has introduced the TWE [33] model, which can simultaneously train vector representations for vocabulary and topics, thus learning different representations of word vectors under different topics and effectively improving the accuracy of mining medical technology report topics.

Analysis of topic evolution is based on topic discovery and involves grasping the rules of topic dynamic development. In topic evolution analysis, improving the accuracy of the analysis of topic evolution paths has been a research hotspot. For instance, in the domain of artificial intelligence journal paper topic extraction, Citation Involved Hierarchical

Dirichlet Process (CIHDP) [59] utilizes citation information to enhance document text representation, determining the number of topics for each period automatically, and identifying more detailed and complete path splits and fusion information. In the case of graphene patent documents, research has been conducted to measure the development degree of topics based on novelty, attention, and topic structure indicators, building on the LDA method [41]. Furthermore, researchers have used time series features in the topic evolution process to improve the accuracy of topic evolution analysis. For example, by using journal papers in the domain of library and information science as data sources, LSTM (Long Short-Term Memory) [13] is applied to model the time series features of the heat evolution of subject topics based on time slices after extracting subject topics using the LDA method. This can effectively improve the accuracy of predicting future trends in the heat of subject topics.

2) TOPIC EXPLORATION OF USER-GENERATED CONTENT

In research related to topic discovery in user-generated content, one approach involves using user comments in the online consumer platform as data sources, with the goal of extracting user opinions on products or services. For instance, Opinion LDA [60] improves document structure by converting word sequences based on user comment content into product feature word sequences based on user viewpoints, effectively identifying user preferences for specific product features. Another approach uses network public opinion data as data sources, with the aim of rumor control. For example, based on LDA analysis of Weibo text topic characteristics, combined with the random forest algorithm, significantly enhancing rumor identification accuracy [61]. However, both types of data usually consist of short texts, with fewer characteristic words reflecting the topic content, making it difficult to fully mine semantic information using the LDA method for topic discovery. To address these issues, commonly used methods primarily involve information integration to increase text length or using topic models more suitable for short texts. Some studies have attempted to combine the LDA model with other methods. For example, after obtaining modelling results for academic app reviews, Glove word vectors are used to calculate word similarity to expand feature words under topics, thereby improving the distinction between topics and mining deeper systematic topic information [62].

Topic evolution analysis of user-generated content focuses on changes in topic trends, which is of practical significance to companies, governments, and other organizations. For user-generated content on online consumer platforms, topic evolution analysis can uncover user focusing on products and services at different time nodes, helping companies improve product and service quality [63]. Topic evolution analysis based on network public opinion data can assist relevant departments in responding to public emergencies [64]. However, social media platforms are dynamic and complex opinion domains, the effectiveness of public opinion control depends on the discovery of key nodes and hot topics during the evolution [25]. Scholars have used the theory of hyper-networks as the basis, identifying Weibo topic sub-nets using the LDA method, then combining corresponding social, content, and emotion sub-nets to construct a Weibo public opinion hyper-network. The HyperEdgeRank algorithm identifies key figures, comprehensively mining key nodes in Weibo public opinion dissemination [65], effectively serving public opinion supervision in social media.

3) TOPIC EXPLORATION OF OTHER WEB INFORMATION RESOURCES

The exploration of topics within web resources, such as news reports and policy texts, holds significant potential for informing corporate and governmental decisions, as well as supporting the tracking of research hotspots. In the realm of news reports [66], due to the issue of imbalanced textual data, some studies [67] have combined feature detection methods (including independence detection, variance detection, and information entropy detection) to enhance the thematic representation capacity of feature words, thereby substantially improving the accuracy of text-based topic identification. Policy texts encompass documents arising from policy-related activities, encompassing official documentation, public document archives, and policy-related sentiment texts [68]. Given the substantial variation in the connotations of policy terms across different contexts, the LDA method addresses the challenge of semantic mining in text clustering by exploiting relationships between text, topics, and words. Consequently, it has been widely applied to the discovery of topics within policy texts, spanning domains such as climate [69] and government open data [70]. Scholars have further leveraged the LDA2vec model to enhance the comprehensiveness of semantic content extraction within policy texts [12]. Furthermore, labels, the terms used to classify or describe web information resources, have been successfully generated using the LDA method in contexts like Weibo [71] and online healthcare [72]. Some studies have extended these models for tailored label generation in specific domains, such as the sureLDA (Surrogate-guided ensemble Latent Dirichlet Allocation) method for generating phenotype labels from electronic health record data, thus expanding the utility of the LDA method [73].

In the analysis of topic evolution within news reports, current research often builds upon the foundation of the LDA method, incorporating other methods to enhance the accuracy of evolution analysis [74]. For instance, the introduction of manifold learning can offer a global temporal perspective to reconstruct relationships between news topics, mitigating issues arising from the use of adjacent time windows [75]. Additionally, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithms can be used to filter out noise text, ensuring the purity of topic extraction via the LDA method and thereby improving the accuracy of topic evolution analysis [76].

Regarding the analysis of topic evolution within policy texts, the integration of the LDA method with algorithms

measuring topic similarity and strength has been applied to quantitative analyses of policy texts in domains such as artificial intelligence [77] and regional technological innovation [78]. This integration provides valuable support for policy formulation and enhancement. Certain studies have incorporated extended models, like the ToT (Topic over Time) model [79], which introduces time factors to derive topic distribution strength across different time slices, thereby bypassing the need for intricate topic alignment procedures.

