

Optimizing Information Retrieval in Dark Web Academic Literature: A Study Using KeyBERT for Keyword Extraction and Clustering

Yosua Setyawan Soekamto^{1,6}, Leonard Christopher Limanjaya², Yoshua Kaleb Purwanto²,
Bongjun Choi³, Seung-Keun Song⁴, Dae-Ki Kang^{5,*}

¹PhD Student at Department of Computer Engineering, Dongseo University, Busan, South Korea

²Master Student at Department of Computer Engineering, Dongseo University, Busan, South Korea

³Professor at Department of Software, Dongseo University, Busan, South Korea

⁴Professor at Department of Visual Contents, Graduate School, Dongseo University, Busan, South Korea

⁵Professor at Department of Computer Engineering, Dongseo University, Busan, South Korea

⁶Lecturer at Department of Information Systems, Universitas Ciputra Surabaya

yosua.soekamto@ciputra.ac.id, leonardchristopher002@gmail.com, yoshuakaleb049@gmail.com,
bjchoi@gdsu.dongseo.ac.kr, songsk@gdsu.dongseo.ac.kr, *dkkang@dongseo.ac.kr

Abstract

The exponential increase in publications and the interconnected nature of sub-domains make traditional methods of information extraction and organization inadequate. This inefficiency can impede scientific progress and innovation. To address these challenges, this research leverages the ability of Bidirectional Encoder Representations from Transformers for keyword extraction (KeyBERT) and integrates with K-Means clustering to organize topics from large datasets effectively. Analyzing a dataset of 47,627 articles from SCOPUS in the domains of Reinforcement Learning and Computer Vision. An ablation study demonstrates the generalizability of the approach across these fields, with the optimal number of clusters determined to be three using the Elbow Method. The results demonstrate that KeyBERT is effective in extracting and organizing topics within these domains, with a particular focus on applications such as medical imaging, autonomous driving, and real-time detection systems. This methodology offers a scalable solution for organizing vast academic datasets, enabling researchers to extract meaningful insights efficiently and apply this approach to other domains.

Keywords: K-Means, KeyBERT, Keyword Extraction, Text Mining, Topic Clustering.

1. Introduction

In today's fast-paced world, the volume of information and technological advancements is growing at an

Manuscript Received: September. 26, 2024 / Revised: October. 2, 2024 / Accepted: October. 7, 2024

Corresponding Author: dkkang@dongseo.ac.kr

Tel: +82-51-320-1724, Fax: +82-51-327-8955

Professor, Department of Computer Engineering, Dongseo University, Busan, Korea

unprecedented rate. The explosion of digital data has transformed various fields, especially in technology and information sciences. This rapid growth is fueled by the continuous development and integration of technologies like artificial intelligence (AI), machine learning (ML), and big data analytics. Researchers and professionals across the globe are generating a vast number of academic papers, technical reports, and datasets daily, which are essential for innovation and development.

However, this abundance of information presents a significant challenge: it becomes increasingly difficult for researchers and practitioners to keep up with the latest developments within their specific domains. The complexity is further compounded by the interconnected nature of various sub-domains, making it challenging to track, organize, and interpret the vast sea of knowledge efficiently [1].

Given the exponential increase in academic publications and the interwoven nature of different sub-domains, traditional methods of information extraction and organization are often inadequate. Traditional information extraction methods, such as Term Frequency-Inverse Document Frequency (TF-IDF) and Latent Dirichlet Allocation (LDA)[2], [3], are widely used for keyword extraction and topic modeling. These methods struggle to effectively capture the nuanced relationships between terms and topics within a specific field. This is particularly evident in fields such as computer science, where domains like reinforcement learning and computer vision are constantly evolving and expanding.

The inability to efficiently extract and organize relevant information can hinder scientific progress and innovation. Therefore, there is a pressing need for advanced methodologies that can automate and enhance the process of information extraction, allowing researchers to stay informed and make meaningful connections between various sub-domains[4], [5]. To address these challenges, this research proposes the use of the power of BERT (Bidirectional Encoder Representations from Transformers) to provide more accurate and contextually relevant keywords from textual data. This research addresses the following contributions:

- Proposing information extraction approach by combining keywords extraction by BERT algorithm (KeyBERT) and K-Means clustering to organize topics from large datasets.
- Research ablation study to show generalizability of the proposed method, by showing the comparison between two topics (reinforcement learning and computer vision).

