IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 25, 2022, accepted March 14, 2022, date of publication March 21, 2022, date of current version March 29, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3160826

Development of an Embedding Framework for Clustering Scientific Papers

SONGHEE KIM, SUYEONG LEE, AND BYUNGUN YOON, (Senior Member, IEEE)
Department of Industrial and Systems Engineering, Dongguk University, Jung-gu, Seoul 100-715, South Korea

Corresponding author: Byungun Yoon (postman3@dongguk.edu)

This work was supported by the National Research Foundation of Korea under Grant NRF-2021R1I1A2045721.

ABSTRACT In this era, research and development are becoming a continuous and accelerating process because technology changes rapidly with a short lifecycle. As a result, various methodologies are being developed to monitor these rapidly changing research trends; In particular, clustering method-related studies in science and technology documents are being developed with a variety of approaches. However, previous studies on document clustering methods focus on a specific field or language but do not take into consideration certain important pieces of information in science and technology documents. Therefore, this study proposes an embedding methodology that uses important content from scientific and technical documents. We took into consideration the importance of information containing core structures in science and technology documents and proposed a clustering methodology that analyzes structured and unstructured data, such as textual information, author information, and citation information. The proposed method combines both textual and structural data from the paper, using a method that focuses on screening important information by sections in science and technology documents. Then, Girvan-Newman clustering and Louvain clustering models are applied to generate embedding vectors and show evaluation results through the clustering indices. As a practical example, we applied the proposed methodology using paper data from the field of hydrogen cell vehicles. The results of this study will be effective in identifying gaps in technology for new technological development, identifying technology trends, and presenting directional information for future technology development.

INDEX TERMS Clustering method, data mining, text mining, text analysis, scientific publishing, fuel cells.

I. INTRODUCTION

Due to the rapid changes in technology with a short life cycle, research and development have also experienced an accelerating process. Therefore, responding quickly and effectively to changes in technology has now become the primary goal of the R&D department of most companies; hence the analysis of emerging technologies should be carried out in future technology planning [1]. It is necessary to review the content of a technological document to predict and prepare for new technology, and to accomplish this; it is necessary to identify the contents of the scientific document of the previous technology. A technical document is a term that covers all documents and materials that contain content about the given technology. Representative technical and scientific documents include patents and scientific papers. A scientific or technological document is a term that encompasses all documents and materials that contain the contents of a given technology or scientific knowledge. Analyzing these documents is essential in identifying technological advances. Patent documents that are the most representative technical documents have standard technical classification codes composed of structural information such as IPC (International Patent Classification) and CPC (Collaborative Patent Classification). Therefore, it is very easy to analyze the latest technology trends or to understand the development trends through the technology fields. In particular, various patent classification studies are underway because documents can be classified and clustered in a relatively straightforward manner [2]–[4]. However, since scientific papers lack structural information, such as technical classification codes, it is difficult to analyze trends by research fields and identify the development trend of the research areas. A more meaningful scientific and technological analysis would be possible

The associate editor coordinating the review of this manuscript and approving it for publication was Jerry Chun-Wei Lin.

if (1) paper clusters are generated based on the papers' contents and (2) the clusters of papers are matched to the technical classification codes of patents.

Even though analysis of scientific papers is difficult, it is crucial to analyze them due to the importance of their research contents related to science and technology. There are various ways to analyze scientific papers, such as clustering, classification, and summarization. One of the important tasks in investigating scientific papers is to compare and contrast them with patent data because many researchers have claimed that scientific papers and patents generally focus on science and technology, respectively. Although the patent data have their own classification codes, the scientific papers do not have a clear classification system. Thus, it is critical to generate the clusters of these papers to compare the results of two data domains. Therefore, an embedding process that enables researchers to express data in a vector space is required to successfully perform such clustering.

Previously, the analysis of scientific and technological papers was performed using only a small amount of structural data due to the limited structural information available. However, in recent years, analyses have been performed using text with unstructured information drawn from scientific papers. In previous studies of structure data analysis in network form, along with the use of unstructured data analysis that examines text [5], [6], technical-economic analysis was conducted through various indicators, such as keyword clusters, mainly by expanding the citation relationship. In the study of text mining-based clustering [7], [8], a document was created to examine the similarity matrix using textual information such as the title and abstract of the scientific paper, which contains unstructured information, to develop technology in a specific field. The study of summarizing scientific and technological papers [9] identified the trend of domain technology through a summary method that incorporated the document clustering method.

Previous research focusing on developing scientific papers' features has analyzed only a small amount of structural data due to a lack of structural information; however, efforts have recently been made to proceed with such analysis through the use of text that contains the nonstructural information in the scientific papers [6], [7]. Within text-based network analysis research, technology economic analysis was conducted through various types of data, such as keyword clusters, and by extending the citation relationship, which is also categorized as structural information. As a text mining-based clustering study, document-to-document similarity matrices were generated using textual information, such as titles and abstracts. In addition, the trends in domain technology were identified through summarizing methods that incorporated document clustering techniques [8]. Various data are currently being used to develop high-performance clustering methods to analyze scientific papers. Clustering methods draw upon citation information or author information, which are categorized as structural information in scientific papers, and proceed to clustering using the information itself [10], [11]. This method has the advantage of being easily accessible and fast in achieving analytical results; however, it does not accurately reflect the contents of the scientific paper. Table 1 presents the limitations of previous studies on paper clustering. There have been studies that additionally used text information in an expanded way, such as extracting keywords from papers published by a specific researcher and the titles and abstracts of cited papers [12] or utilizing the text contents of cited papers [13]. In addition, a study that proposed a coupling approach for the social use of scientific literature on an online information dissemination platform used both social data and paper data [14]. Furthermore, the clustering results of bibliographic coupling (BC) and co-citation (CC) are derived by using citation frequency and distribution and proximity between documents [15] or by using author and citation information in [16]. However, most of the existing studies can be classified as studies focusing on citation information and text data. There are limitations due to insufficient use of various data such as text and metadata together. Furthermore, it was difficult to reflect core information using the full-text of papers, considering only a subset of text data such as title/abstract in analysis.

TABLE 1. Limitations of previous studies.

Reference	Data type	Data Reflection			Core Structure Reflection
		Text	Citation	Author	
[12]	Paper	O	O	X	X(Citation's Tit/abs)
[13]	Paper	O	O	X	X(Citation's Tit/abs)
[14]	Paper, Media	O	O	X	X
[15]	Paper	O	O	X	X
[16]	Paper	X	O	O	X
Proposed Method	Paper	O	O	O	O

This study proposes a scientific paper embedding method that reflects the core structure of scientific documents using both text and metadata. Furthermore, it aims to discuss the results by combining the embedding methodology with clustering approaches so that it can create high-performance clusters that successfully reflect the contents of scientific papers. To this end, we conducted this research with the following research questions: Q1) It is possible to vectorize a scientific paper's text data and metadata (author and citation information) and embed it into a single vector? Q2) Is there a method to reflect the core structure of the scientific paper? And Q3) What is the clustering methodology that can help the best cluster this data?

In these research questions, it is necessary to deal with a method of integrating data with various characteristics at once and embedding it by reflecting the core structure of the paper. Through the proposed methodology, it will be possible to use various types of data and reflect the core structure of the paper in the embedding to cluster papers based on the main contents of scientific and technological papers.

The remainder of the study is structured as follows. Section 2 presents the literature survey of methods that reflect the core structure of the scientific paper, as well as the embedding model and clustering of the scientific paper. In Section 3, we propose an embedding methodology for optimal scientific paper clustering and a framework to explore an appropriate clustering methodology. In Section 4, we apply the proposed approach to the actual data, and in Section 5, we conduct validation and discussion to show that the embedding methodology is suitable for scientific paper clustering. Finally, in Section 6, the study concludes with this study's limitations and suggestions for future works.

II. BACKGROUND

A. STRUCTURES OF SCIENTIFIC PAPERS BY SECTION

Scientific and technological papers are written within a set of structures and templates for clear and concise delivery, so the content written differs depending on the specific structure. The currently applied document vectorization methods do not show strong performance when analyzing scientific and technological papers, as they take into consideration word information but do not include structural information. Therefore, studies related to the structure of each section in scientific and technical documents can be divided into a study on the structure of each section and a study on how to extract the structure for each section.

