Technology Trend Forecasting and Idea Support Incorporating Generative AI Using Patent Information

Yurie Okuhara
Department of Liberal Arts, Informatics
The Open University of Japan
2-11 Wakaba, Mihama-ku, Chiba 261- 8586, Japan
email address: 2010022122@campus.ouj.ac.jp

António Oliveira Nzinga René
Department of Data Science
Toyama Prefectural University
5180 Kurokawa, Imizu, Toyama 939-0398, Japan
email address: rene@pu-toyama.ac.jp

***Abstract*—The rapid accumulation of patent documents presents challenges for traditional analysis methods due to the high dimensionality and volume of data. This study proposes an integrated framework that combines Transformer-based Sentence-BERT embeddings, UMAP dimensionality reduction, and k-medoids clustering to extract and visualize meaningful patent clusters. Selected clusters are analyzed through 2D and 3D visualizations to reveal core and peripheral technological concepts, semantic densities, and inter-cluster relationships. Furthermore, generative AI is employed to provide descriptive summaries and predictive insights for each cluster, enabling the identification of emerging technological trends and potential applications. The proposed approach enhances interpretability, supports strategic decision-making in research and development, and demonstrates a scalable method for AI-assisted patent landscape analysis.**

***Index Terms*—Patent Analysis, Sentence-BERT, UMAP, k-Medoids, Generative AI, Clustering, Technology Trend Prediction.**

I. Introduction

This study extends the framework of patent document analysis by combining Sentence-Bidirectional Encoder Representations from Transformers (Sentence-BERT) for vectorization, Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) for dimensionality reduction, and k-medoids for clustering, and integrating generative AI to enhance cluster interpretation and the extraction of strategic insights. The contemporary technological and business environment is characterized by volatility, uncertainty, complexity, and ambiguity (VUCA) [1], making it crucial to extract actionable knowledge from large-scale patent datasets to support decision-making in research and development (R&D) and technology management.

Patent documents are first transformed into high-dimensional vectors using a Transformer-based Sentence-BERT model [5], preserving sentence-level semantic relationships [2]. These high-dimensional embeddings are then projected into a lower-dimensional space using UMAP, enabling 2D and 3D visualization of the clusters. Visualization enables the capture of the structural features of each cluster, including key terms, peripheral concepts, distances to other clusters, and semantically dense regions.

While 2D visualization is suitable for quick overview and interpretation, 3D visualization provides a spatial representation of complex relationships and multi-layered structures, aiding the understanding of potential interactions between clusters and the direction of technological development [4]. By combining both approaches, users can intuitively grasp central and peripheral technologies within each cluster.

Furthermore, representative terms from each cluster are input into a generative AI model, which automatically produces descriptive summaries that encompass latent term relationships, related technologies, potential applications, and future technological trends [3]. This approach can reveal subtle relationships and emerging directions that may be overlooked by conventional co-occurrence analysis or statistical methods. The outputs from generative AI can be further evaluated and supplemented by experts to improve reliability and interpretability.

In addition, the framework allows exploration of inter-cluster relationships and scenario-based predictions, providing guidance for identifying promising technology areas and research directions worth pursuing. By integrating visualization and generative AI-based interpretation, the overall landscape of technological domains can be comprehensively understood, enabling the acquisition of practical knowledge and strategic insights that were not previously achievable.

II. Analytical Methods and Clustering Process

A. Overview of the Text Data Used in this Study

In this work, we performed extensive textual analysis on patents retrieved from Google Patents. Patents present unique challenges for natural language processing (NLP) due to their large volume, specialized terminology, and frequent use of complex compound expressions. Prior studies have highlighted the usefulness of patent data for identifying technological trends, informing R&D strategies, and performing competitive analyses, emphasizing the importance of systematic patent text mining. Each patent contains multiple layers of information —such as titles, abstracts, International Patent Classification

(IPC) codes, claims, and detailed descriptions—which increases the complexity of preprocessing.

Because Google Patents generates content dynamically, conventional static HTML scraping alone is insufficient for comprehensive data collection. To overcome this limitation, we automated browser-based operations using Selenium in combination with ChromeDriver and, using BeautifulSoup, extracted relevant sections, including titles, abstracts, IPC codes, claims, and detailed descriptions. Furthermore, patent documents often include extraneous information, such as variations in application numbers, company names, multilingual entries, and paragraph numbering. We therefore leveraged the HTML structure to selectively extract essential content, thereby improving data accuracy and reliability.

