The Frontiers of Society, Science and Technology
ISSN 2616-7433 Vol. 5, Issue 17: 71-78, DOI: 10.25236/FSST.2023.051713

Topic Discovery in Operations Management Research Area Based on LDA Topic Modeling

Chongjian Song

School of Economics and Management, Guangxi Normal University, Guilin, China
scj199700@163.com

Abstract: *To summarize and analyse the relevant topics of massive information content by human power alone will consume a huge amount of time and energy. On the contrary, through the development of information science and other disciplines, the use of big data analytics to extract and analyse the themes of massive text content has appeared in many studies, and LDA is one of the more widely used themes, which can be extracted from a large number of Chinese or English texts to extract the potential themes, and has been applied to a variety of research. It is not uncommon to apply LDA theme model to extract potential topics in a research field. In this paper, we collect relevant literature information in the field of operations management research through crawler technology. Through a series of preprocessing processes, the text content can be processed by the LDA topic model to extract the five topic contents of the text. In order to show the theme content more intuitively, the LDA theme results are further visualized. Finally, the research hotspots in the domestic operation management research field are obtained.*

Keywords: *Operations management; Visualization; LDA topic modelling; Text mining*

1. Introduction

The development of big data and information science so that many previously unprocessable unstructured text data can also be processed and analyzed through the analysis of big data, to provide researchers with a wider range and greater degree of freedom of the object of study at the same time, researchers in many areas of research do not have to adhere to the traditional structured text of the small data analysis methods[1]. However, most big data analysis methods are implemented based on software such as R-studio, Python or Hadoop. These software are in turn difficult for many researchers who do not specialize in information science or computer science to master quickly. Therefore, there will be a research gap in many research fields due to the difficulty of researchers in mastering big data analytic methods. And through the development and gradual iteration of big data analytics methods, it is easier for researchers without the foundation of related disciplines to master, and gradually become a research hotspot. Therefore, applying big data analysis methods to research in other research fields has gradually become a research hotspot. Text Miming is a very mature and effective big data technology for extracting and discovering knowledge from unstructured, voluminous and complex text information[2]. Many scholars apply text mining methods to research, such as: research in psychology[3], research in organizational management[2], research in social media language[4] and so on. It also reflects to some extent the effectiveness of text mining techniques for text information processing and analysis. Latent Dirichlet Allocation (LDA) topic extraction model is one of the text mining methods, after a long time of improvement and development, in the field of text mining is gradually becoming mature, and in the process of recognizing the topic of long text to achieve good results. It has achieved good results in the process of recognizing topics in long text[5]. In this paper, we will use the LDA topic model to analyze the relevant literature information in the field of operations management research in order to identify the research hotspots in this research area.

2. Theoretical foundations

Latent Dirichlet Allocation was proposed by Blei[6] in 2003, is a generative probabilistic approach to unsupervised machine learning for modeling corpora, and is now more mature in the field of text mining research, and it can achieve good results for the extraction and recognition of topics in documents with a large amount of text. In the LDA model, the document is regarded as a mixed probability distribution of potential topics, and the topics are regarded as the probability of several words. Given a corpus D of

M documents, where document d has N_d words $d(d \in 1,\dots,M)$. LDA models D according to the following generative process:

(1) Select the multinomial distribution φ_t of subject $t(t \in 1,\dots,T)$ from the Dirichlet with hyperparameter β;

(2) Select the multinomial distribution θ_d of document $d(d \in 1,\dots,M)$ from the Dirichlet with hyperparameter α;

(3) For a word $W_n(n \in 1,\dots,N_d)$ in a document d,

a. Select topic Z_n from θ_d.

b. Select a word W_n from φ_{zn}.

In the above generation process, the words in the document knowledge observed variables, while the others are latent variables (φ and θ) and hyperparameters (β and α).