Studies that classify the structure of the contents of each of the scientific papers' sections contain various different viewpoints. The most representative and customary viewpoint was a perspective on the writing of the scientific paper. In the study of Eger [17], the scientific paper is summarized as a structure that distinguishes the process when writing a scientific paper. It is divided into an introduction (background knowledge, purpose, research questions, answers to questions), the main body (case study, methodology), an article (background information, results, comparisons, conclusions), and a list of references. In Cuschieri's study [18], the paper was explained by dividing it into a paper name, summary, keyword, introduction, methodology, experimental result, and conclusion. In the study by Vitse and Poland [19], title (author information), abstract (introduction, background knowledge, methodology, results, conclusion), Introduction (Assumptions), Methodology (Details, Reproducibility, etc.), Results (Statistical Results), Papers, and History were all examined. While this has the advantage of providing an easier method to divide scientific and technological papers, there is a limit to identifying the important content. Studies dividing the scientific paper from a rhetorical point of view were then conducted. In the study conducted by Teufel and Moens [20], the paper was divided into background knowledge, subject, related research, purpose and problem, solution and methodology, result, and conclusion. In a relatively recent study [21], the paper was divided into strengths, methodologies, solutions, opinions, conclusions, comparisons, and support. Although this has the advantage of providing data that both experts and non-experts can fully understand, it has limitations in that it is somewhat difficult to apply overall because the structure only works for a specific field of study. To solve this problem, studies have recently been conducted to divide the structure of the scientific paper according to the real core structures of the scientific paper [22]. The sections of the paper are classified into the prerequisites for conducting the experiment, the tools used to conduct the experiment, the means for evaluating the study, the results achieved in the study, the limitations of the study, and the expansion of future factors.

For our research, based on the core structure of scientific and technological papers, we chose to cluster papers by focusing on the most essential pieces of information [23]. If the structure of each section of the scientific paper is well-defined, it is necessary to extract the text for each defined structure. The method of extracting the important sentences from a document is largely divided into two: the method of extracting at the phrase level and the method of extracting at the sentence level [24], [25].

In addition, various studies exist that allow for the construction of a dictionary for extraction based on sentences. Previous studies that focus on the process of extracting content from papers can be broadly classified based on the data cutting levels and data extraction methods, which are representative methods for the cutting of data. There are phrase-level analysis methods, such as conditional random fields (CRFs) and support vector machines (SVMs), and sentence-level analysis methods that use machine learning algorithms, such as Bayesian classifiers and SVMs [26]. Although old-level analysis has not been studied due to the lack of datasets until recently, rule-based algorithms and CRF methods have been developed recently. However, there is still a limitation because these are still in their infancy [27]. On the other hand, the studies on sentence-level analysis have been conducted for a longer period of time. Rule-based algorithms are the mainstream, and most recently, studies using machine learning algorithms such as Bayesian classifiers and SVMs are also in use. Machine learning-based information extraction methods (support vector machines, Bayesian classifiers, etc.) have also been developed based on linear separators (linear function-based separators) that use surrounding words to extract information extraction patterns [28]. It was confirmed that the extraction method learned through this process exhibits good performance in extracting all types of information [26]. However, it is difficult to classify just the necessary data. In addition, many datasets are required for learning. The rule-based text information extraction method, which is the most popular method, defines a rule for extracting the structure of a paper using some data of the paper to be processed, and the structure's texts are extracted from the rest of the scientific paper's data according to the defined scientific paper structure rules [29]. In the study by Tkaczyk *et al.* [30], the metadata and reference data of the paper are thus extracted using the rule-based metadata extraction process and the reference extraction process, respectively, after the basic structure extraction.

B. SCIENTIFIC DOCUMENT EMBEDDING MODEL

The most popular natural language processing technique, document embedding, focuses on the extension of the word embedding method from a text and can map a document to a vector space [31], [32]. The document embedding methodology then converts the text data, which are unstructured data, into structured data, enabling a quantitative analysis for machine learning tasks and quick analysis [33]. Initially, the one-hot-vector method was used for embedding the text. The one-hot-vector method is a vector expressed by creating an N-dimensional vector that can express a specific word in a word dictionary containing N-words and placing 0 and 1 in the presence or absence of a word. Its limitation is that the information loss is large because only word occurrence can be checked. To solve this problem, studies using the frequency of occurrence of words have been proposed. The frequency-based methodology uses information on the frequency of occurrence of a word, and the TF-IDF method uses the information on the frequency of occurrences of a word in a specific document [34], [35]. Recently, deep learning-based methodologies that can take into consideration the context between words in a document have also been proposed [29], [36]. Doc2Vec methodology is an algorithm that extends a Word2Vec methodology, which learns words based on an artificial neural network with respect to the number of occurrences of words at both the paragraph and whole document levels. However, the methods that take into consideration the context between the words in a document and the context between the sentences have different characteristics. Where ever the greater the number of data, the higher the learning result; hence the method is specialized for information in the units of subsequent sentences and paragraphs. Therefore, this document embedding methodology has limitations in expressing the scientific paper text extracted by its structure.

The methodology for embedding the scientific paper document can be divided into two categories: an embedding method that uses citation information, which consists of structured information, and an embedding method that uses text, which consists of unstructured information. The embedding methodology based on citation information is applied based on a network, and the majority of the embedding results in the form of a network are derived [12]. Recently, a methodology for embedding both citation information and text information was proposed. It has an advantage where the blank parts of each data can be filled with each other [37], [38]. However, in the case of scientific papers, it is important to apply an appropriate text embedding methodology. Because documents can be classified by applying a clustering technique to the embedding results, an embedding methodology suitable for a paper can provide a solid basis for generating good clusters. Existing paper embedding studies have only considered text data or used both citation and text data [37], [38]. In addition, only some texts, such as abstracts and titles, are used for analysis, failing to reflect more substantial contents. These approaches have limitations in terms of data comprehensiveness because scientific papers include a lot of core information. Based on the research requirements, the use of heterogeneous data, such as citation and author information, must be effective [10], [11] to improve the clustering performance of scientific and technological papers. Therefore, paper embedding methods that use various types of important data should be studied. Hence, we propose an embedding framework that reflects both author and citation data as well as full-text data that reflects the papers' core structural information.

C. CLUSTERING METHODS AND THE CLUSTERING INDEX

Clustering methodologies have a variety of different methods that can be used according to input data. We explored various clustering studies to obtain high-performance clustering results using the embedded data. In this study, the document-to-document similarity matrix of the scientific papers was used as input data for clustering [7], [8].

Girvan-Newman (GN) clustering, which is mainly used for the purposes of network analysis, is a method of constructing and analyzing a network with the clustering target as the node and the relationship between the targets serving as the edges. Girvan-Newman clustering is a method to determine the optimum through the concepts of edge betweenness, centrality, and modularity. The edge betweenness centrality can be defined as the number of the shortest paths that go through an edge in a graph or network. When a node with high node betweenness centrality is considered a node that plays an important role in the information flow, then the betweenness centrality of an edge can also be seen as an edge that connects a pair of nodes that reflects the information flow well. The modularity calculation is as follows:

$$Modularity = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$$

(A_{ij} : $edgeweightbetweeniandj$, k_i : $sum\ of\ edgeweightincludnodei$, c_i : $clustercontaini$,

m : $totaledge$, $\delta\,(u, v)$: $ifu = vresultis1isnot0$)

The G-N clustering method takes a greater amount of time compared to other clustering methods, but its overall performance is high [39].

Finally, Louvain clustering methodology [40], which uses the document-to-document similarity as an input, is the process for clustering within a large-scale network. Louvain clustering performs clustering based on the connectivity between nodes, regardless of the nodes being cycled through. For a cluster created in the network graph of clustering, the modularity Q is defined as a measurement scale that can indicate the density of the connections within a cluster for connectivity with other networks. The larger the value of Q, the higher the intra-cluster connectivity and the lower the connectivity.

$$\boldsymbol{Modularity} = \left[\frac{\sum \boldsymbol{in} + 2\boldsymbol{k}_{i,in}}{2\boldsymbol{m}} - \left(\frac{\sum \boldsymbol{tot} + \boldsymbol{k}_i}{2\boldsymbol{m}}\right)^2\right]$$

$$- \left[\frac{\sum in}{2m} - \left(\frac{\sum tot}{2m} \right)^2 - \left(\frac{k_i}{2m} \right)^2 \right]$$

($\sum in$: *weight sum of cluster innerlink*,
$\sum tot$: *sum of all link weight from cluster*,
k_i : *sumof link weight from node i*,
$k_{i,in}$: *sum of weightwhich linkbetween C an di′snode*,
m : *sum of all link*).