B. KNOWLEDGE ORGANIZATION

The LDA method, by describing information resources' content using topics and topic words as units through unsupervised means, facilitates the refinement of analytical units from documents to topic words. As such, it has been widely adopted in research pertaining to knowledge organization. Knowledge organization is centered around the semantic information within text, emphasizing the inter-connectedness between different pieces of knowledge. The LDA method, through its ability to mine latent topic features, constructs relationships between documents and feature words, which in turn enables knowledge inference. Consequently, it has found application in constructing knowledge graphs and topic maps by scholars in the domain of library and information science.

1) KNOWLEDGE GRAPH CONSTRUCTION

Contemporary knowledge graph construction typically involves methods such as named entity recognition and template matching, which yield comprehensive entity and relationship extractions from specialized domain corpora. However, when the content of a corpus spans diverse topics, relying solely on methods like named entity recognition to extract local information as entities might lead to semantic gaps and other issues [80]. By utilizing the LDA method to treat text topics extracted from global information as entities, a finer representation of knowledge graph can be achieved. This process allows for the structured organization of text data with complex and weakly related topics, enhancing the relationships between entities and ultimately improving the effectiveness of knowledge inference. For instance, in the construction of a knowledge graph within the domain of electronic government, Sun et al. [80] utilized the LDA method to obtain topic entities, thereby augmenting the domain-specific entities related to electronic government. This approach addresses the issue of semantic gaps within entity extraction algorithms, facilitating more effective knowledge inference and governmental decision-making. Apart from employing topics as entities in knowledge graph construction, joint utilization with topic words is also possible. For example, Zhou et al. [81] first identified topics within the medical and health information domain through the LDA method. They then employed social network analysis to uncover core topic words, ultimately constructing a knowledge graph of the medical and health information domain based on the co-occurrence relationships of core topic words. This approach aids in the analysis of domain-related knowledge associations.

2) TOPIC MAP CONSTRUCTION

A topic map serves as a knowledge repository, storing topics along with their logical relationships and hierarchical structures [82]. Unlike knowledge graphs, topic maps are more suitable for organizing networked information resources that are unordered, unstructured, and characterized by divergent topics. When handling unstructured textual information, traditional topic clustering methods like co-occurrence analysis are susceptible to the influence of word frequency and the complexity of the text domain. Consequently, these methods struggle to explain the semantic relationships between documents and vocabulary terms, leading to challenges in achieving high-quality topic identification. The LDA method, however, excels in the modelling of unstructured textual information, describing latent document features and semantic relationships between vocabulary terms. This capability helps mitigate issues related to word mismatch (i.e., synonymy and polysemy) to some extent. In comparison to traditional topic clustering methods, the LDA method more effectively mines semantic information within text, making it widely applicable in topic map construction. For instance, it has been applied to construct topic maps of clinical medical course knowledge [83]. Furthermore, topic maps find utility in sentiment analysis [84]. In the context of network sentiment control based on Weibo information, the LDA method is employed to cluster user comment and retweet texts into topics. This process allows for the construction of user topic maps where common topics serve as nodes and the similarity of topic distributions serves as edges [85]. Alternatively, users can be treated as nodes, and retweet relationships as edges, constructing user topic maps [86]. This approach enables the identification of key topics and key users during sentiment evolution, aiding regulatory authorities in achieving precise sentiment responses.

Despite the LDA method's proficiency in unstructured textual information for topic map construction, attention must be given to enhancing the relevance of topic words within specific domain contexts and ensuring the timeliness of topic maps, thereby elevating their application value.

C. ACADEMIC EVALUATION

Academic evaluation encompasses aspects such as the assessment of document influence and author impact. Existing quantitative academic evaluation methods primarily integrate traditional bibliometric indicators, such as citation frequency, network-based metrics like PageRank [87], H-index [88], and Altmetrics evaluation metrics [89]. However, due to variations in the topics of document content and author research domains [53], these indicators struggle to effectively reflect the actual influence of documents and authors within specific research topics. Consequently, research has begun to assess documents and authors on a finer granularity, focusing on

topics as the basis for evaluation. Classifying documents or authors based on their research topics is a pivotal step in topic-oriented academic evaluation. The LDA method, through probabilistic inference, deduces the document-topic distribution parameters, objectively categorizing documents into several topic categories. By subsequently mapping relationships between documents and authors, it achieves classification of authors' research topics, subsequently combining other metrics for impact calculation [90]. Certain studies directly employ pertinent influence evaluation models, such as the Collective topic PageRank Model (CTPM) [91]. This model, by identifying document topics and inter-topic correlations, introduces metadata like citation counts and journal impact factors to effectively reflect document influence within specific topics. In terms of author impact evaluation, the ATM model [92] has been utilized to achieve more precise categorization of authors' research topics. In the face of continual subdivision within the realm of scientific research, the LDA method can unearth research topics of both documents and authors, addressing the shortcoming of neglecting content information in traditional academic evaluation. As such, it advances the development of nuanced academic evaluation.

D. SENTIMENT ANALYSIS

The primary application of the LDA method lies in extracting topic information from corpora and is inadequate in discerning user sentiment underlying relevant topics. Delving into the emotional tendencies behind topics necessitates the incorporation of corresponding sentiment analysis methods or the construction of sentiment-topic models. The outcomes of such research bear significant real-world implications for enterprise and government decision-making. For instance, utilizing LDA to obtain topics of public interest concerning Samsung smartphone products on the Reddit platform, subsequent coupling with the sentiment analysis tool AlchemyAPI [93] enables the mining of public sentiment attitudes towards these topics [44]. This analysis aids in gauging sentiment tendencies under different levels of interest, assisting enterprises in pinpointing user demands and market pain points, thereby supporting business decisions. Certain studies directly apply sentiment and behavior joint topic models, such as the Sentiment and Behavior topic Model (SBTM) [16]. This model simultaneously integrates user sentiment and interactive behavioral patterns to facilitate complex topic discovery, resulting in enhanced distinctiveness in topic modelling outcomes. The realm of research pertaining to governmental decision-making is primarily reflected in online sentiment control and government service platform construction [94]. For example, the Online Topic and Sentiment Recognition Model (OTSRM) [95] utilizes the transitivity of sentiment intensity to construct a temporal-based topic-sentiment distribution. Using relative entropy, it calculates the maximum sentiment value of topic focus in adjacent time segments, dynamically identifying topic sentiment trends in text and enhancing the precision of sentiment warnings. User sentiment assumes varying meanings and orientations within diverse contexts, characterized by multi-dimensionality, strength, and subtlety. Developing optimized sentiment-topic models to enhance the precision of sentiment recognition within topics remains an ongoing exploration.

