applied sciences

MDPI

Article

A Suggestion on the LDA-Based Topic Modeling Technique Based on ElasticSearch for Indexing Academic Research Results

Mi Kim and Dosung Kim *

Department of Computer Science, Graduate School, Soongsil University, Seoul 06978, Korea; pytwoori@gmail.com
* Correspondence: k3510h@gmail.com; Tel.: +82-10-8700-5738

Abstract: Most academic researchers use the academic information system when they want to write a reference, such as a related research for a paper. Specific classification rules are applied based on vast amounts of data and the latest references to classify and search keywords. Meta information is designed for specific classification rules and search results are restructured. The search results can be classified and rearranged to suit academic research paper keywords by applying the restructured classification system and the LDA-based topic modeling technique. To implement this, the ElasticSearch classification method and topic-based LDA model were applied to extract the characteristics of academic papers in this study. Stable topics that could detect topic estimation and keyword search results within the minimum time were extracted to classify the paper search results. In addition, by analyzing the distribution of document weight among topics, the system performance was proven to be excellent.

Keywords: machine learning; natural language; text-mining; LDA model; ElasticSearch; meta-data

Citation: Kim, M.; Kim, D. A Suggestion on the LDA-Based Topic Modeling Technique Based on ElasticSearch for Indexing Academic Research Results. *Appl. Sci.* **2022**, *12*, 3118. https://doi.org/10.3390/app12063118

Academic Editors: Yosoon Choi and Manuel Armada

Received: 25 January 2022
Accepted: 11 March 2022
Published: 18 March 2022

1. Introduction

With the development of the latest information and communication technology in conjunction with the Fourth Industrial Revolution, there are many information searches through academic information-search portal sites that host mass information production. Accordingly, researchers use academic search engines such as the Google Scala to search for large amounts of information. The Google scholar search engine can be used to efficiently find more articles than SCI(E) even though this specific library search engine is available for searching papers. Topic modeling in the field of machine learning is a statistical model for discovering abstract topics and is one of the text-mining techniques used to discover the hidden semantic meaning structures of text [1].

Researchers use the topic modeling technique to classify abstracts through academic information searches and summarize them into suitable journals for further research [2,3]. Keyword searches through the academic information system infer potential topics and removes rare words of the topic. However, this has a disadvantage in that the classification system through morpheme is insufficient. As a way to solve these problems, a dictionary of rare words can be created focusing on the researcher's specific keyword and when requesting the literature information, the latent Dirichlet allocation (LDA) model can be used to categorize and present the search results based on the automatically inputted keywords. In addition, the academic research classification system that can understand the academic research flow using Google Scala can be automatically classified. For trend analysis such as the summary of collected academic papers, new technology trends, frequency by word, and similarity by search keyword can be calculated through word-cloud analysis. Through this, the characteristics and efficiency of new technology terms in the latest trend papers based on LDA modeling can be proven. Therefore, the classification system was redefined using the meta information for the searched keywords and summary in this study and

the search results of the extracted keywords are connected to meta information for natural language search and the classification system of academic papers using the ElasticSearch (ES) technique.

When academic researchers write their papers, they find references through Google Scala. The ES engine and LDA model were applied to classify the searched papers by category and they play a role as a useful paper writing tool when writing a paper. Through this, the purpose of this paper is to introduce a system that automatically classifies the relevant category when writing a key reference.

To organize the paper, Section 2 summarizes the topic modeling research cases and comparative analysis of other researchers through related studies. In Section 3, the topic modeling technique proposal for summarizing actual academic research papers and utilization methods related to ES are defined. In Section 4, the topic weight and distribution of academic research data classification and data analysis using LDA modeling are analyzed and the conclusion is drawn in Section 5.

2. Related Works

2.1. Classification Modeling for Summarizing Academic Information Research Papers

More than 400 domestic and foreign organizations and cooperative networks such as NDSL provide the latest information notification service for each individual, so that information resources can be utilized [4]. In addition, efficient modeling can save time by finding the necessary information for authors in academic research activities and provide an efficient method for obtaining information. The supervised learning method for classification based on the category of paper includes the Support Vector Machine (SVM) and Naïve Bayes (NB) classification for semi-supervised learning using the existing system [5,6]. These two methods showed higher classification performance than other supervised learning methods [7].