To check the performance of the clustering, clustering index-based verification is performed. To evaluate the clustering results derived through the clustering process, the Silhouette index and Dunn index are used because these two indices showed the highest overall performance among the various existing clustering validity indexes that can evaluate clusters [41]. The Silhouette index is a method for calculating how high the cohesion within a cluster is when compared to the cohesion between clusters, having a value between −1 and 1. The formula is as follows:

$$s(i) = \frac{b(i) - a(i)}{\max a(i), b(i)} \quad -1 \leq \mathrm{s(i)} \leq 1$$

(b(i): average of the distances between the i-th individual and elements belonging to other clusters is obtained for each given cluster, and the smallest value is found; a(i): average of the distances between elements belonging to the same cluster with the i-th object;).

The Dunn index is a method of checking how high the density is divided. It is calculated as the ratio of the minimum and maximum distances between clusters. The greater the distance between the clusters and the smaller the variance value within the clusters, the higher the overall value. The formula is as follows:

$$DI_m = \frac{min_{1 \leq i \leq j \leq m} \delta(C_i, C_j)}{max_{1 \leq k \leq m} \Delta k}$$

($\delta\left(C_i, C_j\right)$: intercluster distance metric, between clusters C_i and C_j, m: number of clusters).

In our study, the clustering method was carried out using various clustering methodologies, and a comparison between the clustering methods was conducted. As an index of comparison, we will check the performance of the clustering results using both the Silhouette index and Dunn index.

III. PROPOSED FRAMEWORK

We propose an embedding methodology that can better reflect the core structure of a scientific paper. Through a vectorization process suitable for each data of the paper, such as the text, author information, and citation information, a vector value that describes a paper well can be obtained. Additionally, we suggest a method for creating a similarity matrix between the documents to obtain improved clustering results. The core structure of the science and technology paper is extracted from the text information, and the extracted text information for each core structure is then embedded into the vector space. A single document-to-document similarity matrix is created by assigning weights to the comparison of keywords for each text of the key structure and abstract.

TABLE 2. Scientific paper structure.

Section	Structure
Introduction	Research Objectives
	Research Significance
	Research Limitation
	Research Problem(Question)
Methods	Research Methods
Background	Research Background
Illustration/ case study	Data
	Research Outcome
Conclusion	Findings
	Implications
	Further Research (Recommendations)

A similarity matrix between the scientific paper documents is created through the analysis of author information and citation information and then finally combined through a hybrid similarity method. The weight for each piece of information is then determined by selecting the case with the highest clustering modularity. The final result of the research will be displayed as one embedding vector based on a similarity matrix containing text, citation, and author information which will show high performance in the clustering method.

The embedding model framework for scientific and technological papers consists of three steps, as shown in Figure 2. First, scientific and technological papers are collected and preprocessed. Since the scientific paper data are collected in PDF format, they are converted into a text format that can be analyzed, and only documents containing the core structure are screened. For the core structure of the scientific paper, the structure defined in previous studies is used, and the text for each structure is extracted by defining the cue word dictionary for each core structure. As a result, in the first step, text, author information, and citation information for each core structure of the paper are indexed. In the second stage, various types of data collected from scientific and technological papers are vectorized and combined into a single vector-based on a hybrid approach. Text data are embedded based on a similarity matrix and a document-term matrix (DTM) that vectorizes the text of a document through a bag of words (BoW)-based representation of each document. For author information, co-authorship is used, and for citation information, the bibliographic coupling method is used. Finally, each data point can be expressed in the same form through a similarity matrix. In the last step, we explore the clustering methodology using the proposed framework by combining the embedding results derived from the similarity matrix using a hybrid approach.

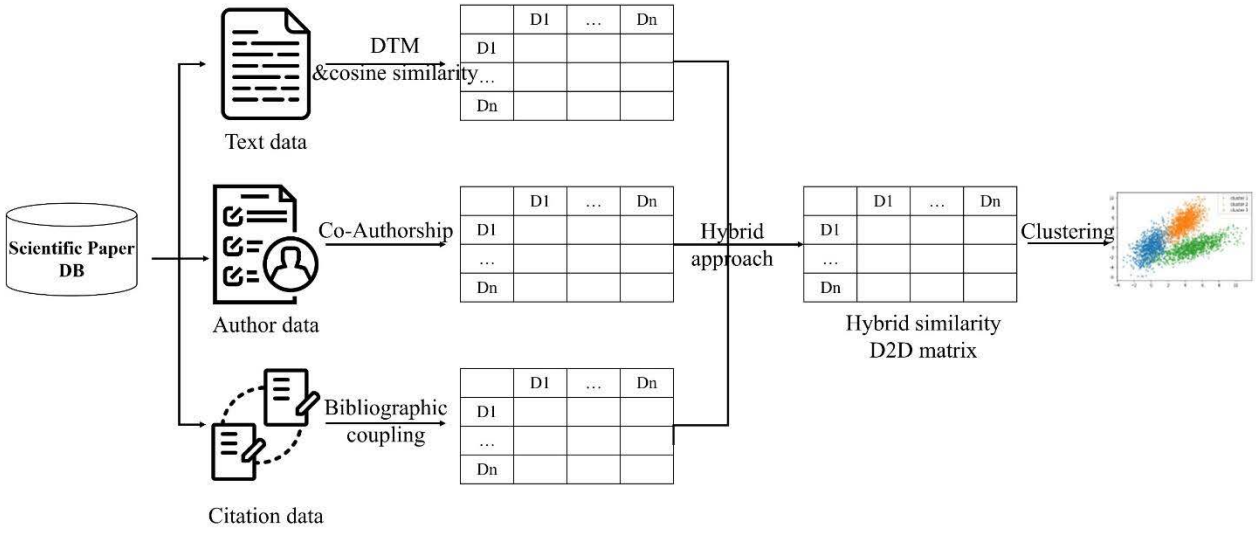

FIGURE 1. The research concept.

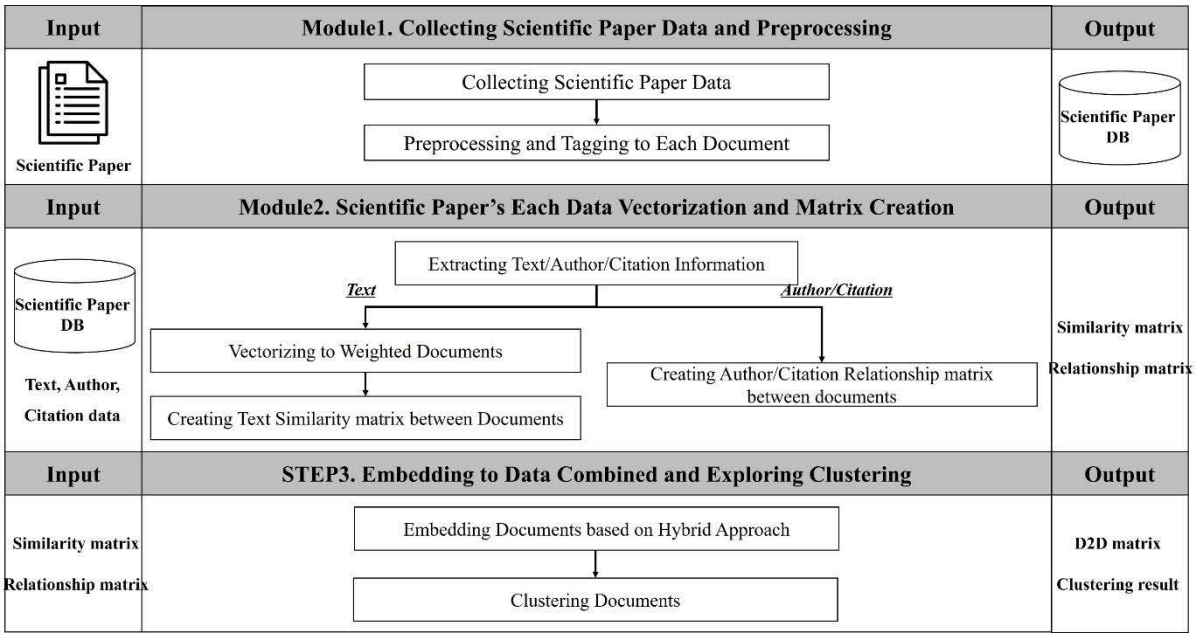

FIGURE 2. Research framework.