E. RECOMMENDATION RESEARCH

Recommendation systems effectively alleviate information overload. Key technologies encompass user modelling, object modelling, and recommendation algorithms. When modelling text information, approaches such as TF-IDF (Term Frequency-Inverse Document Frequency) [96], Bayesian classifiers, and k-nearest neighbors struggle to identify deeper semantic features within text [97]. Recommendation algorithms based on collaborative filtering and network structure often suffer from inadequate recommendation effectiveness due to data sparsity. The LDA method excels in data dimensionality reduction and latent semantic feature mining, efficiently identifying key information within user interests and recommended objects. This model is widely applied in user and recommended object modelling, including social network friend recommendations and personalized news recommendations [35]. In the era of big data, data sparsity and the prominence of massive dynamic features further underscore the challenges to topic mining and information recommendation performance based on the LDA method. For instance, Wang et al. [22] employed the Hadoop platform for structured processing of Weibo data. Subsequently, they utilized the LDA method to extract user Weibo topic information, effectively enhancing the effectiveness of big data information recommendations.

The LDA method effectively resolves the lack of semantics in traditional user and recommended object modelling processes and enhances recommendation accuracy through its excellent dimensionality reduction capability. However, challenges arising from information redundancy and overload in the big data environment impact the recommendation performance of the LDA method. As such, how to merge different methods to further enhance recommendation effectiveness remains an area of exploration.

V. DISCUSS AND CONCLUSION

Through the synthesis of literature review and architecture construction, we systematically expound upon the current application status and potential value of the LDA method in the domain of library and information science. Specifically, we present a systematic overview of the research progress in the domain of library and information science through the application of the LDA method, based on comprehensive domestic and international literature surveys from January 1, 2012, to December 31, 2022. Subsequently, by meticulously dissecting the key stages, we construct an analytical architecture for the application research of the LDA method. This architecture encompasses crucial steps such as text pre-processing, topic model construction (including topic model selection and optimal topic number determination),

and model solving. Building upon this foundation, we further delve into the applications of the LDA method in areas such as topic exploration, knowledge organization, academic evaluation, sentiment analysis, and recommendation research, including analyses of scientific literature, user-generated content, and online information resources. As a powerful topic modeling technique, the LDA method holds substantial promise in text analysis and knowledge discovery. We observe that while the general research directions both domestically and internationally exhibit similarities, specific application aspects display certain distinctions. For instance, in the selection of the optimal number of topics, scholars abroad endeavor to overcome the uncertainty associated with traditional probabilistic evaluation metrics (such as perplexity), focusing on introducing new model evaluation indicators that prioritize the reliability, stability, and interpretability of topic modeling results. On the other hand, scholars in China predominantly rely on perplexity, empirical methods, and other established indicators, with fewer attempts at incorporating novel evaluation metrics. In terms of application domains, the international community largely employs the LDA method for fundamental information organization research, while researchers in China have been more active in exploring knowledge graph and topic map construction. Overall, the following issues merit further investigation:

1) Tackling Complex Processing Challenges: Exploring the application value of the LDA topic model further in the face of challenges posed by massive-scale data, multi-modal data, etc. Currently, LDA method applications mainly focus on text data modelling, lacking exploration in audio, image, video, and other resource types. As the features of big data become more apparent in the domain of topic modelling, the LDA topic model confronts challenges related to data volume and increasing data complexity. Moreover, compared to single-modal text data, audio, images, videos, and other multi-modal data are richer in content and possess stronger topic representation abilities. Consequently, research into topic extraction from multi-modal data presents an important direction for development [2]. Future research can attempt to incorporate distributed and parallel computing from the domain of computer science to reduce the time taken by the LDA method in processing massive-scale documents, enhancing its capability to handle multi-source heterogeneous data. Additionally, further exploration of deep learning topic models that integrate word vectors, language models, and neural network structures could be conducted to advance research into multi-modal topic extraction, thereby improving the depth of topic mining and discovery capabilities and enhancing accurate information services in the domain of library and information science.

2) Enhancing Feature Word Extraction in Text Preprocessing: Ensuring the semantic quality of feature words by focusing on feature word extraction in the text preprocessing stage. The semantic quality of topic modelling results directly affects the reliability of topic analysis and subsequently influences practical application outcomes in domains like sentiment analysis and recommendation research. High semantic quality topic modelling results possess high inter-topic word associations within the same topic, as well as strong semantic differentiability between topics. This ability facilitates clear representation of content topics within the corpus. Text pre-processing is a foundational step in topic modelling and exerts a direct impact on the readability and interpretability of modelling results. However, existing research does not prioritize the quality of feature word extraction in the text pre-processing stage and often fixates on existing technical tools and semantic resources. There exists significant room for improvement in domain purity in feature word extraction. Leveraging appropriate NLP techniques to construct domain dictionaries, semantic resources, and high-quality large-scale annotated datasets for different topic mining tasks is an important future research direction.