2.2. LDA Model for Improving the Limits of Supervised Learning

LDA topic modeling is one of the data-mining techniques, and is a model that infers latent topics based on unstructured text and discovers hidden semantic structures [8]. In addition to academic journals, the LDA model is useful for understanding the latent meaning of topics in various media such as SNS, Facebook, Twitter, and newspaper articles. LDA is a model-based estimation technique that makes it easy to identify how many latent classes are between topics. The topic modeling technique includes LDA and probabilistic latent semantic analysis (PLSA), which combines word distribution based on latent semantic analysis of words and topic distribution of documents [9,10]. In this study, search keywords are classified using ES [11,12], and the extracted summary and results are provided to the authors.

2.3. ElasticSearch Technique

ES is a text search project using open source, a JSON-based distributed search engine and an RDBMS-based distributed storage environment. ES is a search method that speeds up searches by first searching the reverse index [13,14]. EA is a distributed search engine developed by Shay Banon and was first announced in 2010. As a distributed system, EA has the advantage of being easy to respond to when the capacity of search target increases. It stores the actual data index and is divided into a primary shard and replica shard. The primary shard is the primary index that constitutes the shard and the replica shade is a replica of the primary shade stored in another distributed node [15]. EA can not only be executed in the Java development environment, but also input, delete, and search data through Restful API. Libraries such as Java, PHP, Perl, JavaScript, Python, and Ruby can be provided, and plug-in installation is also supported. In addition, one of the characteristics of EA is that various functions can be easily extended and installed.

2.4. Technique Related to Meta Information in the Academic Information Search Classification System

The special structural topic model (STM) of the LDA model uses a logistic normal distribution and assumes that the distribution of topics can be correlated. The defined document metadata with author information, thesis summary, and citation information is tested by a linear regression model (LRM) in which K topic prevalence is designated as an outcome variable and how much the information of metadata affects the possibility of a specific latent topic can be verified through statistical significance [16]. The introduction and communication method of data provider and service provider are as follows: The data provider retains the original, manages the information and provides metadata to the service provider by encoding it in XML [17,18]. The accuracy of syntax analysis can be improved by using the characteristics of agglutinative words formed by the combination of Korean syllables through the study of syntax analyzer using the LDA modeling-based morpheme-analysis for author information, summary, introduction, and citation [19–21].

2.5. Summary

In this chapter, the search technique for the paper search in KCI and SCI-registered journals using NDSL's domestic and overseas cooperation network of over the 400 institutions was reviewed. In particular, the ES engine technique and other techniques for re-establishing the meta-structure of paper summary using the indexing technique to summarize the paper search result were reviewed. Papers were arranged through LDA topic modeling which summarizes semantic analysis by classifying keywords of latent paper topics, and existing papers were studied to arrange by meta structure. In the case of Korean papers, citation information, introduction, and summary were arranged, and morpheme-analysis was performed using the LDA topic modeling. In particular, a LDA topic modeling that gave latent meaning by extracting related keywords except the agglutinative words, insoluble words, and banned words was proposed [22–24].

3. A Study on LDA Topic Modeling Technique Based on ES

3.1. Definition of Index Structure of Papers Based on ElasticSearch

In this paper, the ElasticSearch engine was used to search for potential paper keywords and store them in a distributed storage structure. In order to define the structure, it was proposed to index the citation information by classifying the collected paper keywords through the crawler into a meta structure. The indexing storage structure consisted of a primary shard structure and was determined when the index was first created. The number of replicas could be changed later. Figure 1 is a sample by indexing the primary shard structure.

```
primary shard 5, Replica 1 by Article/_bulk Indexing create
$ curl -XPUT "http://localhost:9200/books" -H 'Content-Type: application/json' -d'
{
  "settings": {
    "number_of_shards": 5,
    "number_of_replicas": 1
  }
}'
```

Figure 1. Primary shard structure.

The paper keyword storage index structure through ElasticSearch was indexed by classifying meta information into the type of paper for keyword storage. As shown in

Figure 2, the article structure was indexed by Journalsid, Scoups, KCI, Conference, Open Access, etc.

```
POST Article/_bulk
{"index":{"_id":1}}
{"model":"Journals","author":475,"date":"2021-01-24"}
{"index":{"_id":2}}
{"model":"Scopus ","author":795,"date":"2015-03-15"}
{"index":{"_id":3}}
{"model":"KCI ","author":859,"date":"2016-02-21"}
{"index":{"_id":4}}
{"model":"Conference","author":959,"date":"2017-03-29"}
{"index":{"_id":5}}
{"model":"Open Access","author":1059,"date":"2018-02-25"}
```

Figure 2. Article query index.