The probability of observing the document set D is derived from the corpus computation. The formula is as follows:

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{Z_{dn}} p(Z_{dn}|\theta_d) p(w_{dn}|Z_{dn},\beta)\right) d\theta_d \qquad (1)$$

In the equation(1), α, β are the hyperparameters that need to be determined. D represents the test document set, M represents the number of texts. θ_d represents the topic distribution of document d, while w_{dn} refers to the nth word in the dth document. Z_{dn} indicates the topic of the nth word in the dth document.

There are many parameters of the LDA model, how to determine these parameters and ensure the rigor of the overall study. Some researchers have given methods to estimate the parameters of LDA model such as Gibbs sampling, Expectation maximization, Variational Bayes inference[7]. In this paper Gibbs sampling method is used to obtain the distribution of topics and words, α,β are used with default values, k (optimal number of topics) is determined in many ways such as: empirical setting method with heuristics, perplexity, thematic consistency analysis way of making judgments. In this paper, Topic Coherence score (Topic Coherence) will be used to determine the k-value.

3. Literature review

3.1 Research on LDA topic modelling

Based on this property of the LDA topic model, the content of representative literature in a research area as the object of study can be analyzed in a quantitative manner to draw conclusions: Asmussen C. B.[8] (2019), on the other hand, used the LDA topic model to analyze specific literatures so as to identify the topics of these literatures, which further resulted in a literature review; YANG J H et al.[9] (2022) compiled a review of the use of text mining techniques in educational research. It can be found that the use of text mining techniques in educational research in general is in its infancy, but the use of these methods is becoming increasingly popular and important in this field of study. ANTONS D et al.[10] (2020) conducted the first combing of innovation research applying text mining and found that text mining methods are in a mature state in innovation research. Based on the combing results, the future focus of text mining applications in innovation research is further depicted. FANG D et al.[11] (2018) analyzed the abstract information of 3,737 articles in accounting journals using the LDA topic model and obtained 32 significant research themes. The results were further measured by a regression analysis, which led to the identification of seven popular themes and six cold themes. Yin B et al.[12] (2022) collected a large amount of literature summary information related to blended learning in WOS literature database, and obtained the main research topics of blended learning as well as its development trend through LDA topic modeling and word cloud analysis methods. This shows that text mining and LDA topic modeling can be a good way to mine and discover potential topics and research hotspots in a research field. This is the topic information generated by the software after processing a large amount of literature content, compared with manual argumentation, this research method can better help researchers to discover the research topics and trends in a research field.

3.2 Research on operations management

In the era of Industry 4.0, the equipment in many production lines have been greatly developed and upgraded, and the production efficiency has been unprecedentedly improved through the role of open operating systems and the Internet of Things, but in the process of these operations and running more need to intervene in the operation and management[13], and the importance of the study of operation and management is particularly important in the development of Industry 4.0. At the same time, a large amount of information and data is generated in these processes, with big data playing a more prominent role and receiving considerable attention in all sectors and fields, and researchers analyze and process this information to derive the hidden information present in the chain. For example, Ogbuke[14] (2022) conducted a study on the application of big data in supply chain management and the corresponding benefits and efficiencies it can bring to the society as well as to the industry; Maheshwari S[15] (2021) sorted out and reviewed the research on big data analytics in inventory, logistics, and supply chain management, clarified the role that big data analytics plays in the above mentioned areas, and further revealed the issues that remain unresolved. Choi[16](2018) systematized the research on applying big data analytics in the field of operations management research, mentioning that data mining methods can be used in the research and analysis of market segmentation and collaborative process. TALWAR S [17](2020) conducted a systematic compilation and assessment of big data analytics approaches present in operations and supply chain management (OSCM) activities, applying a systematic literature review methodology to reveal existing research trends, distill key themes and identify areas for future research. XU J [18] (2023) using the Delphi method identified the correlation between big data analytics methods and supply chain planning activities, especially the positive impact on planning accuracy, response time and supply chain planning flexibility. From the research of these scholars, the methods of big data analysis and data mining are less applied in the field of operation management research or in the real operation management link, but it will become a more popular research trend in the future research. In this paper, we will take the data mining method to analyze the text of the literature on operation management, so as to get the potential subject content of this research field and reveal its hot research content.