A. COLLECTING SCIENTIFIC PAPER DATA AND PREPROCESSING

In the first module, the data collection, preprocessing, and tagging processes are performed. Within the first process, i.e., the data collection process, the designated specific field of data is first collected. In this study, text analysis is conducted based on the structure of the paper; therefore, it is easy to analyze papers with identical structures. In this case, a well-known journal in a given field is selected from within the JCR database that is designated as a scientific journal for the analysis. The collected text data are collected in PDF format, whereas for author information or citation information, the database consisting of the text format is used. In the second process, the data are preprocessed into a form that can be analyzed. Preprocessing is performed to facilitate the analysis of the collected information. First, the verbal data from the science and technology papers stored in PDF format are converted into a text format, and the basic section structure (Introduction, Background, Method, Case Study (Illustration), Conclusion, etc.) information is obtained (excluding documents such as articles). Each database is then created by indexing the title, full-text information, citation information, and author information.

B. SCIENTIFIC PAPER'S EACH DATA VECTORIZATION AND MATRIX CREATION

This second process mainly involves embedding the data collected in the first process and then preprocessing it into a vector space that can be easily clustered. Through this specific process, the text, author, and citation data are converted into data types that are more suitable for clustering. The second module mainly consists of the following: 1. text analysis-based structure, 2. Author network-based analysis, and 3. Bibliographic coupling.

During the text vectorization process, text data extraction from scientific and technological papers is first carried out. Within this study, the core structure of the scientific paper is first defined, and then the sentences corresponding to the specifically defined structures are extracted and analyzed. Extracting the text structure of the paper is reasonably important for the core structure of the paper, that is, for the actual clustering process. The core structure that we used for the research is shown in Table 2. Based on various previous research projects that analyze the core structure of scientific papers, we extract the correct text information that we require from papers [17], [19], [20], [27], [42], [43]. We define the structures that are important factors for clustering and give their definitions.

Based on the definition of the structure, rules and cue word-based defined structures are extracted. A cue word is a word that leads to a sentence; it is defined as a word that can assign a role for each sentence and then facilitate research to extract the desired contents. The rules and cue words used in this study are prepared with specific references to the research [22], [44]. The final cue words and rules are further defined by closely checking the extracted results through the prepared cue words and rules and then supplementing the missing portions.

The method of embedding the extracted text data is shown in Figure 3. To proceed with the embedding of the extracted text data, the term frequency method was implemented. This method was used before the term frequency-inverse document frequency method was developed. The term frequency-inverse document frequency method is the product of the frequency of a word, in addition to the reciprocal of the frequency of the given document. It is a methodology that gives great weight to certain words, mainly by excluding words that are too general rather than targeting to increase the number of overall words. In this study, it is not practical to use the inverse document frequency, a method for excluding common words in documents, because it is more essential to include the common main contents than to exclude them. Therefore, in this study, vectorization is carried out based on the term frequency, such that the number of keywords whose rarity is lowered is selected, and vectorization is performed for a total of 11 different structures.

A method is used to combine the vector data for each structure in each document, mainly by assigning a higher weight to the information that is most important for clustering. For instance, the abstract is the section where the author puts great effort in writing a scientific and technological document, and weights are assigned to each structure accordingly. The weight is determined based on the keyword similarity between the text information of each structure and the information of the given summary, and the chosen formula is as follows. As a result of the text analysis, one vector is generated for each scientific and technological paper document. This is then defined as the scientific and technological paper text vector. The text similarity matrix between the scientific and technological papers is created by calculating the cosine similarity between the generated vectors to improve text clustering performance [45].

$$\begin{aligned} V_{D_1} = {} & KSimilarity\left(Title_{D_1} + Abs_{D_1}, RO_{D_1}\right) \\ & + KSimilarity\left(Title_{D_1} + Abs_{D_1}, RS_{D_1}\right) \\ & + \cdots + KSimilarity\left(Title_{D_1} + Abs_{D_1}, Ref_{D_1}\right) \end{aligned}$$

$$(\boldsymbol{KSimilarity}\,(\boldsymbol{A}, \boldsymbol{B}) = \frac{\boldsymbol{A} \cap \boldsymbol{B}}{\boldsymbol{A} \cup \boldsymbol{B}})$$

To vectorize the citation information, we make use of normalized bibliographic coupling, which is a method used to calculate the bibliographic coupling relationship. The bibliographic coupling relationship indicates the degree to which the references are shared. When comparing the similarity of the two papers, it can be said that the more common the references are, the stronger the bibliographic coupling relationship is. Essentially, it is a concept that judges that a paper is similar if it is co-cited from another paper. Since the co-citation relationship has to determine the relationship for whether it is co-cited by the papers published in the future, the similarity value between the chosen papers may change if the papers are cited in the future, whereas the bibliographic coupling relationship is based on the citation of each given paper. As a reference, this is preferred because it has an advantage where the value does not change due to the ability to calculate the shared relationship. In this study, the following formula was used to calculate the normalized surge bond strength. Finally, a similarity value between each scientific and technological paper is derived.

$$N_{norm} = \frac{N_{AB}}{\sqrt{N_A} N_B} \tag{1}$$

(N_A, N_B: Cited paper from papers A, B, N_{AB}: number of commonly cited papers).

The author information and the final data we use are calculated based on both the author's network and the distance between the authors of the paper. The network node is formed as the author, and an edge exists if the authors have the co-authorship of a paper. The similarity between papers can be calculated by using the average distance between all of the authors of two papers.

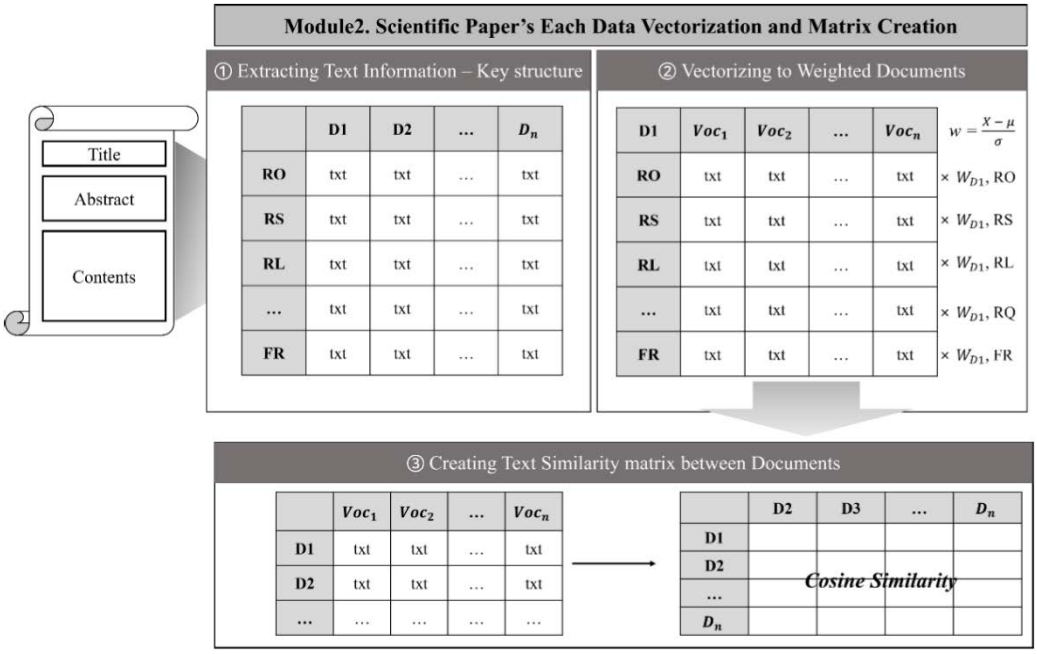

FIGURE 3. Scientific paper's each data vectorization and matrix creation.