3.2. Keyword Classification through the LDA Modeling Definition

LDA topic modeling was applied to preprocess the text after storing the paper keyword index using ElasticSearch. The distribution values and word sets for the stored paper keywords and any topics in the Abstract were confirmed by applying the indexed paper storage structure to the topic model and the DT (dynamin topic) model and the result was given as a new topic through semantic reasoning. The number of topics was decided between 30 and 50. Morphological analysis of the selected topic was conducted to remove insoluble words and banned words and frequency analysis was conducted through morpheme analysis, word embedding based on vector, and word vector expression based on text and document. Finally, meaningful keyword separation was completed in the topical reasoning by the topic of search paper.

3.3. Analysis of Paper Search Keyword Trend through the DT Model

The keywords were classified by year and the paper trend was analyzed to determine the literature trend of the paper and academic search information. The number of topics was designated as 50, and 150 topics were selected by analyzing the data for 5 years. Figure 3 is a basic formula for analyzing the paper keyword trend of the DT model. The basic structure is as follows.

$$\alpha_t \sim \mathbf{N}\left(\alpha_t|\alpha_{t-1},\sigma^2 I\right)$$

$$\Phi_{k,t} \sim \mathbf{N}\left(\Phi_{k,t}|\Phi_{k,t-1},\beta^2 I\right)$$

$$\eta_{d,t} \sim \mathbf{N}\left(\eta_{d,t}|\alpha_t,\psi^2 I\right)$$

$$Z_{d,n,t} \sim \mathrm{Mult}\left(Z_{d,n,t}|\mathrm{softmax}(\eta_{d,t})\right)$$

$$W_{d,n,t} \sim \mathrm{Mult}\left(W_{d,n,t}|\mathrm{softmax}(\Phi_{Z_{d,n,t},t})\right)$$

Figure 3. DT model formula.

1. αtαt is calculated for T years. Φk,tΦk, and t is calculated for K subjects in T years. ηd, tηd, and t are calculated for all the literature d in T years.
2. A word is created about the literature d of t year as follows.

a. First, a topic k is determined. The topic k is calculated by the polynomial distribution softmax(ηd,t)softmax(ηd,t).
b. Then, a word w is calculated using the calculated topic k. The word w is calculated from the polynomial distribution softmax(Φk,t)softmax(Φk,t).
c. The calculated w is written. This process goes back to a and repeats.

4. Data Analysis

4.1. Analysis of Search Results

This chapter describes the data analysis results. The results of the conducted paper search through the ElasticSearch base were stored as index. The analysis environment was tested in Jupyter notebook using Anaconda 3.0. The title and abstract of search results through the Google Scala, RISS, NTIS, etc., were calculated and recorded in text form using the stored index. Figure 4 shows the results of applying the pyLDAvis library to 50 topics calculated through the LDA model and the DT model based on the original source data of text type. In the first index, keywords such as AI-based CNN and deep learning were indexed. The most frequent words appeared in the order of Image, Detection, and Recognition.