4. Research design

The flowchart of the study is shown in *Figure 1*. A brief description of the processes in *Figure 1* as well as the software processing is shown in *Table 1*.

Table 1 Software and libraries used for process treatment

No.	Process	Software	Introduction
P1	Crawling texts	Octopus V8	Save research content locally in large batches
P2.1	Text preprocessing	Microsoft Word 2016	Delete fixed-format text in bulk using "find and replace".
P2.2	Add user defined dictionaries	TXT document	Ensure that proper nouns consisting of multiple words are not automatically broken up by the Jieba library.
P2.3	Segmentation	Jupyter Notebook5.7.8	Split all coherent sentences in the text into separate words for further analysis by the software.
P2.4	Rejecting stop words		Remove words from the text that do not contribute to the purpose of the study.
P3	Calculate topic consistency		Calculate the topic consistency of the LDA model to determine the number of topics in the LDA model.
P4.1	Optimal number of topics		Set the number of topics that maximize topic consistency
P4.2	Extracting topics		Run the LDA theme model to extract the theme information of the text.
P5	LDA topic visualization		Visualize the results of extracting themes from the LDA theme model.

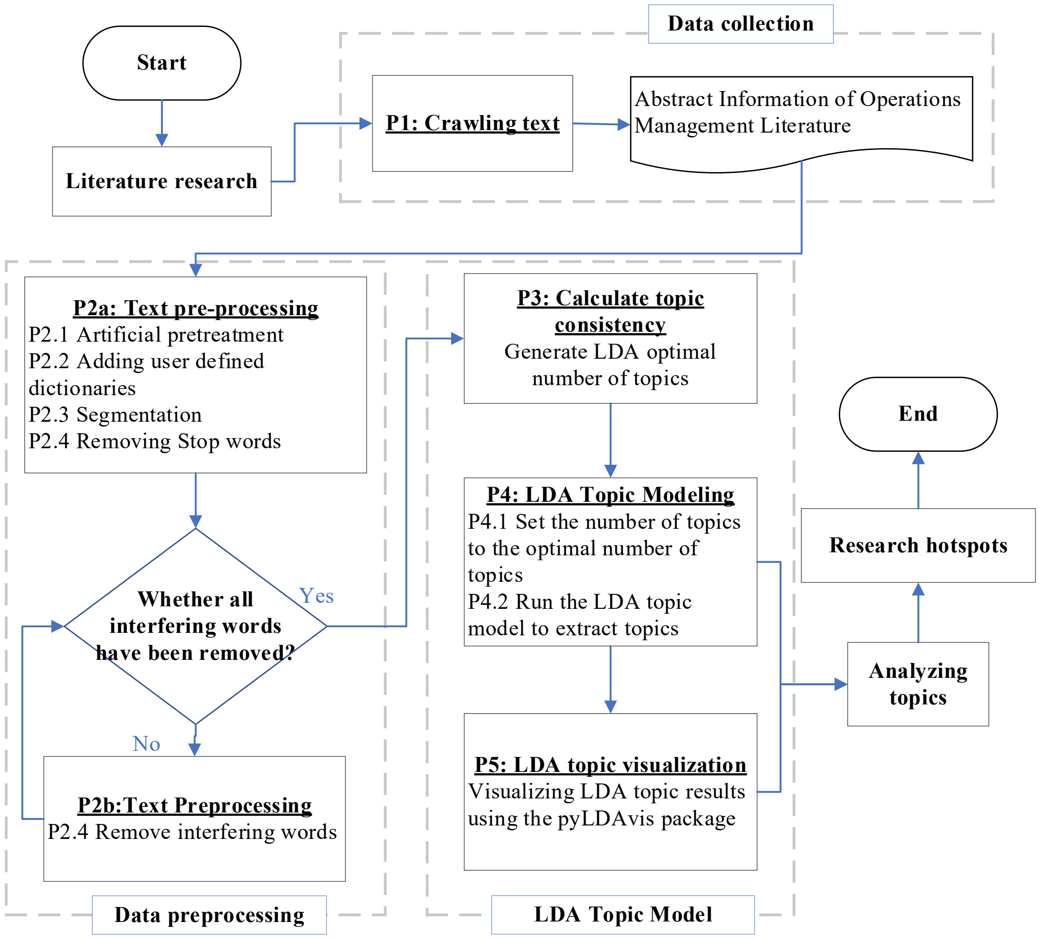

Data collection
Start
Literature research
P1: Crawling text
Abstract Information of Operations Management Literature
P2a: Text pre-processing
P2.1 Artificial pretreatment
P2.2 Adding user defined dictionaries
P2.3 Segmentation
P2.4 Removing Stop words
Whether all interfering words have been removed?
Yes
No
P2b:Text Preprocessing
P2.4 Remove interfering words
P3: Calculate topic consistency
Generate LDA optimal number of topics
P4: LDA Topic Modeling
P4.1 Set the number of topics to the optimal number of topics
P4.2 Run the LDA topic model to extract topics
P5: LDA topic visualization
Visualizing LDA topic results using the pyLDAvis package
End
Research hotspots
Analyzing topics
Data preprocessing
LDA Topic Model

Figure 1: Research Flowchart

4.1 Data sources

Source: China National Knowledge Infrastructure (CNKI);

Collection method: The search terms were set as "operation management" and "supply chain management", and the document type was set as "academic journal". The source categories were set as "CSSCI" and "CSCD";

Collection date: December 1, 2022;

Collection results: A total of 2078 documents were collected, and after reviewing them one by one and deleting the papers with missing abstracts and keywords, 1806 valid documents were obtained.

4.2 Research process

In *Figure 1*, the crawled text step was used to collect the study content using Octopus V8 collector; the data preprocessing step was processed using Microsoft Word 2016 and Python software; and the LDA topic modeling step was processed using Jupyter Notebook software.