2. Literature Review

2.1 Keyword Extraction with BERT

Keyword extraction is a fundamental task in natural language processing (NLP) that involves identifying significant terms or phrases within a text [6]. The advent of transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) has revolutionized this field [7]. BERT's ability to understand the context of words in a bidirectional manner makes it exceptionally effective for keyword extraction. KeyBERT, an extension of BERT, leverages this capability by generating embeddings for the text and extracting keywords that are contextually relevant. Studies have shown that KeyBERT can significantly improve the quality of keyword extraction by capturing the nuanced meanings of words within their specific contexts [8].

Moreover, recent advancements like AdaptKeyBERT and LLM-TAKE demonstrate how BERT's embeddings can be fine-tuned or adapted for domain-specific applications, further enhancing the precision and relevance of extracted keywords [9], [10]. These developments highlight the critical role of BERT-based models in advancing keyword extraction methodologies.

2.2 Topic Clustering

Topic clustering is an essential technique in text mining and NLP that involves grouping similar texts or keywords into coherent topics. This process helps in uncovering the underlying structure of large text corpora and identifying prevalent themes. Unsupervised learning algorithms, such as K-Means and hierarchical clustering, are commonly used for this purpose [11]. By clustering keywords extracted from texts, researchers can identify major topics and sub-domains within a dataset.

The integration of BERT-based models like KeyBERT with clustering algorithms has been shown to enhance the coherence and interpretability of the clusters [12]. For instance, in this study, keywords extracted from academic articles on Reinforcement Learning and Computer Vision were clustered to reveal prominent research themes and emerging trends [13]. Visualization tools, such as word clouds and cluster maps, further aid in interpreting these clusters by providing a visual representation of keyword frequencies and relationships. The continuous evolution of clustering techniques and their application in various domains underscore their importance in extracting meaningful insights from vast amounts of textual data [14].

3. Methodology

3.1 Data Collection and Preprocessing

The data for this study was collected from the SCOPUS database, a comprehensive source of academic research articles. The collecting process uses SCOPUS API-Key and Python Beautiful-Soup library. SCOPUS provides abstracts and citation information from a wide range of publishers, making it an ideal source for our analysis. The dataset spans from 2012 to 2023, focusing on articles related to "Reinforcement Learning" and "Computer Vision." A total of 47,627 articles were included in the dataset, encompassing keywords such as "Image Recognition," "Object Detection," "Pattern Recognition," "Image Classification," and "Real-Time Detection." This extensive collection of abstracts served as the primary text data for subsequent keyword extraction and topic clustering.

Before performing keyword extraction, the collected abstracts underwent preprocessing to enhance the quality of the text data. This involved the removal of custom stop words, which included both common English stop words and domain-specific terms that could skew the results. This step ensured that the text data was refined and relevant, improving the accuracy of keyword extraction. The preprocessing phase is crucial as it prepares the data by eliminating noise and irrelevant information, thus allowing for more precise and meaningful keyword extraction.

3.2 Keyword Extraction and Topic Clustering

The core of our methodology involved the use of KeyBERT for keyword extraction. KeyBERT, a keyword extraction tool based on BERT embeddings, was employed to identify and extract keywords from the abstracts. The process began by embedding the text using BERT, after which KeyBERT extracted keywords that were most representative of the content. This resulted in the extraction of 148,393 keywords from the dataset. The top-used keywords included terms such as "recognition," "image," "detection," "segmentation," and "classification." KeyBERT's capability to leverage BERT embeddings ensured that the extracted keywords were contextually relevant and representative of the underlying research topics.

Following keyword extraction, the next phase involved clustering these keywords to identify prevalent topics and sub-domains within the dataset. Unsupervised clustering algorithms, such as K-Means, were utilized to group similar keywords together. This approach facilitated the identification of coherent topics based on

keyword similarities. Visualization tools, including word clouds, were employed to represent the frequency and relationships between keywords within these clusters. An analysis of the clusters was conducted to find centroids and inertia between words, explaining relationships and prominent topics such as "Segmentation," "Human," and "Health." This step provided a clear and visual understanding of the main research areas and emerging topics in the domains of Reinforcement Learning and Computer Vision.