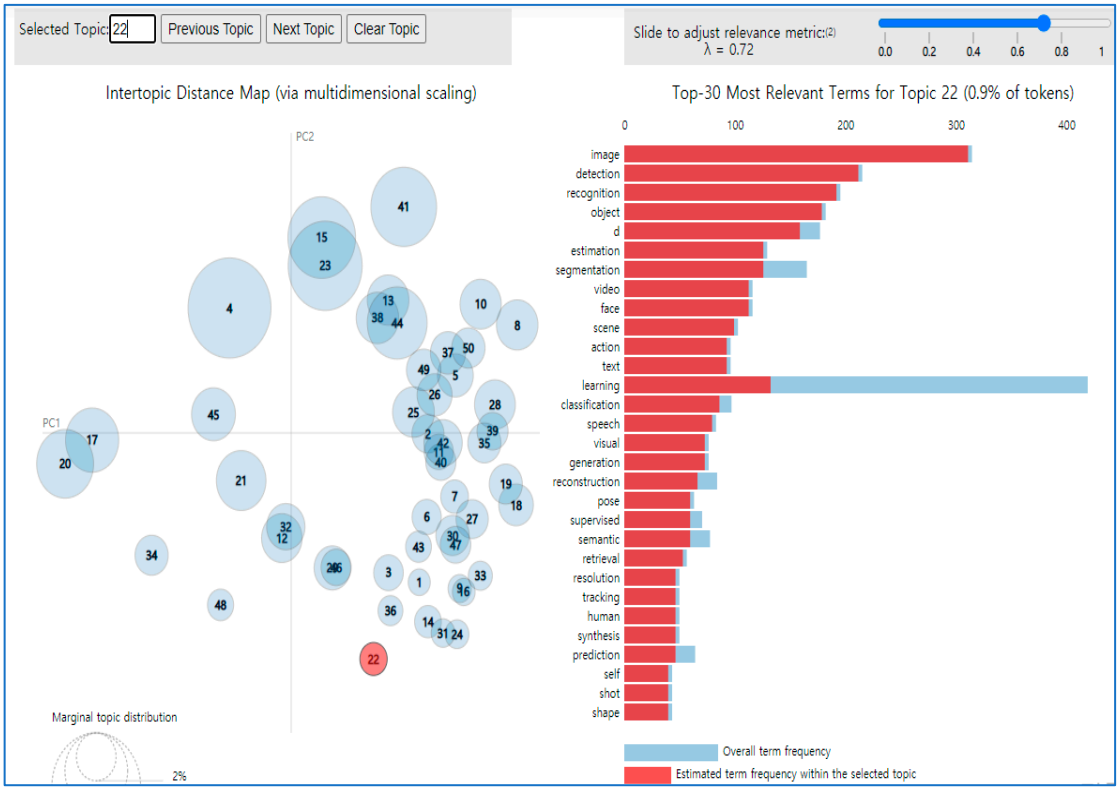

Figure 4. Topic modeling analysis.

As shown in Figure 5, the keywords except for insoluble words, and banned words were extracted as separate texts and extracted as a word cloud, a representative word technique in the IT deep-learning field. The word-cloud analysis was conducted based on ghostwhite background color because the word cloud can be very useful of representing frequent displayed keywords as the LDA chart that extracts detailed topics. Therefore, authors can find easier topics for their study when they write papers.

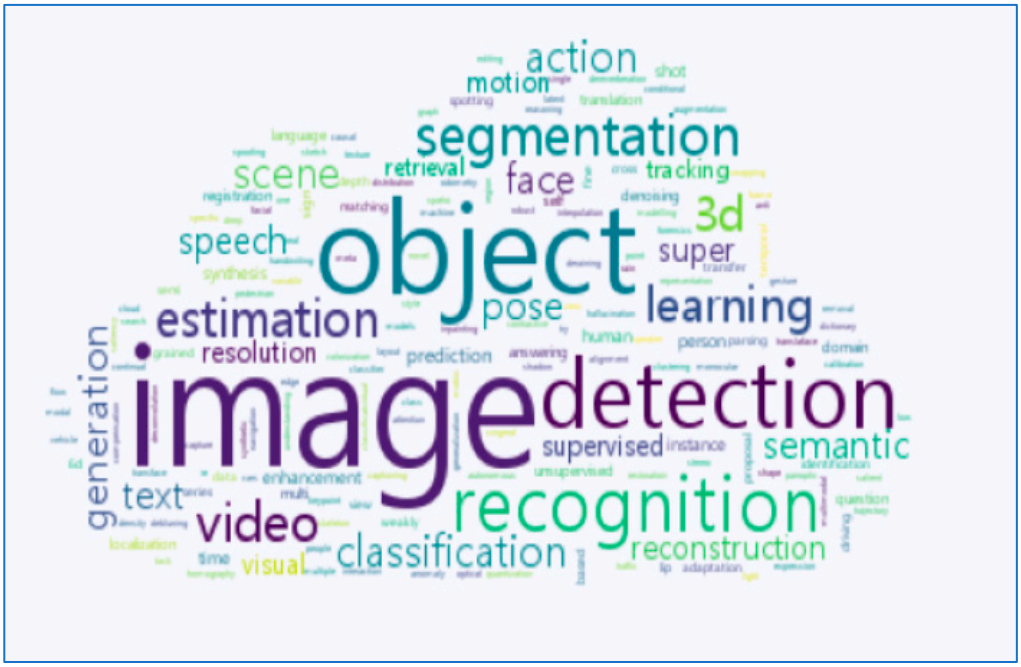

Figure 5. Main keyword word-cloud.

The meaning of topics can be inferred for LDA topic modeling and the topic weight among keywords through the paper research is shown. The total number of topics was 30, and topics such as classification, predictive analysis, and meaning were derived from topic 17 based on topics that appeared frequently, as shown in Table 1. As shown in Figure 6, the result of the tree-structure map was calculated through Jupyter notebook based on the derived values in Table 1. The average distribution map of the main topics derived from Figure 6 is shown in Figure 7. Figure 8 shows the trends of corresponding potential keyword classification of the topic "System" with high frequency. Figure 9 shows the frequency tracking of OBJECT topic, which has the highest weight among the top 20 topics.