The text pre-processing process needs to clean the text, is for the research will produce interference, fixed-format text, no real meaning of the word (see Table 2) to be deleted to ensure the accuracy of the research. After deleting the fixed-format text, the content of the text needs to be divided into words. Segmentation is the process of dividing the sentences that make up the information into words, and it is only through segmentation that the software can continue to process the text subsequently. The segmentation tool is Jiaba, which has the best segmentation effect and the highest recognition in the current research. And the word segmentation process should be an iterative process, because at the beginning of the study, it is not possible to a priori remove all the words that interfere with the study at once and accurately, and it is necessary to remove the words that appear more frequently but are not meaningful to the study one by one through several trials. The deletion of meaningless words can be realized by removing deactivated words with the help of the software, and the deactivated word list is adopted from the HITC deactivated word list. However, before removing deactivated words and word frequency statistics, it is necessary to realize the operation of word division, which is the process of dividing the sentences that constitute the information into words, and only after the word division can the software continue to the text for subsequent processing.

Table 2: Examples of words to be deleted

Word type	**Examples of words**	**Treatment**
Text Fixed Format	Keywords, abstract, purpose, etc.	Delete with the help of Microsoft Word 2016
Modal particle	Ah, oh, and so on.	Delete with the help of a disambiguation operation
Nonsense word	Of, with, all, etc.	Delete with the help of a disambiguation operation

4.3 LDA topic modelling

The LDA topic model needs to be implemented with the help of Gensim package, which firstly needs to construct the DT Matrix (Document-Term Matrix), through which words are turned into computer-understandable encodings, and the LDA model is run and trained on the DT Matrix, and the topic consistency of the overall text is computed to determine the number of topics for the LDA model. Topic consistency score refers to whether the high probability words corresponding to each topic generated by the model are semantically consistent or not, and the higher the score is, the better the model is. Therefore, combining with *Figure 2*, we know that the model fits best when the number of topics is set to 5. Therefore, when running the LDA algorithm subsequently, the number of topics is set to 5 and the number of topic words is set to 10.