4. Result and Discussion

This study aims to measure the capability of KeyBERT in identifying relevant keywords in each domain. Therefore, experiments were conducted separately for each domain. The experimental results indicated that the optimal number of clusters is three in each domain. The search for the best number of clusters was performed using the elbow method, as shown in Figure 1. After determining the number of clusters, the research continued by grouping the documents based on the keywords obtained by KeyBERT.

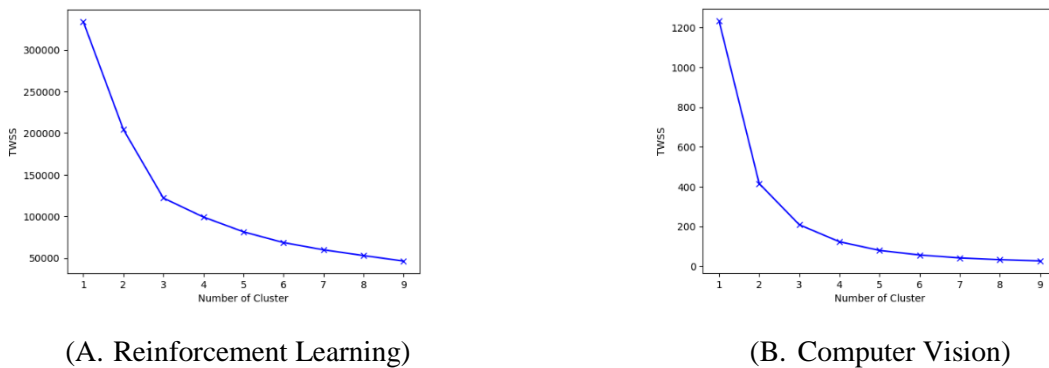
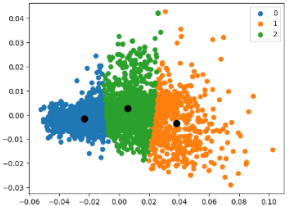
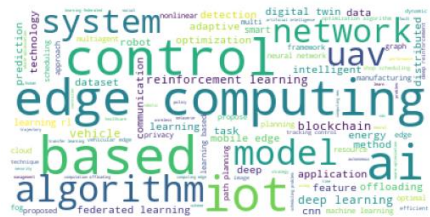
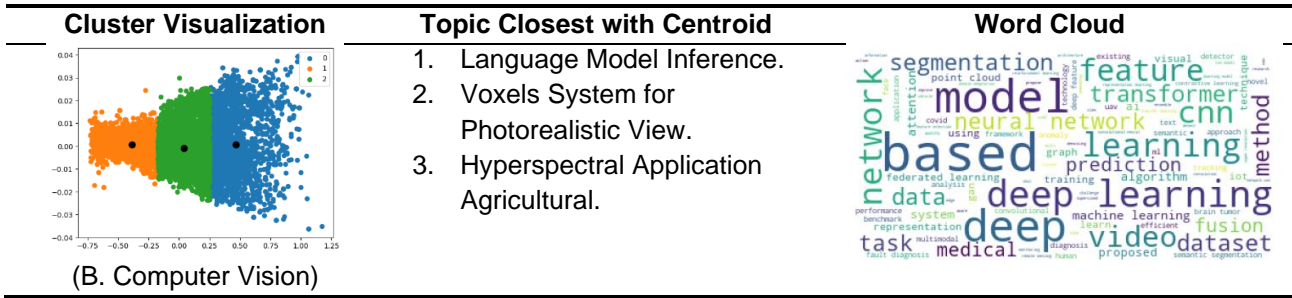


Figure 1. Best N-Cluster Using Elbow Method

The experimental results can be seen in Table 1. This study found that within the topic of reinforcement learning, the subtopics closest to the centroid are multimodal retrieval imagery, swarm optimization algorithm application, and fuzzy consensus tracking controller. Meanwhile, within the topic of computer vision, the subtopics closest to the centroid are language model inference, voxels system for photorealistic view, and hyperspectral application in agriculture.