3) Developing a Comprehensive LDA method Quality Evaluation System and Optimizing Topic Number Selection Methods: Presently, model quality evaluation for topic models predominantly relies on perplexity and empirical methods. However, model performance varies significantly under different evaluation indicators, making it difficult to objectively and effectively evaluate model quality by relying solely on one method. Improving traditional evaluation methods, introducing new methods, and adopting multi-indicator joint application for model quality evaluation is a notable trend. Future research can attempt to construct a more comprehensive topic model quality evaluation system, optimize topic number selection methods, and enhance the quality of topic modelling results.

4) Diversifying LDA method Application Modes and Deepening Model Application Research: Current research overly relies on the traditional LDA method, leading to a lack of exploration of emerging extended models' applications. Given the numerous parameters and complex structures inherent in various extended models, researchers in the domain of library and information science face higher demands for computer technology application capabilities. Future research should strive to optimize the time or space complexity of extended models, develop user-friendly open-source toolkits, and construct integrated model application tool systems to enhance model application efficiency and universality. Furthermore, considering the current application status of the model, the results of LDA method modelling often serve as intermediate stages for related research tasks. Different methods and tools should be employed based on the application context to solve specific research problems. For handling large-scale data, more attempts can be made to extend LDA methods based on distributed computing and deep learning. For processing small to medium-sized data, the coordinated application of the LDA method with traditional co-occurrence analysis, clustering analysis, community detection, and other topic analysis methods can ensure topic mining accuracy and achieve complementary effects. For example, co-occurrence analysis has higher readability

in topic clustering results on small to medium-sized datasets (document_count < 1,000), while although the LDA topic model lacks flexibility compared to co-occurrence analysis in choosing representative topic words, it can reflect the underlying topic structure in its most original state, thereby reducing bias. In summary, further exploration of extended model effectiveness in the domain of library and information science, synergistic application of the LDA method with traditional topic mining analysis methods, or utilizing the LDA method as a foundational component in combination with machine learning, knowledge graphs, big data, and relevant specialized algorithms represents a significant trend in current LDA method application research.

5) Exploring the Applications of Large Language Models in the Field of Library and Information Science: Large language models (LLMs) are a type of deep learning model that, after extensive training on vast amounts of data, can understand and generate natural language text. These models are typically built on neural network architectures and possess a large number of parameters, enabling them to perform a wide range of NLP tasks, including text generation, text classification, machine translation, sentiment analysis, and more. What distinguishes these models as ''large'' is their substantial scale, often featuring billions to trillions of parameters. One of the most prominent examples of a large language model is GPT-3 (Generative Pre-trained Transformer 3), developed by OpenAI, which boasts 175 billion parameters and is widely utilized across various natural language processing tasks. These models are typically constructed in two phases: pre-training and fine-tuning. During pre-training, the model is trained on extensive text corpora, learning rich language knowledge and grammar. In the fine-tuning phase, the model is further trained on task-specific data to adapt to particular applications, such as text classification or question answering. The introduction of these models has brought about a revolutionary transformation in the field of NLP, making applications involving the processing of natural language text more intelligent and efficient. These models find applications in a wide range of domains, including document summarization, intelligent customer support, search engine optimization, automatic translation, and more. In this work, we focus on LDA-based topic modeling methods. While LLMs excel in many NLP tasks, LDA retains unique advantages in certain situations. First, LDA is a generative model capable of uncovering the underlying topic structure of textual data, making it highly valuable in tasks like topic modeling and text classification. Second, LDA still holds advantages in scenarios where topic modeling or topic analysis is required, such as text summarization, text retrieval, and text classification. Furthermore, LDA can provide more intuitive topic models, aiding in the comprehension of the structure of textual data. Nevertheless, it is foreseeable that LLMs have numerous potential applications in the field of library and information science. They can enhance information retrieval systems, facilitate text summarization, improve text classification, support question-answering systems, enable semantic search, aid in constructing knowledge graphs, assist in term extraction, enhance recommendation systems, handle multilingual processing, and enhance topic modeling tasks. These applications have the potential to increase the efficiency and accuracy of information retrieval, document management, and knowledge analysis, ultimately providing better services to users in the realm of library and information science.

REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, ''Latent Dirichlet allocation,'' *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.
[2] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, ''Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey,'' *Multimedia Tools Appl.*, vol. 78, no. 11, pp. 15169–15211, Jun. 2019.
[3] Y. Qi and J. He, ''Application of LDA and word2vec to detect English off-topic composition,'' *PLoS ONE*, vol. 17, no. 2, Feb. 2022, Art. no. e0264552.
[4] L. Huang, Z. Hou, Y. Fang, J. Liu, and T. Shi, ''Evolution of CCUS technologies using LDA topic model and derwent patent data,'' *Energies*, vol. 16, no. 6, p. 2556, Mar. 2023.
[5] Sakshi and V. Kukreja, ''Recent trends in mathematical expressions recognition: An LDA-based analysis,'' *Expert Syst. Appl.*, vol. 213, Mar. 2023, Art. no. 119028.
[6] R. K. Gupta, R. Agarwalla, B. H. Naik, J. R. Evuri, A. Thapa, and T. D. Singh, ''Prediction of research trends using LDA based topic modeling,'' *Global Transitions*, vol. 3, no. 1, pp. 298–304, Jun. 2022.
[7] E. Quatrini, S. Colabianchi, F. Costantino, and M. Tronci, ''Clustering application for condition-based maintenance in time-varying processes: A review using latent Dirichlet allocation,'' *Appl. Sci.*, vol. 12, no. 2, p. 814, Jan. 2022.
[8] Y. Zhang and L. Zhang, ''Movie recommendation algorithm based on sentiment analysis and LDA,'' *Proc. Comput. Sci.*, vol. 199, pp. 871–878, Jan. 2022.
[9] D. Zheng, Z. Hong, N. Wang, and P. Chen, ''An improved LDA-based ELM classification for intrusion detection algorithm in IoT application,'' *Sensors*, vol. 20, no. 6, p. 1706, Mar. 2020.
[10] H. Wang, R. He, H. Liu, C. Wu, and B. Wang, ''Topic model on microblog with dual-streams graph convolution networks,'' in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2022, pp. 1–8.
[11] J. Chen, Z. Gong, and W. Liu, ''A nonparametric model for online topic discovery with word embeddings,'' *Inf. Sci.*, vol. 504, pp. 32–47, Dec. 2019.
[12] G. Tang, X. Chen, N. Li, and J. Cui, ''Research on the evolution of journal topic mining based on the BERT-LDA model,'' in *Proc. SHS Web Conf.*, vol. 152, 2023, p. 3012.
[13] M. Ćirić, B. Predić, D. Stojanović, and I. Ćirić, ''Single and multiple separate LSTM neural networks for multiple output feature purchase prediction,'' *Electronics*, vol. 12, no. 12, p. 2616, Jun. 2023.
[14] B. Ozyurt and M. A. Akcayol, ''A new topic modeling based approach for aspect extraction in aspect based sentiment analysis: SS-LDA,'' *Expert Syst. Appl.*, vol. 168, Apr. 2021, Art. no. 114231.
[15] V. Suarez-Lledo and J. Alvarez-Galvez, ''Prevalence of health misinformation on social media: Systematic review,'' *J. Med. Internet Res.*, vol. 23, no. 1, Jan. 2021, Art. no. e17187.
[16] X. Peng and Q. Xu, ''Investigating learners' behaviors and discourse content in MOOC course reviews,'' *Comput. Educ.*, vol. 143, Jan. 2020, Art. no. 103673.
[17] H. Liang, U. Ganeshbabu, and T. Thorne, ''A dynamic Bayesian network approach for analysing topic-sentiment evolution,'' *IEEE Access*, vol. 8, pp. 54164–54174, 2020.
[18] Y. Tian, Y. Song, F. Xia, T. Zhang, and Y. Wang, ''Improving Chinese word segmentation with wordhood memory networks,'' in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 8274–8285.
[19] K. Watanabe and A. Baturo, ''Seeded sequential LDA: A semi-supervised algorithm for topic-specific analysis of sentences,'' *Social Sci. Comput. Rev.*, May 2023.
[20] I. Savin, I. Ott, and C. Konop, ''Tracing the evolution of service robotics: Insights from a topic modeling approach,'' *Technol. Forecasting Social Change*, vol. 174, Jan. 2022, Art. no. 121280.

[21] Y. Qin, X. Qin, H. Chen, X. Li, and W. Lang, "Measuring cognitive proximity using semantic analysis: A case study of China's ICT industry," *Scientometrics*, vol. 126, no. 7, pp. 6059–6084, Jul. 2021.

[22] M. Wang and Q. Li, "A multi-agent based model for user interest mining on Sina Weibo," *China Commun.*, vol. 19, no. 2, pp. 225–234, Feb. 2022.

[23] L. Huang, Y. Cai, E. Zhao, S. Zhang, Y. Shu, and J. Fan, "Measuring the interdisciplinarity of information and library science interactions using citation analysis and semantic analysis," *Scientometrics*, vol. 127, no. 11, pp. 6733–6761, Nov. 2022.

[24] S. Kempeneer, A. Pirannejad, and J. Wolswinkel, "Open government data from a legal perspective: An AI-driven systematic literature review," *Government Inf. Quart.*, vol. 40, no. 3, Jun. 2023, Art. no. 101823.

[25] T. Kolajo, O. Daramola, and A. A. Adebiyi, "Real-time event detection in social media streams through semantic analysis of noisy terms," *J. Big Data*, vol. 9, no. 1, pp. 1–36, Dec. 2022.

[26] T. Ha, B. Beijnon, S. Kim, S. Lee, and J. H. Kim, "Examining user perceptions of smartwatch through dynamic topic modeling," *Telematics Informat.*, vol. 34, no. 7, pp. 1262–1273, Nov. 2017.

[27] M. Valeri and R. Baggio, "Italian tourism intermediaries: A social network analysis exploration," *Current Issues Tourism*, vol. 24, no. 9, pp. 1270–1283, May 2021.

[28] B.-H. Leem and S.-W. Eum, "Using text mining to measure mobile banking service quality," *Ind. Manage. Data Syst.*, vol. 121, no. 5, pp. 993–1007, Apr. 2021.

[29] A. Sharma, N. P. Rana, and R. Nunkoo, "Fifty years of information management research: A conceptual structure analysis using structural topic modeling," *Int. J. Inf. Manage.*, vol. 58, Jun. 2021, Art. no. 102316.

[30] Y.-H. Jiang, "Research on collaborative classification of E-commerce multi-attribute data based on weighted association rule model," in *Proc. 4th EAI Int. Conf. Adv. Hybrid Inf. Process. (ADHIP)*, Binzhou, China, Springer, 2021, pp. 377–388.

[31] B. Lu, B. Ma, D. Cheng, and J. Yang, "An investigation on impact of online review keywords on consumers' product consideration of clothing," *J. Theor. Appl. Electron. Commerce Res.*, vol. 18, no. 1, pp. 187–205, Jan. 2023.

[32] F. Benita, "Human mobility behavior in COVID-19: A systematic literature review and bibliometric analysis," *Sustain. Cities Soc.*, vol. 70, Jul. 2021, Art. no. 102916.