Table 1. Result of weight analysis among LDA Topics.

Topic	Weight	Main Topics
Analysis	0.21%	Learning
Correspondence	0.363%	Classification
Word	0.144%	Model
Data	0.117%	Analysis
Emotion	0.0876%	Prediction
Engine	0.0518%	Frequency
CMF	0.0415	Equipment

Figure 6. DT-model tree map.

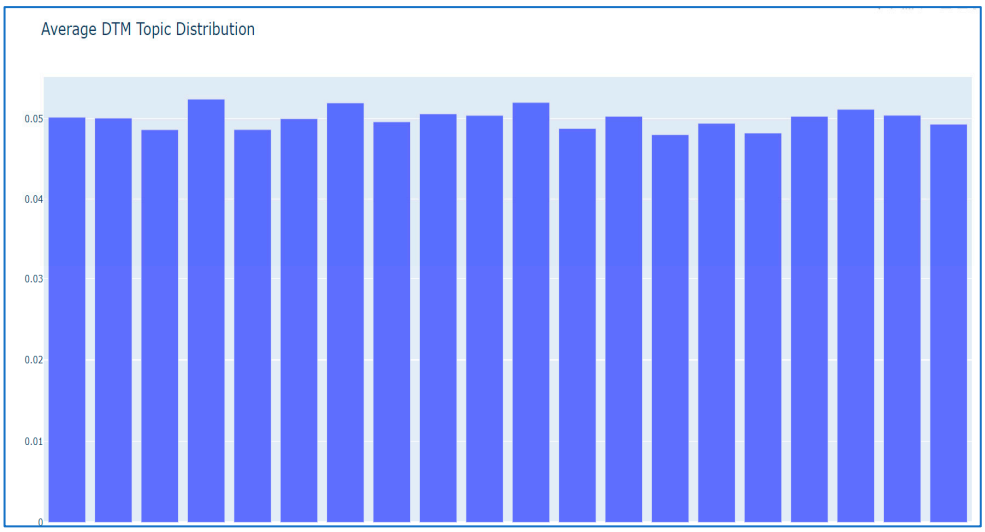

Figure 7. DTM topic average distribution map.

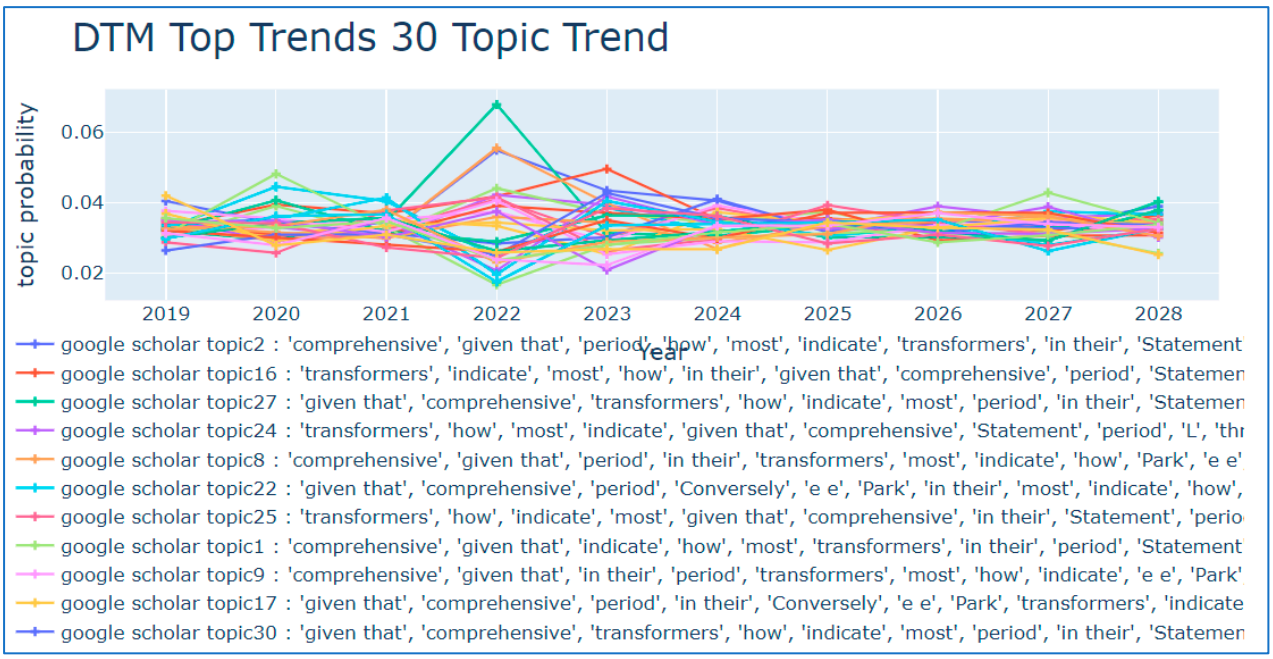

Figure 8. System topic probability.