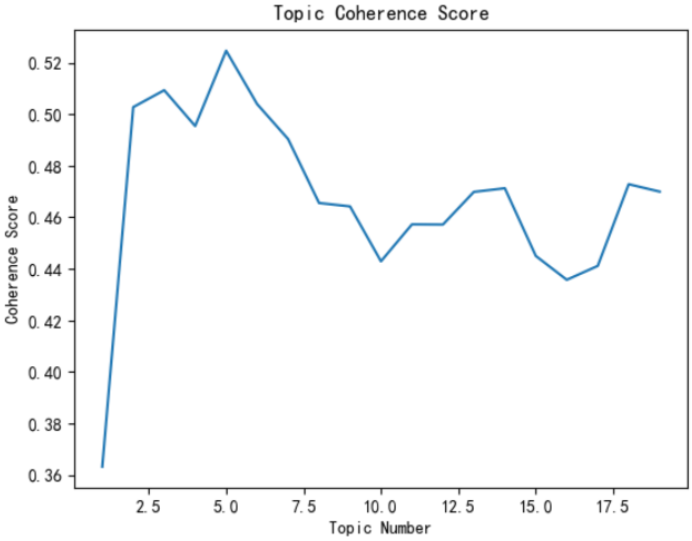


Figure 2: Trend chart of topic coherence scores

4.4 Experimental results

4.4.1 LDA topic modelling results

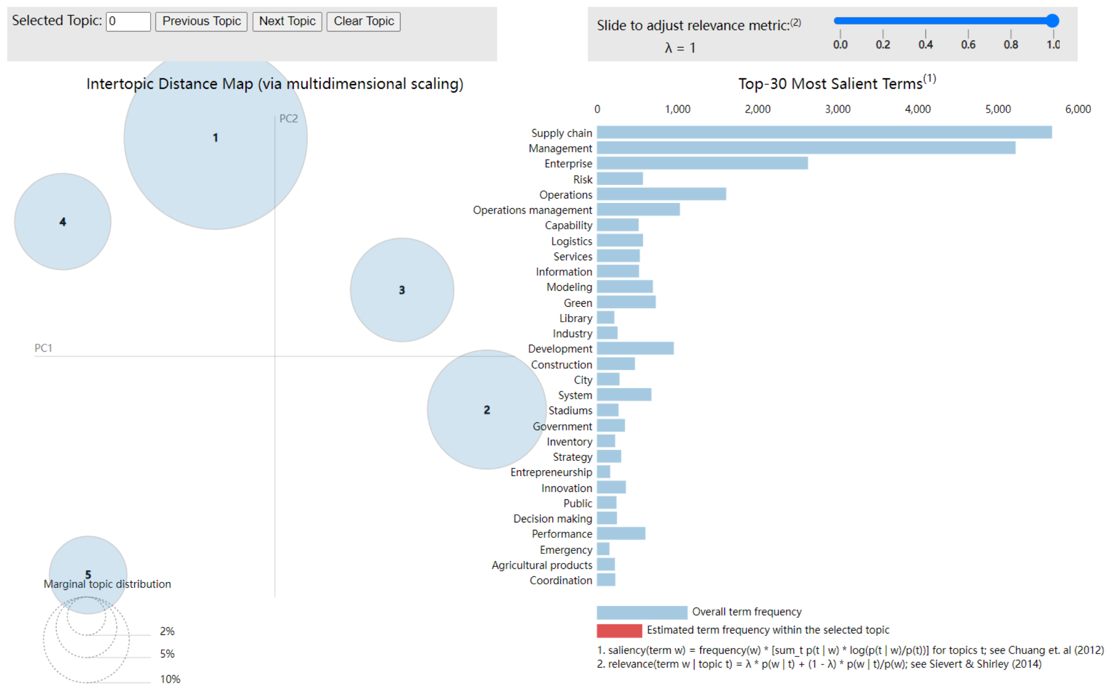
Determining the number of topics for the LDA topic model based on the topic consistency scores calculated in the previous section allows determining the value of k (number of topics). Meanwhile, default values are used for α and β. Once the main parameters are determined, the LDA topic information can be output, and each topic contains word item information under it. These lexical items are the high probability feature words of each topic, and each topic can be generalized by these high probability feature words to match the topic identity of the topic's high probability feature words(See Table 3).