[33] D. Li, W. Guo, X. Chang, and X. Li, "From earth observation to human observation: Geocomputation for social science," *J. Geographical Sci.*, vol. 30, no. 2, pp. 233–250, Feb. 2020.

[34] E. Adamopoulou and L. Moussiades, "Chatbots: History, technology, and applications," *Mach. Learn. Appl.*, vol. 2, Dec. 2020, Art. no. 100006.

[35] L. Zhong, W. Wei, and S. Li, "Personalized news recommendation based on an improved conditional restricted Boltzmann machine," *Electron. Library*, vol. 39, no. 4, pp. 553–571, Nov. 2021.

[36] U. Chauhan and A. Shah, "Topic modeling using latent Dirichlet allocation: A survey," *ACM Comput. Surv.*, vol. 54, no. 7, pp. 1–35, Sep. 2022.

[37] W. Li and E. Suzuki, "Adaptive and hybrid context-aware fine-grained word sense disambiguation in topic modeling based document representation," *Inf. Process. Manage.*, vol. 58, no. 4, Jul. 2021, Art. no. 102592.

[38] M.-S. Chen, J.-Q. Lin, X.-L. Li, B.-Y. Liu, C.-D. Wang, D. Huang, and J.-H. Lai, "Representation learning in multi-view clustering: A literature review," *Data Sci. Eng.*, vol. 7, no. 3, pp. 225–241, Sep. 2022.

[39] M. Tan, Y. Xu, Z. Gao, T. Yuan, Q. Liu, R. Yang, B. Zhang, and L. Peng, "Recent advances in intelligent wearable medical devices integrating biosensing and drug delivery," *Adv. Mater.*, vol. 34, no. 27, Jul. 2022, Art. no. 2108491.

[40] R. Hu, W. Ma, W. Lin, X. Chen, Z. Zhong, and C. Zeng, "Technology topic identification and trend prediction of new energy vehicle using LDA modeling," *Complexity*, vol. 2022, pp. 1–20, Mar. 2022.

[41] W. Seo, "A patent-based approach to identifying potential technology opportunities realizable from a firm's internal capabilities," *Comput. Ind. Eng.*, vol. 171, Sep. 2022, Art. no. 108395.

[42] L. Wright, E. Paul, A. Steptoe, and D. Fancourt, "Facilitators and barriers to compliance with COVID-19 guidelines: A structural topic modelling analysis of free-text data from 17,500 U.K. adults," *BMC Public Health*, vol. 22, no. 1, pp. 1–22, Dec. 2022.

[43] B. Karas, S. Qu, Y. Xu, and Q. Zhu, "Experiments with LDA and Top2Vec for embedded topic discovery on social media data—A case study of cystic fibrosis," *Frontiers Artif. Intell.*, vol. 5, Aug. 2022, Art. no. 948313.

[44] M. Giannakis, R. Dubey, S. Yan, K. Spanaki, and T. Papadopoulos, "Social media and sensemaking patterns in new product development: Demystifying the customer sentiment," *Ann. Oper. Res.*, vol. 308, nos. 1–2, pp. 145–175, Jan. 2022.

[45] T. Li, Z. Zeng, J. Sun, and S. Sun, "Using data mining technology to analyse the spatiotemporal public opinion of COVID-19 vaccine on social media," *Electron. Library*, vol. 40, no. 4, pp. 435–452, Aug. 2022.

[46] G. Orosz, Z. Szántó, P. Berkecz, G. Szabó, and R. Farkas, "HuSpaCy: An industrial-strength Hungarian natural language processing toolkit," 2022, *arXiv:2201.01956*.

[47] B. Gencoglu, M. Helms-Lorenz, R. Maulana, E. P. W. A. Jansen, and O. Gencoglu, "Machine and expert judgments of Student perceptions of teaching behavior in secondary education: Added value of topic modeling with big data," *Comput. Educ.*, vol. 193, Feb. 2023, Art. no. 104682.

[48] X. Ge, Y. Yang, J. Chen, W. Li, Z. Huang, W. Zhang, and L. Peng, "Disaster prediction knowledge graph based on multi-source spatiotemporal information," *Remote Sens.*, vol. 14, no. 5, p. 1214, Mar. 2022.

[49] H. Zhou, J. Zhang, and Y. Zhao, "Wind power policies in China: Development themes, evaluation, and recommendations," *Energy Sci. Eng.*, vol. 11, no. 3, pp. 1044–1059, Mar. 2023.

[50] O. Ballester and O. Penner, "Robustness, replicability and scalability in topic modelling," *J. Informetrics*, vol. 16, no. 1, Feb. 2022, Art. no. 101224.

[51] K. Taghandiki and M. Mohammadi, "Topic modeling: Exploring the processes, tools, challenges and applications," Tech. Rep., 2023.

[52] H. Liu, "On the training mode of innovative and entrepreneurial talents in higher vocational finance and economics professional groups under the background of 'big wisdom and cloud,'" *Adv. Multimedia*, vol. 2022, pp. 1–11, Oct. 2022.

[53] X. Li and Y. Kuang, "Research on the evolution of logistics identification marking technology based on mallet-LDA model and social network," in *Proc. WHICEB*, 2022.

[54] M. Baghmohammad, A. Mansouri, and M. Cheashmehsohrabi, "Identification of topic development process of knowledge and information science field based on the topic modeling (LDA)," *Iranian J. Inf. Process. Manag.*, vol. 36, no. 2, pp. 297–328, 2022.

[55] C. Hennesy and D. Naughton, "Computational topic models of the library quarterly," *Portal, Libraries Acad.*, vol. 22, no. 3, pp. 745–768, Jul. 2022.