After collection, the text underwent normalization, symbol removal, and stopword filtering, followed by morphological tokenization. Japanese patents, in particular, contain numerous compound nouns and technical terms, which make dictionary-based tokenization inadequate for accurate word segmentation. To address this, we applied the termextract library to identify and standardize domain-specific terms and compound expressions, resulting in a vocabulary better suited for patent analysis. Additionally, the processed text was formatted for direct embedding with Sentence-BERT, preserving contextual information.

By combining dynamic acquisition, structured text extraction, and specialized term handling, we established a preprocessing pipeline that produces a high-quality dataset optimized for patent document analysis. This dataset subsequently supports downstream tasks, including embedding generation, dimensionality reduction, clustering, and semantic interpretation via generative AI, thereby enhancing the robustness and reliability of the overall analytical workflow.

B. Document Embedding

Patent documents accumulate continuously, resulting in an enormous volume that complicates efficient analysis. To address this challenge, it is effective to convert each patent into a numerical vector that captures its semantic content. In this study, we employed Sentence-BERT to obtain sentence-level embeddings for patent texts.

Sentence-BERT builds on the BERT architecture, generating sentence vectors that capture the overall meaning of sentences while preserving word order. BERT itself uses an encoder-only Transformer design and captures word relationships through attention mechanisms [6]. It is pretrained on Masked Language Modeling and Next Sentence Prediction tasks, providing general contextualized embeddings suitable for downstream applications such as classification, clustering, and semantic analysis.

Unlike traditional recurrent or convolutional networks, the Transformer leverages self-attention, enabling parallel computation and scalability across NLP tasks. Multi-Head Attention allows a model to simultaneously focus on multiple representation subspaces Eqs. 1, 2 and 3, while positional encoding injects sequence information into the embeddings Eqs. 4 and 5.

Multi-Head Attention

$$\begin{aligned}\text{Multi-Head Attention}(Q,K,V)\\ = \text{Concat}(head_1, head_2, \cdots, head_h)\,W_o\end{aligned} \tag{1}$$

$$\begin{aligned}\text{head}_i = ScaledDotProductAttention(QW_i^Q,\\ KW_i^K, VW_i^V)\end{aligned} \tag{2}$$

<Scaled Dot-Product Attention>

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

Positional Encoding

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \tag{4}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \tag{5}$$

To capture semantic similarities between multiple sentences, Sentence-BERT employs a Siamese network structure. Sentence embeddings generated by BERT are pooled and subsequently compared or classified using a Softmax layer. These embeddings serve as the foundation for identifying technological domains, tracking innovation trends, and feeding generative AI modules for further interpretation.

This modified approach extends our previous work by incorporating a more sophisticated preprocessing and embedding strategy, providing enhanced representation of patent semantics while facilitating downstream clustering, visualization, and semantic forecasting using LLMs.

C. Dimensionality Reduction with UMAP

The patent embeddings generated in this study are 768-dimensional vectors, which can lead to challenges related to the curse of dimensionality during clustering. To mitigate this issue and facilitate meaningful clustering, it is necessary to reduce dimensionality while retaining the data's essential geometric and topological structure. Dimensionality reduction techniques can be categorized as linear or nonlinear. While linear approaches are computationally efficient, they are often insufficient for capturing the complex nonlinear relationships present in patent embeddings. Nonlinear methods, on the other hand, can model these intricate structures but require higher computational resources.