Table 3: LDA Topic Results Table

No.	Topic Description	Topic Items
Topic1	Supply Chain Management	supply chain, management, enterprise, research, green, analysis, performance, knowledge, development, strategy
Topic2	Government Operations	operation, operations management, construction, model, government, development, China, cities, stadiums, public
Topic3	Industrial Operations	operations management, operations, industry, development, innovation, platform, research, analysis, impact, resource
Topic4	Operation Risk Study	risk, model, system, analysis, strategy, supply chain, decision making, cooperation, agricultural products, coordination
Topic5	Operations Industry	capacity, service, library, logistics, entrepreneurship, information, inventory, emergency, operations management, hospitals

4.4.2 Visualization of LDA topic results

With the LDA topic visualization loading package pyLDAvis developed by Sievert and Shirley[19], it is very intuitive to present LDA topics graphically as the relationship between each lexical item and the topic as well as the interconnections between the subjects. The LDA topic model is visualized with the help of LDAvis loading package and the results are shown in *Figure 3*. In the *Figure 3*, there are five shaded circles representing five subjects, and the serial number inside the circle corresponds to the serial number of the subject in *Table 3*, and the larger the area of the circle indicates that the number of texts present in the subject is also larger, and the more influence it has[20]. The farther the distance between the circles, the less similarity between the two themes; conversely, the higher similarity between the two topics. In the right bar graph, the blue long bar indicates the probability that the word belongs to the corresponding theme, and the red long bar indicates the degree of association between the word and the corresponding topic. In *Figure 3*, there is a certain distance between the distributions of topic, i.e., the overlap between themes is small, which also confirms the rationality of the number of themes set side by side.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 3: LDA Topic Visualization Results

5. Research findings

Overall, it can be seen from the table and figure that operations management is widely used in subjects such as libraries, logistics, hospitals, highways, networks, agricultural products, and also has a very important role in behaviors or subjects such as entrepreneurship, suppliers, and inventory. From *Table 3* and the explanation for the formation of the regions in *Figure 3*, it can be seen that in the field of operations management research, these five subjects are the ones that have received the attention of a large number of researchers in this field of research. Moreover, the content of texts belonging to supply chain management and government operations is much more than the remaining three themes, which shows that in the field of operations management research, there is a large proportion of research on supply chain management as well as government operations, which mainly involves thinking and discussing about environmental protection, performance, and technological and theoretical issues (see *Table 3*). In addition to this, the operations management research area also includes studies on industry, government and operations risk assessment.

Undoubtedly, big data methods have surfaced in the field of operations management research in foreign countries, intervening in that research in order to analyze and solve related problems[21], but it is difficult to see the use of big data methods in research in the domestic literature related to operations management. This can be seen in the figure and table, where the words big data or big data methods do not appear. In addition, in the process of the transformation of Industry 4.0, more and more information will be carried or occur in the enterprise or production line, so the use of big data methods in domestic operations management research may be a new research direction and research hotspot.

6. Conclusion

In this paper, the literature related to operations management included in the CNKI database (searched with specific search criteria) was used as the research content, and 2078 documents were collected. The Octopus V8 collector was used to collect the abstract and title information of these literatures, and the content of the literatures with incomplete information was deleted, and finally 1,806 valid literatures were obtained. After data preprocessing and LDA topic modeling, five topics (i.e., Supply Chain Management, Government Operations, Industrial Operations, Operational Risk Research, and Operational Industries) were obtained in the field of operations management research. The results of LDA themes were visualized using the LDAvis tool, and while evaluating the LDA themes through the visualization results, it was further found that the two themes of supply chain management and government operations have more text content and influence than the other three themes. The trend of combining big data methods with operations management research is also explored. The results show that the LDA topic model can well extract the research hotspots and topics in the field of operation management research in China, which can help researchers grasp the research hotspots in the field of research and discover the emerging research topics.