C. EMBEDDING TO DATA COMBINED AND EXPLORING CLUSTERING

In this study, we try to combine text, citation, and author information based on a hybrid approach. The study that combines the embedding results derived from the similarity matrix between the text information and citation information of scientific papers mainly consists of studies that calculate the general sum. Since this study makes use of three types of information, the hybrid approach study [37], which combines both existing heterogeneous data and text data, was extended to be applied. The procedure of the hybrid approach is as follows: 1. Set the first weight for each text, author, and citation ($\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$). 2. Calculate the combined embedding vector 3. Apply the clustering method (Louvain, Girvan-Newman) 4. Calculate the modularity of the cluster results. 5. Repeat steps 1-4 and selects the highest modularity of all. With this approach, the result shows the most appropriate clustering and displays the importance of each piece of information (text, author, and citation) [37], [38]. The formula for combining the three pieces of data is as follows:

$$HS_{(D_i,D_j)} = cos\left(\alpha \cdot arccos\left(TS_{(D_i,D_j)}\right) + \beta \cdot arccos\left(RS_{(D_i,D_j)}\right) + \gamma \cdot arccos\left(AS_{(D_i,D_j)}\right)\right)$$

$\{HS : hybrid\ similarity, TS : Text - based\ similarity, RS : Reference - based\ similarity$

$AS : Authorship - based\ similarity, \alpha \in [0,1], \beta \in [0,1], \gamma \in [0,1], \alpha + \beta + \gamma = 1\}$.

The papers are clustered based on the embedding results that are derived from the similarity matrix of previous scientific and technological papers. First, to assign an optimal weight to each data point, the clustering is performed by changing the weight by a value of 0.05 based on the simulation, and when the optimal weight is derived, it is adopted. The clustering methods were Girvan-Newman (G-N) clustering and Louvain clustering. G-N clustering configures a network in which the clustering object is a node, and the relationship between the objects is an edge. Additionally, it is a method that can determine the optimum method using the concepts of edge betweenness centrality and modularity, which are network analysis concepts. The edge mediation center is a concept that is applied to edge centrality, and it defines the degree to which the centrality of the node can be placed between each node while taking into consideration the shortest path between the nodes. When a node with high node mediation centrality is taken as a node that plays an important role in information flow, the mediation centrality of an edge can be seen as an edge that can connect a pair of nodes which reflects the information flow accurately. Modularity takes a longer period of time when compared to other network analysis methods but has the advantage of showing higher performance. The Louvain clustering method is a method that can be used to calculate modularity by solving the time-consuming method of finding the existing modularity. As the computation time is high, the size of the community of generated results is not large. In this study, clustering was performed using this method.

To check the performance of each clustering result, we focus on clustering index-based verification. To evaluate the clustering results based on clustering, we use the Silhouette index, which is a method of calculating how the high intra-cluster cohesion relates to inter-cluster cohesion.. In addition, we use the Dunn index, which calculates the ratio of the smallest distance between two entities belonging to

TABLE 3. Search equations.

keyword	(fuel or hydro) and (cell) and (automobile or car or vehicle)

different clusters to the largest distance between two entities belonging to the cluster. The Silhouette index and Dunn index formulas are described in Section B.

This study intends to examine whether a cluster that represents the clustering results accurately can be derived using the IPC code within the relevant patent so that the cluster that best represents the technology can be identified. The IPC code of a patent is an internationally unified patent classification system that can indicate the technical field of an invention. In other words, as the IPC code is divided into technical fields, the contents of the cluster of scientific papers can be analyzed in this study. First, the summary and keywords of the papers in the cluster are extracted. Then, topic modeling is conducted with LDA analysis in an attempt to match IPC codes with the results of topic modeling.

IV. CASE STUDY

A. COLLECTING SCIENTIFIC PAPER DATA AND PREPROCESSING

Scientific papers to be analyzed were selected from the field of hydrogen battery vehicles. The data were collected from the SCOPUS database, which contains scientific journals, books, and seminar materials, using keywords that were chosen by expert reviews, as shown in Table 3. Papers published from 2001 to 2020 were collected among the papers related to fuel cell vehicles, along with text information, citation information, and author information.

The title, EID, summary, full text, author, and citation information of all the collected papers are then combined into one dataset. A total of 1397 papers were collected, and the analysis was performed using a total of 986 papers, excluding cases where no information was available to be used.

B. SCIENTIFIC PAPER'S EACH DATA VECTORIZATION AND MATRIX CREATION

The first step in the second module, scientific paper vectorization, and matrix creation, focuses on the process of analyzing the text by its structure. Structural extraction is performed for text analysis, and the rules and cue words for extraction are defined in this step. The rules are prepared based on information from previous studies. The rules for extracting the defined structures are shown in Table 5.

The text data of each structure of the extracted scientific paper are vectorized based on the term frequency. For vectorization, a term frequency matrix is created for each structure, and the keywords are created using the top 2000 main keywords. One document embedding result is derived from the combination of the document embedding results for each keyword-based structure that was created by weighting the similarity with the document summary as a weight. Then,

TABLE 4. Collected results.

Rank	Journal	SJR	No. of papers
1	Nature Energy	46.49	184
2	Energy and Environmental Science	30.28	193
3	Joule	29.15	202
4	Progress in Energy & Combustion Science	28.93	173
5	Advanced Energy Materials	26.25	49
6	ACS Energy Letters	19.00	254
7	Renewable & Sustainable Energy Reviews	12.11	122
8	Journal of Materials Chemistry A	11.30	132
9	Applied Energy	8.84	130
10	Journal of Power Sources	8.24	48

D0 =	1	0	0	0.002	0.001	0.002	...	0.002
D1 =	0	1	0	0.76	0	0	...	0
D2 =	0	0	1	0.52	0	0	...	0
D3 =	0.002	0	0	1	0	0.053	...	0.07
D4 =	0	0	0	0	1	0	...	0
D5 =	0.002	0	0	0.053	0	1	...	0.077
D6 =	0	0	0	0.066	0	0.072	...	0.097
D7 =	0.002	0	0	0.009	0	0.01	...	0.013
D8 =	0.002	0	0	0.066	0	0.072	...	0.097
D9 =	0.002	0	0	0.071	0	0.077	...	0

FIGURE 4. Text vector based cosine similarity.

to proceed with the clustering in the form of a network, the similarity calculation between documents for each core structure is performed. The cosine similarity method is selected for the similarity calculation from which a scientific paper-scientific paper similarity matrix for each core structure is created. The document-to-document similarity matrix-based embedding results of scientific paper text, derived by weighting keywords in the abstract of the scientific paper and the keyword similarity of each core structure as weights, which constitute the scientific paper-scientific paper similarity matrix for each structure of the one document-to-document matrix, is shown in Figure 4. It was confirmed that the maximum value of the document-to-document similarity matrix-based embedding result was equal to 1, the minimum value was 0, and the average similarity was 0.399.

Second, citation information analysis is performed using the normalized bibliographic coupling method, and the derived document-to-document similarity matrix-based embedding result is as follows (Figure 5):

The maximum value of the document-to-document similarity matrix-based embedding result between the prepared

TABLE 5. Cue words and rules.