In this study, we applied UMAP [7], a nonlinear dimensionality reduction technique that efficiently preserves both local and global relationships among high-dimensional vectors. UMAP constructs a weighted k-nearest neighbor graph for the high-dimensional data points and optimizes a low-dimensional representation that minimizes the discrepancy between the high- and low-dimensional proximities Eqs. 6. This process enables clustering algorithms to operate more effectively on reduced-dimensional data while preserving meaningful semantic relationships.

$$L = \sum_{i,j} \left[v_{ij} \log \frac{v_{ij}}{w_{ij}} + (1 - v_{ij}) \log \frac{1 - v_{ij}}{1 - w_{ij}} \right] \tag{6}$$

D. Clustering with k-Medoids

To organize the high-dimensional patent embeddings obtained in this study, we used the k-medoids clustering algorithm. Clustering helps to identify groups of patents with similar semantic features by assigning a representative center for each cluster [8]. Unlike k-means, k-medoids chooses the cluster center from actual data points, called medoids, minimizing the sum of distances to all other points in the cluster. This approach is remarkably robust against outliers and irregular data distributions:

$$\mathbf{m}_k = \arg \min_{\mathbf{x}_j \in X_k} \sum_{\mathbf{x}_i \in X_k} d(\mathbf{x}_i, \mathbf{x}_j) \tag{7}$$

where $d(\mathbf{x}_i, \mathbf{x}_j)$ represents a chosen distance metric. The optimal number of clusters was determined by silhouette analysis [9], which evaluates both intra-cluster cohesion $a(i)$ and inter-cluster separation $b(i)$ for each data point $x(i)$. The silhouette coefficient $s(i)$ is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{8}$$

We selected the number of clusters that maximized the average silhouette coefficient. In this study, silhouette scores were calculated for cluster numbers ranging from 2 to 30, and the cluster configuration yielding the highest score was adopted for subsequent analysis.

E. Analysis and Visualization of Clustered Patent Terms

To gain insights into the characteristics of patent clusters, we examined term co-occurrence patterns within each cluster. Co-occurrence indicates how frequently specific terms appear together in the same context, which provides information about their semantic relationships [10]. In this study, we used several similarity indices to quantify co-occurrence, including the Jaccard, Dice, and Simpson coefficients. The Jaccard coefficient measures the proportion of shared terms between sets A and B:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{9}$$

The Dice coefficient evaluates the shared terms relative to the average size of the two sets [11]:

$$DSC(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{10}$$

Additionally, the Simpson coefficient considers the overlap relative to the smaller set [12]:

$$\text{overlap}(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)} \tag{11}$$

The quantified co-occurrence data were visualized as networks to facilitate intuitive interpretation of term relationships. For 2D visualizations, we used the pyvis library, which allows flexible customization of node and edge attributes and interactive exploration of the network. To represent a larger amount of information spatially, 3D visualizations were created using Three.js. While 3D representations can increase information density, potentially reducing readability, combining them with 2D visualizations ensures both detailed and comprehensible views. This integrated visualization approach supports the interpretation of patent cluster semantics using both quantitative and visual perspectives.

III. AI-Based Cluster Interpretation and Future Prediction

A. Positioning of Generative AI

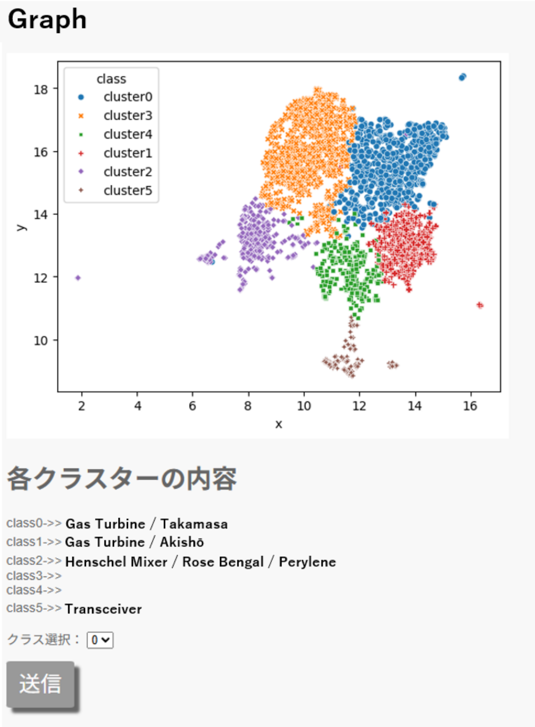
In conventional cluster analysis, the interpretation of technological characteristics has primarily relied on word frequencies and co-occurrence relationships within clusters. However, this approach relies solely on word occurrences and co-occurrences, making it difficult to understand the semantic background and latent relationships within clusters fully. Especially for high-dimensional, specialized datasets such as patent documents, simple frequency-based information may not capture key aspects, including technological applicability, research trends, or societal impact. Moreover, manual interpretation heavily relies on domain expertise and time, posing challenges for reproducibility and efficiency.