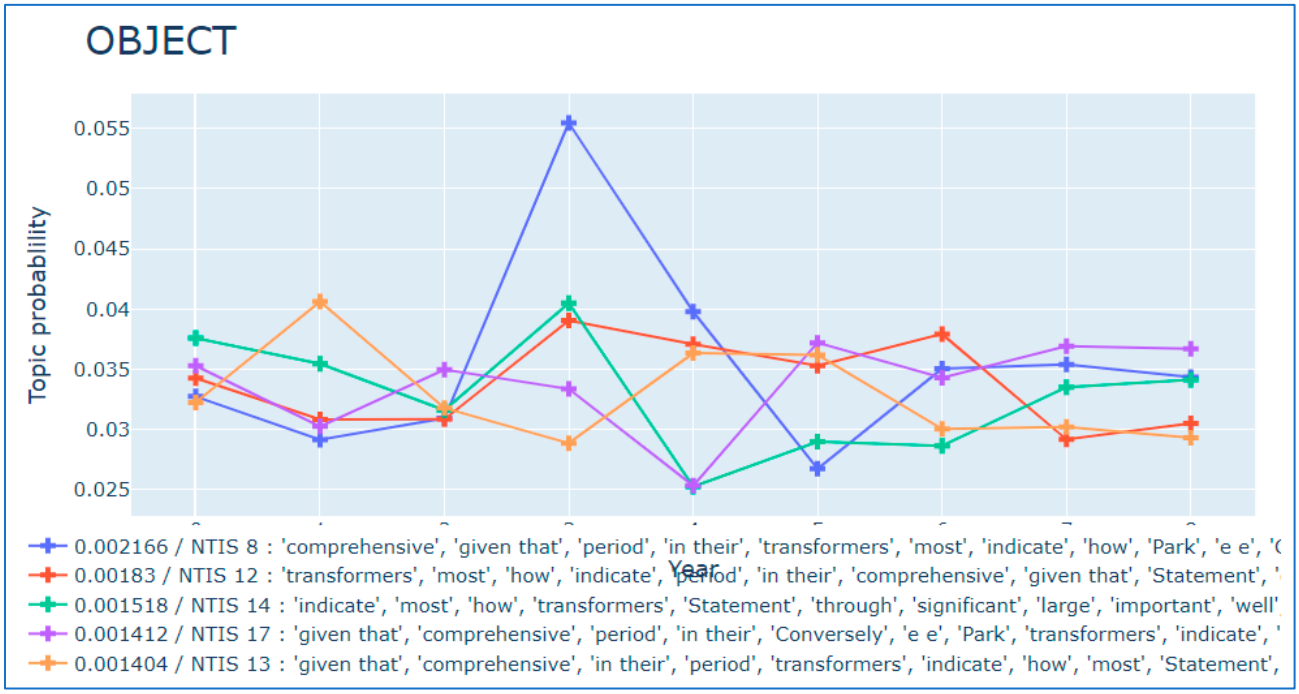

Figure 9. DTM OBJECT topic probability.

4.2. Experiment Result

As an application method through the search results in the Section 4.1, the ES index was registered through search keywords such as deep learning, CNN, RNN, and DNN for the search word using the ES engine. For the presented topics, the topic weight among keywords was analyzed through the LDA and DT models. The correlation analysis of semantic keywords automatically extracted through keyword semantic inference among the topics was conducted. In addition, the significant probability values of the extracted topics for each specific word were calculated through the tree map in order to check the correlation between words and words, documents and documents. Through the presented technique in this study, IT researchers can proceed with the interpretation of related words and meaningful topic terms that can be referenced by topic when writing a paper. Through this, semantic similarity can be applied through keyword related word extraction based on a new paper-search engine.