References

[1] YUCHENG Z, SHAN X, LONG Z, et al. Big data and human resource management research: An integrative review and new directions for future research [J]. Journal of Business Research, 2021, 133.
[2] KOBAYASHI V B, MOL S T, BERKERS H A, et al. Text Mining in Organizational Research [J]. Organizational Research Methods, 2018, 21(3).
[3] EVAN C E, P W S. A practical guide to big data research in psychology [J]. Psychological methods, 2016, 21(4).
[4] L K M, GREGORY P, C E J, et al. Gaining insights from social media language: Methodologies and challenges [J]. Psychological methods, 2016, 21(4).
[5] NIELSEN M W, BöRJESON L. Gender diversity in the management field: Does it matter for research outcomes[J]. Research Policy, 2019, 48(7): 1617-1632.
[6] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet Allocation [J]. Journal of machine learning research, 2003, 3(4/5).
[7] JELODAR H, WANG Y, YUAN C, et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey [J]. Multimedia Tools and Applications, 2019, 78(11).
[8] ASMUSSEN C B, MøLLER C. Smart literature review: a practical topic modelling approach to exploratory literature review [J]. Journal of Big Data, 2019, 6(1).
[9] YANG J H, KINSHUK, AN Y. A survey of the literature: how scholars use text mining in Educational Studies[J]. Education and Information Technologies, 2023, 28(2): 2071-2090.
[10] ANTONS D, GRüNWALD E, CICHY P, et al. The application of text mining methods in innovation research: current state, evolution patterns, and development priorities [J]. R & D Management, 2020, 50(3): 329-351.
[11] FANG D, YANG H, GAO B, et al. Discovering research topics from library electronic references

using latent Dirichlet allocation [J]. Library Hi Tech, 2018, 36(3): 400-410.

[12] YIN B, YUAN C-H. Detecting latent topics and trends in blended learning using LDA topic modeling [J]. Education and Information Technologies, 2022, 27(9): 12689-12712.

[13] IVANOV D, TANG C S, DOLGUI A, et al. Researchers' perspectives on Industry 4.0: multi-disciplinary analysis and opportunities for operations management [J]. International Journal of Production Research, 2021, 59(7): 2055-2078.

[14] OGBUKE N J, YUSUF Y Y, DHARMA K, et al. Big data supply chain analytics: ethical, privacy and security challenges posed to business, industries and society [J]. Production Planning & Control, 2022, 33(2-3): 123-137.

[15] MAHESHWARI S, GAUTAM P, JAGGI C K. Role of Big Data Analytics in supply chain management: current trends and future perspectives [J]. International Journal of Production Research, 2021, 59(6): 1875-1900.

[16] CHOI T M, WALLACE S W, WANG Y L. Big Data Analytics in Operations Management [J]. Production and Operations Management, 2018, 27(10): 1868-1883.

[17] TALWAR S, KAUR P, WAMBA S F, et al. Big Data in operations and supply chain management: a systematic literature review and future research agenda [J]. International Journal of Production Research, 2021, 59(11): 3509-3534.

[18] XU J, PERO M, FABBRI M. Unfolding the link between big data analytics and supply chain planning [J]. Technological Forecasting and Social Change, 2023, 196: 122805.

[19] Sievert C, Shirley K E. LDAvis: A Method for Visualizing and Interpreting Topics [J]. Procedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, 2014: 63-70.

[20] LIGORIO L, VENTURELLI A, CAPUTO F. Tracing the boundaries between sustainable cities and cities for sustainable development. An LDA analysis of management studies [J]. Technological Forecasting and Social Change, 2022, 176.

[21] GöLZER P, FRITZSCHE A. Data-driven operations management: organisational implications of the digital transformation in industrial practice [J]. Production Planning & Control, 2017, 28(16).