Section	Structure	Cue word Rule
Introduction	Research Objectives(RO)	(purpose, aim, describe, objective) + In this (paper, study, research)
	Research Significance(RS)	(significant*, important*+advantage*) + (paper, study, research)
	Research Limitation(RL)	(Limit*, weak*, vulnerable*, prone*, drawback*) + In this (paper, study, research)
	Research Question(RQ)	hyposcientific paper, problem, question
Background	Research Background(BG)	(previous, recent) + (study, research) + (show*, cover*, reveal*, discover*)
Methods	Research Methods (RM)	(method, how, propose, introduce, way, suggest, proposal, structure, built*) + (paper, study, research)
Illustration	Data (D)	data, use*, resource* + (paper, study, research)
	Research Outcome (Rout)	(result, outcome*, final*,consequence*, Issue*)+(paper, study, research)
Conclusion	Findings (Find)	Finding*, discover*, detect*, unearth*
	Implications (Impli)	implication, imply, suggest, inference+(paper, study, research)
	Further Research (FR)	further, further effort, overcome, develop

D0 =	1	0.95	0.62	0.75	0.59	0.78	...	0.86
D1 =	0.95	1	0.65	0.76	0.68	0.79	...	0.86
D2 =	0.62	0.65	1	0.52	0.64	0.74	...	0.67
D3 =	0.71	0.76	0.52	1	0.5	0.5	...	0.8
D4 =	0.59	0.68	0.64	0.5	1	0.7	...	0.63
D5 =	0.78	0.79	0.74	0.65	0.7	1	...	0.82
D6 =	0.71	0.74	0.57	0.74	0.49	0.67	...	0.76
D7 =	0.74	0.57	0.63	0.61	0.64	0.91	...	0.76
D8 =	0.67	0.63	0.62	0.57	0.54	0.88	...	0.7
D9 =	0.54	0.56	0.59	0.45	0.9	0.68	...	0.6

FIGURE 5. Citation vector based cosine similarity.

citation information document was equal to 1, the minimum value was equal to 0, and the average similarity was equal to 0.09.

Third, for author information analysis, author analysis was performed by extending the author network to the document level, and the derived document-to-document

D0 =	1	0	0	0	0.143	0.362	...	0
D1 =	0	1	0	0	0	0	...	0
D2 =	0	0	1	0	0	0	...	0
D3 =	0	0	0	1	0	0	...	0
D4 =	0.134	0	0	0	1	0.22	...	0
D5 =	0	0	0	0	0	1	...	0
D6 =	0.293	0	0	0	0.251	0	...	0
D7 =	0	0	0	0	0	0	...	0
D8 =	0.283	0	0	0	0.2	0.362	...	0
D9 =	0	0	0	0	0	0	...	0

FIGURE 6. Author vector based author similarity.

TABLE 6. Overall results of clustering.

		GN	Louvain
Text + Citation + Author (Hybrid approach)	Silhouette	0.034	0.386
	Dunn	7.56e-17	7.25e-17
	Valid N-cluster (N-cluster)	3(3)	5(18)

similarity matrix-based embedding results are shown as follows (Figure 6):

It was confirmed that the maximum value before the normalization of the author information document-to-document similarity matrix-based embedding result was 2.67, the minimum value was 0, and the average value was 0.0069.

C. EMBEDDING TO DATA COMBINED AND EXPLORING CLUSTERING

The document-to-document similarity matrix is calculated based on the previously derived information drawn from text, author, and the reference document-to-document similarity matrices. Through this process, we check which data have the highest influence on clustering performance among the papers' text information, citation information, and author information and then present the most appropriate clustering algorithm. The calculation proceeds through the mentioned formula (Section 3), changing each α, β, and γ value to 0.05 and running the clustering then selecting the best performance of α, β, γ. To select a cluster with the best performance among G-N clustering and Louvain clustering, the weights α, β, and γ with the highest modularity values are adopted by using their modularity values. For the data used in the study, when Girvan Newman clustering was performed, the highest clustering performance was shown when the text, citation information, and author information had weights of 0.15, 0.3, and 0.55, respectively. When Louvain clustering was performed, the highest modularity was shown when the text, citation information, and author information had weights of 0.30, 0.40, and 0.30, respectively. The results from calculating the index for each data clustering result were then confirmed.

We present the optimal clustering methodology based on the Silhouette index, Dunn index, and number of clusters. First, when checking the Silhouette index, Louvain had a higher value, and the Dunn index was derived at a similar level for both Girvan Newman and Louvain.

TABLE 7. Examples of louvain clustering results.

Cluster	Number of papers	Titles of Papers
1	24	Air-cathode structure optimization in separator-coupled microbial fuel cells, …
2	4	Effect of co gas and anode metal loading on h2 oxidation in proton exchange membrane fuel cell, …
3	4	Adipose triglyceride lipase and hormone-sensitive lipase are the major enzymes in adipose tissue triacylglycerol catabolism, …
4	406	An improved synthesis method of ceria-carbonate based composite electrolytes for low-temperature sofc fuel cells, …
5	2	New fuels as alternatives to methanol for direct oxidation fuel cells, …

The results of clustering using the proposed methodology can be summarized as follows. Although the Silhouette index is the most used index among various indices to evaluate cluster validity levels, the optimal cluster result was confirmed by including both the index and the number of clusters. First, in Table 6, the optimal results were derived using the Silhouette index. When using the Louvain clustering methodology, the Silhouette index was highest. Second, in Table 6, when using Girvan-Newman clustering, the Dunn index was slightly higher than that of the Louvain cluster, but it was slightly different. Last, when we checked the valid number of clusters, the Girvan-Newman clustering result had 3 valid clusters. In contrast, the Louvain clustering result had 5 valid clusters. The valid cluster is the cluster that has more than 5 documents. The number of clusters should be neither too high nor too low. In this study, the optimal number of clusters is approximately 22, which is similar to that of the Louvain clustering method. The optimal number of clusters can be calculated by the rule of thumb, and the calculation formula is:

$$k \approx \sqrt{n/2}$$

(k=*number of cluster*, n=*number of document*).
The optimal clustering method is Louvain clustering in this study, where the overall results from the clustering method are shown in Table 7. When Louvain clustering was performed, the weights for each data point were 0.3 for text information, 0.4 for citation information, and 0.3 for author information by including text, citation, and author information that were significant for clustering. We can confirm that among these, citation information is the most important data for clustering.

V. VALIDATION AND DISCUSSION

Clustering is a representative unsupervised learning method that has the advantage of receiving various types of data as input. However, it possesses a strong limitation: the verification method is quite difficult and complex to interpret. Therefore, in this study, the clustering evaluation index was used to confirm the results. The clustering evaluation index identifies only the relationship between clusters; therefore, it is necessary to review the contents of each data point. In this study, we attempt to confirm the clustering results using methods other than the clustering evaluation index.

TABLE 8. Results of topic modeling.

No	Topic	IPC	Number of Paper
1	Fuel, cell, BCY, electrolyte, battery, powertrain, system, fatty, using, optimized, proved, lithium, capacity, control, CO2, temperature, acids, storage, lipid	H01F	24
2	Fuel, cell, hydrogen, power, A, binding, using, high, In, energy, cells, DNA, lipid, carbon, system, different, two, CNP	B60R	406
3	Hydrogen, cell, system, fuel, design, cells, compounds, cancer, delivery, FA-DABA-SMA, use, energy, performance, using, evaluated, hybrid, hydrophobic	X	42
4	Fuel, high, stack, carbon, support, gas, performance, CNF, in, all, purging, cell rights, reserved, rate, effect, this, catalytic	F01N	60
5	Fuel, cell, hydrogen, solar, energy, production, power, in, use, carbon, liquid, cells, using, also, MIF, direct, all	G01F	408

A. VALIDATION USING TOPIC MODELING AND IPC CODE

To check the validity and robustness of the derived clustering results, we used the LDA (latent Dirichlet allocation) methodology, which is a representative topic modeling method. As shown in Table 8, we applied LDA to the results of Louvain clustering, which performed better in a hybrid approach. LDA analysis is a probabilistic model that expresses the probability that a particular word will appear in a particular topic. In addition to topic distribution, it has the advantage of estimating the distribution of words per topic. In this study, we performed LDA for each cluster to validate, compare, and contrast the characteristics of the clusters. The result of 10 topic keywords in each cluster, which were derived by extracting only the top effective clusters with five or more documents in the cluster, is the same as the topic item in Table 8. In addition, to match the technical classification codes of the patent to the results of the paper clusters, we mapped the IPC codes to the topics using the industry–patent linkage table provided by the Korean Intellectual Property Office. As a result, we confirmed that the different IPC codes for each cluster were identical.

B. VALIDATION BETWEEN DATA

For the second validation, we applied GN and Louvain clustering methodologies to each data point, and the results for three performance metrics were quantitatively confirmed. As seen from the results in Table 9, high Silhouette index values were confirmed in both clustering models when the text, citation, and author data were all considered by applying

TABLE 9. Validation between data.