Table 1. Experiment Result

Cluster Visualization	Topic Closest with Centroid	Word Cloud
 <p>(A. Reinforcement Learning)</p>	<ol style="list-style-type: none"> Multimodal Retrieval Imagery Swarm Optimization Algorithm Application Fuzzy Consensus Tracking Controller 	



To illustrate the keywords formed from the dataset, this study used word cloud visualization. The word cloud results show that keywords in the reinforcement learning domain are mostly related to “control”, “UAV”, “IoT”, “AI”, and “blockchain.” In the computer vision domain, the prominent keywords are “segmentation”, “transformer”, “feature”, “medical”, “prediction”, “CNN”, “point cloud”, and “federated learning.”

5. Conclusion

The study demonstrated the effectiveness of using KeyBERT for keyword extraction and K-Means clustering for topic identification in a large dataset of academic articles. By leveraging these techniques, we were able to extract meaningful keywords and identify key research themes and emerging trends in the fields of Reinforcement Learning and Computer Vision. The findings highlight the growing focus on applications such as medical imaging, autonomous driving, and real-time detection systems. These insights are valuable for researchers and practitioners, providing a comprehensive overview of the current state and future directions in these rapidly evolving domains.

Applied	<ul style="list-style-type: none"> Image classification Surface Detection/Roughness Fingerprint/Pattern Recognition Object Detection/Tracker 	<ul style="list-style-type: none"> Image Classification Object Detection/Tracker Person Identification/Pattern Recognition Video Recognition 	<ul style="list-style-type: none"> Object Detection/Tracker Biomedical Imaging/Retinal/ECG Signal Video Recognition/Emotion Recognition 	<ul style="list-style-type: none"> Model Mapping/CNN Vehicle Tracking/Video Recognition Pattern Recognition Fusion Model/Multi Models 	<ul style="list-style-type: none"> Image Enhancement Robot Automation Surveillance Processing/Surface Detection UAV Automation/Video Recognition Ensemble Computer Vision 	<ul style="list-style-type: none"> Traffic Prediction/CCTV Recognition Ultrasound Image Processing/Surface Detection UAV Automation/Video Recognition
Enhancement	<ul style="list-style-type: none"> Kernel Scaling Saliency Model Image Segmentation Point Edge Matching 	<ul style="list-style-type: none"> Deep Dense Face Detector Video Enhancement Image Enhancement (Remove Blur) Background Modeling Dehazing Image Processing 	<ul style="list-style-type: none"> Saliency Model Spatial Temporal Feature Video Super-Resolution Adversarial Boosting AutoML 	<ul style="list-style-type: none"> Similarity Measurement Method Improvement Computational Cost Enhancement Transfer Motion Pose Object Joint Enhancement (Hand Object Manipulation) 3D Shape Representation 	<ul style="list-style-type: none"> Semantic Segmentation Improvement Part Aware Transformer Re-identification Masked Language Modeling VIT 	<ul style="list-style-type: none"> Hyperspectral Imagery (Transformer) Photorealistic (3D Reconstruction) GAN in Diffusion Joint Tracking Improvement

(A. Reinforcement Learning)

Applied	<ul style="list-style-type: none"> Robot Path / Direction Problem 	<ul style="list-style-type: none"> Robot Path / Direction Problem Image / Video Classification 	<ul style="list-style-type: none"> Eye Tracking EEO Recording PCMDP Hardware Analysis 	<ul style="list-style-type: none"> News / Text Summarization Image / Video Path / Direction Problem (Vehicle) QnA / Recommend System Simulation / Scheduler 	<ul style="list-style-type: none"> Path / Direction Problem (Navigation) QnA / Recommend System Anomaly / Phishing Detection Robot (Sensor) 	<ul style="list-style-type: none"> Robot / Video Core Biometric Recognition Trajectory UAV Path
Enhancement	<ul style="list-style-type: none"> Policy Optimization Algorithm Memory Usage 	<ul style="list-style-type: none"> Policy Optimization Algorithm (Evolutionary) Control Online / Offline Process Activity / Task 	<ul style="list-style-type: none"> Policy Optimization Algorithm Parameters Kernel Online / Offline Process Pruning Method Planning Abstraction Inverse Model Network Architecture 	<ul style="list-style-type: none"> Policy (Navigation, Sarsa) Optimization Algorithm (Pre-Trained) Control (Trained) Network Architecture Provisioning 	<ul style="list-style-type: none"> Control (Quantum) Actor / Critic Activity / Task (Pre-Trained) Intelligent Spectrum Management Network Architecture 	<ul style="list-style-type: none"> Optimization Algorithm (Swarm) Control (Fuzzy Consensus Tracking, Mobility, Hybrid Disassembly) Network Architecture Trajectory Self-Supervised Multimodal Imaginary