[56] A. S. Miner, S. A. Stewart, M. C. Halley, L. K. Nelson, and E. Linos, "Formally comparing topic models and human-generated qualitative coding of physician mothers' experiences of workplace discrimination," *Big Data Soc.*, vol. 10, no. 1, pp. 1–17, Jan. 2023.

[57] K. Balog, D. Maxwell, P. Thomas, and S. Zhang, "Report on the 1st simulation for information retrieval workshop (Sim4IR 2021) at SIGIR 2021," *ACM SIGIR Forum*, vol. 55, no. 2, pp. 1–16, Dec. 2021.

[58] Q. Huilin and S. Bo, "Research on identification methods of scientific research hotspots under multi-source data," *Library Inf. Service*, vol. 64, no. 5, p. 78, 2020.

[59] H. Liu, Z. Chen, J. Tang, Y. Zhou, and S. Liu, "Mapping the technology evolution path: A novel model for dynamic topic detection and tracking," *Scientometrics*, vol. 125, no. 3, pp. 2043–2090, 2020.

[60] Q. Shen, S. Han, Y. Han, and X. Chen, "User review analysis of dating apps based on text mining," *PLoS ONE*, vol. 18, no. 4, Apr. 2023, Art. no. e0283896.

[61] Y. Guo, C. Jia, C. Wu, and Y. Tu, "Social media rumor identification based on random forest classification and feature engineering: Case study on Weibo platform: Social media rumor identification based on random forest classification," in *Proc. 7th Int. Conf. Big Data Comput.*, May 2022, pp. 109–118.

[62] G. Zhu, Z. Pan, Q. Wang, S. Zhang, and K.-C. Li, "Building multi-subtopic bi-level network for micro-blog hot topic based on feature co-occurrence and semantic community division," *J. Netw. Comput. Appl.*, vol. 170, Nov. 2020, Art. no. 102815.

[63] C. N. Novera, Z. Ahmed, R. Kushol, P. Wanke, and M. A. K. Azad, "Internet of Things (IoT) in smart tourism: A literature review," *Spanish J. Marketing*, vol. 26, no. 3, pp. 325–344, Dec. 2022.

[64] X. Han and J. Wang, "Modelling and analyzing the semantic evolution of social media user behaviors during disaster events: A case study of COVID-19," *ISPRS Int. J. Geo-Inf.*, vol. 11, no. 7, p. 373, Jul. 2022.

[65] C. Shuting et al., "Discovery and evolution of hot topics of network public opinion in emergencies based on time-series supernetwork," *J. Tsinghua Univ., Sci. Technol.*, vol. 63, no. 6, pp. 968–979, 2023.