Data		GN	Louvain
Text + Citation + Author (Hybrid Approach)	Silhouette	0.034	0.386
	Dunn	7.56e-17	7.25e-17
	N-cluster	3	18
Text	Silhouette	0.126	-0.717
	Dunn	0	0
	N-cluster	108	110
Authorship	Silhouette	-0.0485	-0.234
	Dunn	0.0	0.0
	N-cluster	395	391
Citation	Silhouette	-0.818	0
	Dunn	0.0	0
	N-cluster	77	1,300

TABLE 10. Validation between embedding models.

Method		GN	Louvain
Text + Citation + Author (Hybrid Approach)	Silhouette	0.034	0.386
	Dunn	7.56e-17	7.25e-17
	N-cluster	3	18
Doc2Vec	Silhouette	0.036	0.036
	Dunn	0.0003	-0.00029
	N-cluster	3	3
TF-IDF	Silhouette	0.1416	0.273
	Dunn	0	0
	N-cluster	4	7

the hybrid approach. In particular, when the hybrid approach was applied to Louvain, we found that a value closest to 1 was obtained and clustered well. On the other hand, when clustering was performed using only citation data, the performance was the worst. This suggests that when the role of each data point is methodologically emphasized, it can be used as a complementary perspective. The results are the same when comparing the Silhouette index values and the number of clusters. When only one data point is considered, the number of invalid clusters is confirmed. However, in the hybrid approach, it can be seen that a valid number of clusters can be verified and used. This indicates that embeddings suitable for clustering have been generated. Therefore, it can be said that the embedding framework based on the hybrid approach proposed in this study is suitable for paper clustering.

C. VALIDATION BETWEEN EMBEDDING MODELS

As a third validation, we compared the clustering results with the Doc2Vec and TF-IDF models. Looking at the results described in Table 10, the hybrid approach shows the highest average index performance among the two clustering models. In the case of Doc2Vec, the text data are reflected in the model according to the window size, so it cannot reflect the core information within the contents of the paper. TF-IDF is a frequency-based method because each word in the DTM (document-term matrix) is given a significant degree of weight. However, the hybrid approach proposed in this study is based on text data reflecting the core structure of the text and citation/author data. In addition, when only text data are used during the hybrid approach in Table 9, it shows better performance than the Doc2Vec model in Table 10. This suggests that text data reflecting the core structure of a paper's information is a more suitable method for paper clustering.

VI. CONCLUSION

In this study, we aimed to propose an embedding methodology that reflects the core structure of scientific papers and includes text, author, and citation data altogether. Furthermore, by applying clustering, we explored the clustering methodology with the best performance for the proposed embedding process. The results of this study showed that the best performance in clustering was achieved when text, author, and citation data were embedded together. In particular, the proposed approach delivered the best performance compared with the existing models, Doc2Vec and TF-IDF, which have been widely used for scientific clustering papers.

Based on the derived results, we can suggest the following implications from academic and practical aspects. Academically, the proposed study consists of an embedding framework that uses various data, including the text data of a paper, its authors, and citations. In fact, when all of these three types of data are considered, the GN and Louvain models give good clustering performance. This result also indicates the possibility of developing advanced models using various data in the future. In addition, instead of considering the entire text of scientific papers, text data for each core structure are extracted and reflected. The result of clustering verification indicates that the proposed model shows good performance. Thus, it can be interpreted that it relevantly reflects the contents of the papers. Furthermore, only the text data of the core structure of the desired paper can be used according to the purposes of clustering in the future. From the practical perspective, there is a contribution in the prospect that the results derived through the paper clustering process can be used to derive a promising research field or devise an R&D strategy. In particular, this study suggests the possibility of analyzing patent and paper data at the same level. This suggestion is supported by the result that the derived paper clusters can be assigned to various IPC codes in the patent data. In general, a paper is a medium that mainly describes basic research that is the basis of scientific knowledge, and a patent is a medium that presents applied research that is relatively close to commercialization. Thus, these two are considered valuable documents with different characteristics. Therefore, it is expected that, when analyzed together, patents and papers can become resourceful for scientific and technological development.

However, this study has several limitations. First, deep learning-based document embedding methodologies have not been applied to this study, although various deep-learning methods have been proposed recently. In the future, it will be necessary to perform clustering by proposing a neural network-based embedding model or by reflecting recent models such as BERT and ELMo. Although the proposed study delivered better performance than the clustering results with

Doc2Vec, which is a representative neural network-based embedding methodology, it can still be improved by using state-of-the-art neural network approaches. Second, when matching and interpreting topic keywords and IPCs by applying LDA to the clustering results, we performed the evaluation based on qualitative judgment. Thus, the subjectivity issue can be solved by using a more systematic model that even generates the names of clusters. Third, one limitation is that patent data are not directly collected and analyzed at the same level as the papers. Since it has become possible to use both data together, future research can deal with promising technology opportunities and technology life cycle analysis based on the combination of paper and patent data. Finally, in the process of collecting and analyzing paper data, the preprocessing process was insufficiently applied. In the study of Jain *et al.* [46], the documents were reviewed and screened on the basis of qualitative criteria such as journal criteria. In addition, the ESI (Essential Science Indicators) field was used to review the consistency and quality of paper contents, and then appropriate data was selected [47]. In this study, scientific papers were screened by just focusing on the structural form of data. However, when the proposed approach is used for technological intelligence, additional work is required for sophisticated data screening, such as the review on the consistency of contents and the quality and suitability of technical fields.

REFERENCES

[1] A. R. Patil and A. A. Manjrekar, "An innovative approach to classify and retrieve text documents using feature extraction and Hierarchical clustering based on ontology," in *Proc. Int. Conf. Comput., Anal. Secur. Trends (CAST)*, vol. 2016, pp. 371–376.

[2] B. Degroote and P. Held, "Analysis of the patent documentation coverage of the CPC in comparison with the IPC with a focus on Asian documentation," *World Pat. Inf.*, vol. 54, pp. S78–S84, Sep. 2018.

[3] S. Adams, "Comparing the IPC and the U.S. classification systems for the patent searcher," *World Pat. Inf.*, vol. 23, no. 1, pp. 15–23, Mar. 2001.

[4] T. Tran and R. Kavuluru, "Supervised approaches to assign cooperative patent classification (CPC) codes to patents," in *Proc. Int. Conf. Mining Intell. Knowl. Explor.* Hyderabad, India: Springer, 2017, pp. 22–34.

[5] F. Shi, L. Chen, J. Han, and P. Childs, "A data-driven text mining and semantic network analysis for design information retrieval," *J. Mech. Des.*, vol. 139, no. 11, p. 139, Nov. 2017.

[6] B. Yoon and Y. Park, "A text-mining-based patent network: Analytical tool for high-technology trend," *J. High Technol. Manage. Res.*, vol. 15, no. 1, pp. 37–50, Feb. 2004.

[7] K. W. Boyack, D. Newman, R. J. Duhon, R. Klavans, M. Patek, J. R. Biberstine, B. Schijvenaars, A. Skupin, N. Ma, and K. Börner, "Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches," *PLoS ONE*, vol. 6, no. 3, Mar. 2011, Art. no. e18029.

[8] C.-K. Yau, A. Porter, N. Newman, and A. Suominen, "Clustering scientific documents with topic modeling," *Scientometrics*, vol. 100, no. 3, pp. 767–786, Sep. 2014.

[9] M. Moradi, "CIBS: A biomedical text summarizer using topic-based sentence clustering," *J. Biomed. Informat.*, vol. 88, pp. 53–61, Dec. 2018.

[10] S. Ma, C. Zhang, and X. Liu, "A review of citation recommendation: From textual content to enriched context," *Scientometrics*, vol. 122, no. 3, pp. 1445–1472, Mar. 2020.

[11] Y. Zhang and Q. Ma, "Citation recommendations considering content and structural context embedding," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Feb. 2020, pp. 1–7.

[12] L. Bornmann, R. Haunschild, and S. E. Hug, "Visualizing the context of citations referencing papers published by Eugene garfield: A new type of keyword co-occurrence analysis," *Scientometrics*, vol. 114, no. 2, pp. 427–437, Feb. 2018.