4.3. Discussion

In this section, the search technique for the paper search in KCI and SCI-registered journals using NDSL's domestic and overseas cooperation network of over the 400 institutions is reviewed. In particular, the ES engine technique and other techniques for re-establishing the meta-structure of paper summary using the indexing technique are reviewed to summarize the paper search result. The papers were arranged through LDA topic modeling which summarizes semantic analysis by classifying keywords of latent paper topics, and existing papers were studied to arrange meta structure. In the case of Korean papers, citation information, introductions, and summaries were arranged in this paper and morpheme-analysis performed using the LDA topic modeling. In particular, an LDA topic modeling that gives latent meaning by extracting related keywords except the agglutinative words, insoluble words, and banned words was proposed. Citation information, introductions, and summaries were arranged in the case of research papers and morphological syntax analysis was performed using the LDA topic modeling. Researchers can apply it as a support tool through a meaningful probability distribution when writing a paper.

5. Conclusions

All academic researchers use academic information systems to write references for related research. Many academic researchers use Google Scala to search through references when writing their papers. The ES engine and LDA model can be applied as a useful paper-writing tools to classify the searched papers by category and write the paper. Through this, the purpose of this paper was to introduce a system that automatically classifies the relevant category when writing a key reference. Through the experiment, various core topics were derived from the keywords that came out through the re-indexing of the collected keywords using the elastic search engine as the LDA model. The derived keywords can be categorized into the category of reference and can be used as a tool to help authors write their paper more easily.

Specific classification rules are applied when searching for vast amounts of data, up-to-date references, and classified keywords. For this, meta information should be designed and search results can be reorganized. The search results can be properly organized for the academic research papers by using the LDA-based topic modeling technique based on the restructured classification system. To apply this, the topic weight of search keywords was analyzed through the ES technique and topic-based LDA model that extracts the characteristics of academic papers in this study. As a result of the analysis, topic estimation and keyword search results could be detected in a shorter time to classify the paper search results. Therefore, the distribution of document weight among the stable topics could be analyzed. In addition, the experimental results and environment could be provided to establish the related words and categories.

The topics of paper search keywords could be estimated through the ES model and the LDA model. The model of this study was able to analyze the semantic similarity and correlation between keywords. The category of paper-search keywords could be structured through the meta information by analyzing the average of topic weight and distribution. In addition, the model of this study can be easily applied through the analysis of related keywords, similarity between keywords, and average weight of paper keywords in research papers that can represent predictive models using the researcher's well-arranged interests. The utilization of ES can be further expanded by inferring topics of categories through various news searches, issue searches, and scientific information searches. However, further research on the automatic classification and inference of paper keywords based on the inferred topics from the meaning of academic information keywords will be required.

Author Contributions: Conceptualization, M.K. and D.K.; methodology, M.K.; software, M.K.; validation, D.K. and M.K.; formal analysis, M.K.; investigation, M.K.; resources, M.K.; data curation, M.K.; writing—original draft preparation, M.K.; writing—review and editing, D.K. and M.K.; visualization, M.K.; supervision, D.K.; project administration, M.K.; funding acquisition, M.K. and D.K.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wu, Z.; Lei, L.; Li, G.; Huang, H.; Zheng, C.; Chen, E.; Xu, G. A Topic Modeling Based Approach to Novel Document Automatic Summarization. *Expert Syst. Appl.* **2017**, *84*, 12–23. [CrossRef]
2. Fiandrino, S.; Tonelli, A. A Text-Mining Analysis on the Review of the Non-Financial Reporting Directive: Bringing Value Creation for Stakeholders into Accounting. *Sustainability* **2021**, *13*, 763. [CrossRef]
3. Ammirato, S.; Felicetti, A.M.; Raso, C.; Pansera, B.A.; Violi, A. Agritourism and Sustainability: What We Can Learn from a Systematic Literature Review. *Sustainability* **2020**, *12*, 9575. [CrossRef]
4. Mustafa, M.; Zeng, F.; Ghulam, H.; Muhammad Arslan, H. Urdu Documents Clustering with Unsupervised and Semi-Supervised Probabilistic Topic Modeling. *Information* **2020**, *11*, 518. [CrossRef]
5. Wahid, J.A.; Shi, L.; Gao, Y.; Yang, B.; Tao, Y.; Wei, L.; Hussain, S. Identifying and Characterizing the Propagation Scale of Covid-19 Situational Information on Twitter: A Hybrid Text Analytic Approach. *Appl. Sci.* **2021**, *11*, 6526. [CrossRef]