(B. Computer Vision)

Figure 2. Chart of Development of Topics

The methodology presented in this study, combining advanced keyword extraction with topic clustering, can be applied to other domains to uncover hidden patterns and trends in large text corpora. Future work could involve integrating more sophisticated clustering algorithms and exploring additional visualization techniques to further enhance the interpretability and applicability of the results [15]. For better understanding, this research also provides an illustration of the development of topics in reinforcement learning and computer vision, as shown in Figure 2.

Acknowledgement

This research was supported by 'The Construction Project for Regional Base Information Security Cluster', a grant funded by the Ministry of Science, ICT and Busan Metropolitan City in 2024.

References

- [1] S. Sun, Z. Liu, C. Xiong, Z. Liu, and J. Bao, "Capturing Global Informativeness in Open Domain Keyphrase Extraction," Apr. 2020, [Online]. Available: <http://arxiv.org/abs/2004.13639>
- [2] T. Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization."
- [3] D. M. Blei, A. Y. Ng, and J. B. Edu, "Latent Dirichlet Allocation Michael I. Jordan," 2003.
- [4] M. Basaldella, E. Antolli, G. Serra, and C. Tasso, "Bidirectional LSTM recurrent neural network for keyphrase extraction," in *Communications in Computer and Information Science*, Springer Verlag, 2018, pp. 180–187. doi: 10.1007/978-3-319-73165-0_18.
- [5] Y. Zhang, M. Tuo, Q. Yin, L. Qi, X. Wang, and T. Liu, "Keywords extraction with deep neural network model," *Neurocomputing*, vol. 383, pp. 113–121, Mar. 2020, doi: 10.1016/j.neucom.2019.11.083.
- [6] T. Nomoto, "Keyword Extraction: A Modern Perspective," *SN Comput Sci*, vol. 4, no. 1, Jan. 2023, doi: 10.1007/s42979-022-01481-7.
- [7] P. Sharma and Y. Li, "Self-supervised Contextual Keyword and Keyphrase Retrieval with Self-Labeling," 2019, doi: 10.20944/preprints201908.0073.v1.
- [8] C. Yoo and H. Lee, "Improving Abstractive Dialogue Summarization Using Keyword Extraction," *Applied Sciences (Switzerland)*, vol. 13, no. 17, Sep. 2023, doi: 10.3390/app13179771.
- [9] A. Priyanshu and S. Vijay, "AdaptKeyBERT: An Attention-Based approach towards Few-Shot & Zero-Shot Domain Adaptation of KeyBERT," Nov. 2022, [Online]. Available: <http://arxiv.org/abs/2211.07499>
- [10] R. Y. Maragheh *et al.*, "LLM-TAKE: Theme Aware Keyword Extraction Using Large Language Models," Dec. 2023, [Online]. Available: <http://arxiv.org/abs/2312.00909>
- [11] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," 2000.
- [12] L. George and P. Sumathy, "An integrated clustering and BERT framework for improved topic modeling," *International Journal of Information Technology (Singapore)*, vol. 15, no. 4, pp. 2187–2195, Apr. 2023, doi: 10.1007/s41870-023-01268-w.
- [13] S. Syed and M. Spruit, "Full-Text or abstract? Examining topic coherence scores using latent dirichlet allocation," in *Proceedings - 2017 International Conference on Data Science and Advanced Analytics, DSAA 2017*, Institute of Electrical and Electronics Engineers Inc., Jul. 2017, pp. 165–174. doi: 10.1109/DSAA.2017.61.
- [14] Q. Xie and L. Waltman, "A comparison of citation-based clustering and topic modeling for science mapping," 2023. doi: <https://doi.org/10.48550/arXiv.2309.06160>.
- [15] D. Sharma, B. Kumar, and S. Chand, "A Trend Analysis of Machine Learning Research with Topic Models and Mann-Kendall Test," *International Journal of Intelligent Systems and Applications*, vol. 11, no. 2, pp. 70–82, Feb. 2019, doi: 10.5815/ijisa.2019.02.08.