[13] B. Aljaber, N. Stokes, J. Bailey, and J. Pei, "Document clustering of scientific texts using citation contexts," *Inf. Retr.*, vol. 13, no. 2, pp. 101–131, Apr. 2010.

[14] S.-U. Hassan, N. R. Aljohani, M. Shabbir, U. Ali, S. Iqbal, R. Sarwar, E. Martínez-Cámara, S. Ventura, and F. Herrera, "Tweet coupling: A social media methodology for clustering scientific publications," *Scientometrics*, vol. 124, no. 2, pp. 973–991, Aug. 2020.

[15] S. Zhang, Y. Xu, and W. Zhang, "Clustering scientific document based on an extended citation model," *IEEE Access*, vol. 7, pp. 57037–57046, 2019.

[16] J. Yun, S. Ahn, and J. Y. Lee, "Return to basics: Clustering of scientific literature using structural information," *J. Informetrics*, vol. 14, no. 4, Nov. 2020, Art. no. 101099.

[17] E. I. Eger, "A template for writing a scientific paper," *Anesthesia Analgesia*, vol. 70, no. 1, pp. 91–96, 1990.

[18] S. Cuschieri, V. Grech, and C. Savona-Ventura, "WASP (Write a scientific Paper): Structuring a scientific paper," *Early Human Develop.*, vol. 128, pp. 114–117, Jan. 2019.

[19] C. L. Vitse and G. A. Poland, "Writing a scientific paper—A brief guide for new investigators," *Vaccine*, vol. 35, no. 5, pp. 722–728, 2017.

[20] S. Teufel and M. Moens, "Sentence extraction and rhetorical classification for flexible abstracts," in *Proc. AAAI Spring Symp. Intell. Text Summarization*, 1998, pp. 89–97.

[21] M. L. Khodra, D. H. Widyantoro, E. Aziz, and R. T. Bambang, "Information extraction from scientific paper using rhetorical classifier," in *Proc. Int. Conf. Elect. Eng. Inform.*, Jul. 2011, pp. 1–5.

[22] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Information extraction from scientific articles: A survey," *Scientometrics*, vol. 117, no. 3, pp. 1931–1990, Dec. 2018.

[23] S. R. Basha, J. K. Rani, and J. P. Yadav, "A novel summarization-based approach for feature reduction, enhancing text classification accuracy," *Eng., Technol. Appl. Sci. Res.*, vol. 9, no. 6, pp. 5001–5005, 2019.

[24] A. Kyriakopoulou and T. Kalamboukis, "Combining clustering with classification for spam detection in social bookmarking systems," in *Proc. Eur. Conf. Mach. Learn. Princ. Pract. Knowl. Discovery Databases (ECML PKDD)*, Antwerp, Belgium, 2008, pp. 47–54.

[25] G. Sinclair and B. Webber, "Classification from full text: A comparison of canonical sections of scientific papers," in *Proc. Int. Joint Workshop Natural Lang. Process. Biomed. Appl.*, 2004, pp. 69–72.

[26] A. Téllez-Valero, M. Montes-y-Gómez, and L. Villaseñor-Pineda, "A machine learning approach to information extraction," in *Proc. Int. Conf. Intell. Text Process. Comput. Linguistics*. Mexico City, Mexico: Springer, 2005, pp. 539–547.

[27] M.-S. Song and Y. M. Ko, "A study on the macro analysis of knowledge structure of the domestic Korean studies for identifying the research fields," *J. Korean Soc. Inf. Manage.*, vol. 32, no. 3, pp. 221–236, Sep. 2015.

[28] S. Rahamat Basha and J. K. Rani, "A comparative approach of dimensionality reduction techniques in text classification," *Eng., Technol. Appl. Sci. Res.*, vol. 9, no. 6, pp. 4974–4979, Dec. 2019.

[29] T. Mikolov, I. C. K. Sutskever, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[30] D. Tkaczyk, P. Szostek, M. Fedoryszak, P. J. Dendek, and Ł. Bolikowski, "CERMINE: Automatic extraction of structured metadata from scientific literature," *Int. J. Document Anal. Recognit. (IJDAR)*, vol. 18, no. 4, pp. 317–335, Dec. 2015.

[31] A. M. Dai, C. Olah, and Q. V. Le, "Document embedding with paragraph vectors," 2015, *arXiv:1507.07998*.

[32] J. Han Lau and T. Baldwin, "An empirical evaluation of doc2vec with practical insights into document embedding generation," 2016, *arXiv:1607.05368*.

[33] J. Tang, M. Qu, and Q. Mei, "PTE: Predictive text embedding through large-scale heterogeneous text networks," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 1165–1174.

[34] A. A. Hakim, A. Erwin, K. I. Eng, M. Galinium, and W. Muliady, "Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach," in *Proc. 6th Int. Conf. Inf. Technol. Elect. Eng. (ICITEE)*, Oct. 2014, pp. 1–4.

[35] L.-P. Jing, H.-K. Huang, and H.-B. Shi, "Improved feature selection approach TFIDF in text mining," in *Proc. Int. Conf. Mach. Learn. Cybern.*, Nov. 2002, pp. 944–946.

[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[37] W. Glänzel and B. Thijs, "Using 'core documents' for the representation of clusters and topics," *Scientometrics*, vol. 88, no. 1, pp. 297–309, 2011.

[38] D. Yu, Z. Xu, Y. Kao, and C.-T. Lin, "The structure and citation landscape of IEEE Transactions on Fuzzy Systems (1994–2015)," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 2, pp. 430–442, Apr. 2018.

[39] C. Lim, D. Yun, I. P. G. Park, S. Koh, and B. Yoon, "Exploring prospective research areas in UI/UX through the analysis of patents," *Korean Manage. Sci. Rev.*, vol. 32, no. 4, pp. 1–18, 2015.

[40] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech.: Theory Exp.*, vol. 2008, no. 10, Oct. 2008, Art. no. P10008.

[41] J. M. Luna-Romera, J. Garcia-Gutierrez, and J. C. Riquelme-Santos, "An approach to silhouette and Dunn clustering indices applied to big data in spark," in *Proc. Conf. Spanish Assoc. Artif. Intell.* Springer, 2016, pp. 160–169.

[42] M. Derntl, "Basics of research paper writing and publishing," *Int. J. Technol. Enhanced Learn.*, vol. 6, no. 2, pp. 105–123, 2014.

[43] N. Kando, "Text-level structure of research papers: Implications for text-based information processing systems," in *Proc. 19th Annu. BCS-IRSG Colloq. IR Res.*, 1997, pp. 1–14.

[44] S. Teufel, A. Siddharthan, and C. Batchelor, "Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2009, pp. 1493–1502.

[45] M. S. Basha, S. Mouleeswaran, and K. R. Prasad, "Sampling-based visual assessment computing techniques for an efficient social data clustering," *J. Supercomput.*, vol. 77, no. 8, pp. 8013–8037, 2021.

[46] P. K. Jain, R. Pamula, and G. Srivastava, "A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews," *Comput. Sci. Rev.*, vol. 41, Aug. 2021, Art. no. 100413.

[47] F. Janssens, L. Zhang, M. B. De, and W. Glänzel, "Hybrid clustering for validation and improvement of subject-classification schemes," *Inf. Process. Manage.*, vol. 45, no. 6, pp. 683–702, 2009.

SONGHEE KIM is currently pursuing the master's degree with the Department of Industrial and Systems Engineering, Dongguk University. Her theme of study has involved artificial intelligence patent analysis, new technology development methodology, and visualization algorithms. Her current research interests include natural language processing, technology opportunity discovery, and scientific paper text analysis techniques.

SUYEONG LEE is currently pursuing the master's degree with the Department of Industrial and Systems Engineering, Dongguk University. Her theme of study has involved patent analysis, technology evolution analysis, and patent embedding. Her current research interests include technology intelligence, natural language processing, and technology DNA analysis with patent mining.

BYUNGUN YOON (Senior Member, IEEE) is currently a Professor with the Department of Industrial and Systems Engineering, Dongguk University. His theme of study has involved technology, patent analysis, new technology development methodology, and visualization algorithms. His current research interests include enhancing technology road mapping, research and development quality, and product designing with data mining techniques.