6. Tharakan, R.A.; Joshi, R.; Ravindran, G.; Jayapandian, N. Machine Learning Approach for Automatic Solar Panel Direction by using Naïve Bayes Algorithm. In Proceedings of the 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 6–8 May 2021; pp. 1317–1322.
7. Kim, P.-J.; Lee, J.-Y. Utilizing Unlabeled Documents in Automatic Classification with Inter-Document Similarities. *J. Korean Soc. Inf. Manag.* **2007**, *24*, 251–271. [CrossRef]
8. Cheng, Q.; Kang, J.; Lin, M. Understanding the Evolution of Government Attention in Response to COVID-19 in China: A Topic Modeling Approach. *Healthcare* **2021**, *9*, 898. [CrossRef] [PubMed]
9. Hofmann, T. Probabilistic Latent Semantic Indexing. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '99, Berkeley, CA, USA, 15–19 August 1999; ACM: New York, NY, USA, 1999; pp. 50–57.
10. Koltcov, S.; Ignatenko, V. Renormalization Analysis of Topic Models. *Entropy* **2020**, *22*, 556. [CrossRef] [PubMed]
11. Bendechache, M.; Svorobej, S.; Endo, P.T.; Mihai, A.; Lynn, T. Simulating and Evaluating a Real-World Elasticsearch System Using the Recap Des Simulator. *Futur. Internet* **2021**, *13*, 83. [CrossRef]
12. Qin, L.; Sun, Q.; Wang, Y.; Wu, K.F.; Chen, M.; Shia, B.C.; Wu, S.Y. Prediction of Number of Cases of 2019 Novel Coronavirus (COVID-19) Using Social Media Search Index. *Int. J. Environ. Res. Public Health* **2020**, *17*, 2365. [CrossRef] [PubMed]
13. Abayomi-Alli, A.; Abayomi-Alli, O.; Misra, S.; Fernandez-Sanz, L. Study of the Yahoo-Yahoo Hash-Tag Tweets Using Sentiment Analysis and Opinion Mining Algorithms. *Information* **2022**, *13*, 152. [CrossRef]
14. Shang, Z.; Luo, J.M. Topic Modeling for Hiking Trail Online Reviews: Analysis of the Mutianyu Great Wall. *Sustainability* **2022**, *14*, 3246. [CrossRef]
15. Murakami, R.; Chakraborty, B. Investigating the Efficient Use of Word Embedding with Neural-Topic Models for Interpretable Topics from Short Texts. *Sensors* **2022**, *22*, 852. [CrossRef] [PubMed]
16. Elasticsearch. Available online: https://www.elastic.co/kr/elasticsearch (accessed on 12 March 2020).
17. Park, J.; Cho, W.; Kim, K. Anomaly Detection Analysis Using Repository Based on Inverted Index. *J. KIISE* **2018**, *45*, 294–302. [CrossRef]
18. Farkhod, A.; Abdusalomov, A.; Makhmudov, F.; Cho, Y.I. LDA-Based Topic Modeling Sentiment Analysis Using Topic/Document/Sentence (TDS) Model. *Appl. Sci.* **2021**, *11*, 11091. [CrossRef]
19. Ingram, C.; Downey, V.; Roe, M.; Chen, Y.; Archibald, M.; Kallas, K.A.; Kumar, J.; Naughton, P.; Uteh, C.O.; Rojas-Chaves, A.; et al. COVID-19 Prevention and Control Measures in Workplace Settings: A Rapid Review and Meta-Analysis. *Int. J. Environ. Res. Public Health* **2021**, *18*, 7847. [CrossRef] [PubMed]
20. Lee, S. A Study on the OAI based Open Digital Library. *J. Inf. Manag.* **2004**, *35*, 139–159.
21. McDonald, R.; Nivre, J.; Quirmbach-Brundage, Y.; Goldberg, Y.; Das, D.; Ganchev, K.; Hall, K.; Petrov, S.; Zhang, H.; Täckström, O.; et al. Universal Dependency Annotation for Multilingual Parsing. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, 4–9 August 2013.
22. Huang, H.-L.; Lin, S.-J.; Hsu, M.-F. An Advanced Decision Making Framework via Joint Utilization of Context-Dependent Data Envelopment Analysis and Sentimental Messages. *Axioms* **2021**, *10*, 179. [CrossRef]
23. Li, C.; Liu, Z.; Shi, R. A Bibliometric Analysis of 14,822 Researches on Myocardial Reperfusion Injury by Machine Learning. *Int. J. Environ. Res. Public Health* **2021**, *18*, 8231. [CrossRef] [PubMed]
24. Truică, C.-O.; Apostol, E.-S.; Șerban, M.-L.; Paschke, A. Topic-Based Document-Level Sentiment Analysis Using Contextual Cues. *Mathematics* **2021**, *9*, 2722. [CrossRef]