[66] S. Karimi, A. Shakery, and R. M. Verma, "Enhancement of Twitter event detection using news streams," *Natural Lang. Eng.*, vol. 29, no. 2, pp. 181–200, Mar. 2023.
[67] F. K. Sufi, "Identifying the drivers of negative news with sentiment, entity and regression analysis," *Int. J. Inf. Manage. Data Insights*, vol. 2, no. 1, Apr. 2022, Art. no. 100074.
[68] P. T. Jaeger, J. Lin, and J. M. Grimes, "Cloud computing and information policy: Computing in a policy cloud?" *J. Inf. Technol. Politics*, vol. 5, no. 3, pp. 269–283, Oct. 2008.
[69] C. Song, J. Guo, F. Gholizadeh, and J. Zhuang, "Quantitative analysis of food safety policy—Based on text mining methods," *Foods*, vol. 11, no. 21, p. 3421, Oct. 2022.
[70] B. Lin and C. Huang, "Analysis of emission reduction effects of carbon trading: Market mechanism or government intervention?" *Sustain. Prod. Consumption*, vol. 33, pp. 28–37, Sep. 2022.
[71] C. Tan, F. Yu, Y. Gao, and B. Hu, "A generation method of enterprise appropriate policy tags based on LDA model," in *Proc. 4th Int. Conf. Electron. Eng. Informat. (EEI)*, Jun. 2022, pp. 1–8.
[72] W. Chen, Z. Li, H. Fang, Q. Yao, C. Zhong, J. Hao, Q. Zhang, X. Huang, J. Peng, and Z. Wei, "A benchmark for automatic medical consultation system: Frameworks, tasks and datasets," *Bioinformatics*, vol. 39, no. 1, Jan. 2023, Art. no. btac817.
[73] Y. Ahuja, Y. Zou, A. Verma, D. Buckeridge, and Y. Li, "MixEHR-guided: A guided multi-modal topic modeling approach for large-scale automatic phenotyping using the electronic health record," *J. Biomed. Informat.*, vol. 134, Oct. 2022, Art. no. 104190.
[74] S. F. Ahmed, M. S. B. Alam, M. Hassan, M. R. Rozbu, T. Ishtiak, N. Rafa, M. Mofijur, A. B. M. S. Ali, and A. H. Gandomi, "Deep learning modelling techniques: Current progress, applications, advantages, and challenges," *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 13521–13617, Nov. 2023.
[75] S. Mishra, P. Shukla, and R. Agarwal, "Analyzing machine learning enabled fake news detection techniques for diversified datasets," *Wireless Commun. Mobile Comput.*, vol. 2022, pp. 1–18, Mar. 2022.
[76] M. I. Nadeem, K. Ahmed, D. Li, Z. Zheng, H. K. Alkahtani, S. M. Mostafa, O. Mamyrbayev, and H. A. Hameed, "EFND: A semantic, visual, and socially augmented deep framework for extreme fake news detection," *Sustainability*, vol. 15, no. 1, p. 133, Dec. 2022.
[77] C. Yang and C. Huang, "Quantitative mapping of the evolution of AI policy distribution, targets and focuses over three decades in China," *Technol. Forecasting Social Change*, vol. 174, Jan. 2022, Art. no. 121188.
[78] P. Zhao, Y. Gao, and X. Sun, "How does artificial intelligence affect green economic growth?—Evidence from China," *Sci. Total Environ.*, vol. 834, Aug. 2022, Art. no. 155306.
[79] S. Bai, W. Yu, and M. Jiang, "Promoting the tripartite cooperative mechanism of E-commerce poverty alleviation: Based on the evolutionary game method," *Sustainability*, vol. 15, no. 1, p. 315, Dec. 2022.
[80] C.-M. Sun, Y. Zhou, and S. Yuan, "Check for updates a visual analysis of E-government research in China based on co-word clustering," in *Proc. 11th 12th EAI Int. Conf. Big Data Technol. Appl. (BDTA)*. Springer, vol. 480, 2023, p. 330.
[81] D. Zhou, B. Zhou, Z. Zheng, A. Soylu, G. Cheng, E. Jimenez-Ruiz, E. V. Kostylev, and E. Kharlamov, "Ontology reshaping for knowledge graph construction: Applied on Bosch welding case," in *Proc. Int. Semantic Web Conf.*, vol. 2022. Cham, Switzerland: Springer, 2022, pp. 770–790.
[82] A. Purnomo, N. Asitah, E. Rosyidah, A. Septianto, and M. Firdaus, "Mapping of computational social science research themes: A two-decade review," in *Intelligent Systems and Sustainable Computing*, 2022, pp. 617–625.
[83] Y. Yin, L. Zhang, Y. Wang, M. Wang, Q. Zhang, and G.-Z. Li, "Question answering system based on knowledge graph in traditional Chinese medicine diagnosis and treatment of viral hepatitis B," *BioMed Res. Int.*, vol. 2022, pp. 1–8, Feb. 2022.
[84] J. Chen, S. Du, and S. Yang, "Mining and evolution analysis of network public opinion concerns of stakeholders in hot social events," *Mathematics*, vol. 10, no. 12, p. 2145, Jun. 2022.
[85] B. Lin, X. Zhu, and J. Hu, "Machine learning-based Weibo user group profiling under hot events," in *Proc. Int. Conf. Culture-Oriented Sci. Technol. (CoST)*, Aug. 2022, pp. 283–287.
[86] N. Sivapriya and R. Mohandas, "Optimal route selection for mobile ad-hoc networks based on cluster head selection and energy efficient multicast routing protocol," *J. Algebr. Statist.*, vol. 13, no. 2, pp. 595–607, 2022.
[87] M. Bianchini, M. Gori, and F. Scarselli, "Inside PageRank," *ACM Trans. Internet Technol.*, vol. 5, no. 1, pp. 92–128, Feb. 2005.
[88] A. Bihari, S. Tripathi, and A. Deepak, "A review on h-index and its alternative indices," *J. Inf. Sci.*, vol. 49, no. 3, pp. 624–665, Jun. 2023.
[89] P. Sud and M. Thelwall, "Evaluating altmetrics," *Scientometrics*, vol. 98, no. 2, pp. 1131–1143, Feb. 2014.
[90] X. Kong, H. Jiang, W. Wang, T. M. Bekele, Z. Xu, and M. Wang, "Exploring dynamic research interest and academic influence for scientific collaborator recommendation," *Scientometrics*, vol. 113, no. 1, pp. 369–385, Oct. 2017.
[91] D. Yu and Z. Yan, "Knowledge diffusion trajectories of PageRank: A main path analysis," *J. Inf. Sci.*, Apr. 2023.
[92] P. William, A. B. Pawar, M. A. Jawale, A. Badholia, and V. Verma, "Energy efficient framework to implement next generation network protocol using ATM technology," *Meas., Sensors*, vol. 24, Dec. 2022, Art. no. 100477.
[93] P. William, A. Shrivastava, P. S. Chauhan, M. Raja, S. B. Ojha, and K. Kumar, "Natural Language processing implementation for sentiment analysis on tweets," in *Mobile Radio Communications and 5G Networks*. Singapore: Springer, 2023, pp. 317–327.
[94] S. Verma, "Sentiment analysis of public services for smart society: Literature review and future research directions," *Government Inf. Quart.*, vol. 39, no. 3, Jul. 2022, Art. no. 101708.
[95] A. R. Pathak, M. Pandey, and S. Rautaray, "Topic-level sentiment analysis of social media data using deep learning," *Appl. Soft Comput.*, vol. 108, Sep. 2021, Art. no. 107440.
[96] H. Christian, M. P. Agus, and D. Suhartono, "Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF)," *ComTech, Comput., Math. Eng. Appl.*, vol. 7, no. 4, p. 285, Dec. 2016.
[97] M. Azam, T. Ahmed, F. Sabah, and M. I. Hussain, "Feature extraction based text classification using k-nearest neighbor algorithm," *Int. J. Comput. Sci. Netw. Secur.*, vol. 18, no. 12, pp. 95–101, 2018.
[98] D. K. Kirtania and S. K. Patra, "OpenAI ChatGPT generated content and similarity index: A study of selected terms from the library & information science (LIS)," *Qeios*, 2023.

XINGZHOU PAN was born in Jinan, Shandong, China, in 1982. He received the master's degree from Xinan University, China. He is currently with the Library of Shandong University. His research interests include artificial intelligence technology and research data management.

YU XUE was born in Jinan, Shandong, China, in 1989. She received the master's degree from Shandong University, China. She is currently with the Library of Shandong University. Her research interests include artificial intelligence technology and research data management.