The use of generative AI offers an effective means to overcome these limitations. Generative AI can automatically generate natural-language summaries and explanations from the words and document information within a cluster, enabling an intuitive and comprehensive understanding of the cluster's technological meaning and potential applications [13]. In particular, large language models that have learned contextual relationships between words can infer subtle differences in meaning and relationships that may be overlooked by human experts, thereby improving both the accuracy and depth of cluster interpretation.

Furthermore, leveraging generative AI enables automation and efficiency in the analysis process. Previously, representative words for each cluster had to be manually organized and interpreted. Generative AI can summarize and analyze information across thousands of clusters almost instantaneously, enabling comprehensive interpretations of large-scale patent datasets within realistic time frames. In this way, generative AI serves as a powerful tool that balances improved interpretive accuracy with analytical efficiency.

B. Generation of Cluster Descriptions

Generative AI-based cluster description involves creating natural language summaries of clusters using representative, frequently occurring words as input. Specifically, prompts such as "This cluster contains words AAA, BBB, and CCC. Describe the characteristics, related technologies, and potential applications of this cluster in detail" are used to guide the



各クラスターの内容

class0->> Gas Turbine / Takamasa
class1->> Gas Turbine / Akishō
class2->> Henschel Mixer / Rose Bengal / Perylene
class3->>
class4->>
class5->> Transceiver

クラス選択：0

Fig. 1. Cluster display interface

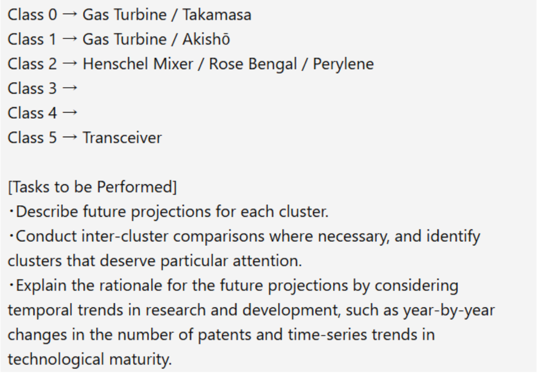
model, enabling output that captures the technological content and scope of the cluster [14].

This approach allows a comprehensive understanding of cluster features and enables applications such as inter-cluster comparison and technological domain classification. Since generative AI considers contextual relationships between words, it can reveal latent relationships and technological directions that conventional statistical methods cannot detect. Additionally, the generated cluster descriptions can be validated and refined by human experts, ensuring accurate and reliable interpretation.

C. Future Prediction and Insight Extraction for Each Cluster

The words within each cluster and their relationships can be organized in textual or tabular formats, which can serve as input for generative AI models. Specifically, representative words, frequently occurring words, and strongly co-occurring word pairs are listed, and the relationships and characteristics of the cluster are summarized in text. By inputting this organized information into a generative AI model, latent patterns, novel ideas, and related technologies within the cluster can be extracted in natural language. This method allows an intuitive, comprehensive understanding of cluster structures and provides insights that traditional statistical analyses alone cannot deliver.

Furthermore, using template formats to organize input information enhances the reproducibility and efficiency of cluster interpretation. Generative AI can also generate future trend predictions in textual form, serving as a reference for decision-makers and researchers to formulate technology strategies quickly. This approach advances beyond conventional cluster analysis, which primarily provides technological categorization, to offer direct support for strategic decision-making and the setting of research and development directions.



Class 0 → Gas Turbine / Takamasa
Class 1 → Gas Turbine / Akishō
Class 2 → Henschel Mixer / Rose Bengal / Perylene
Class 3 →
Class 4 →
Class 5 → Transceiver

[Tasks to be Performed]
・Describe future projections for each cluster.
・Conduct inter-cluster comparisons where necessary, and identify clusters that deserve particular attention.
・Explain the rationale for the future projections by considering temporal trends in research and development, such as year-by-year changes in the number of patents and time-series trends in technological maturity.

Fig. 2. Examples of templates to input into generative AI

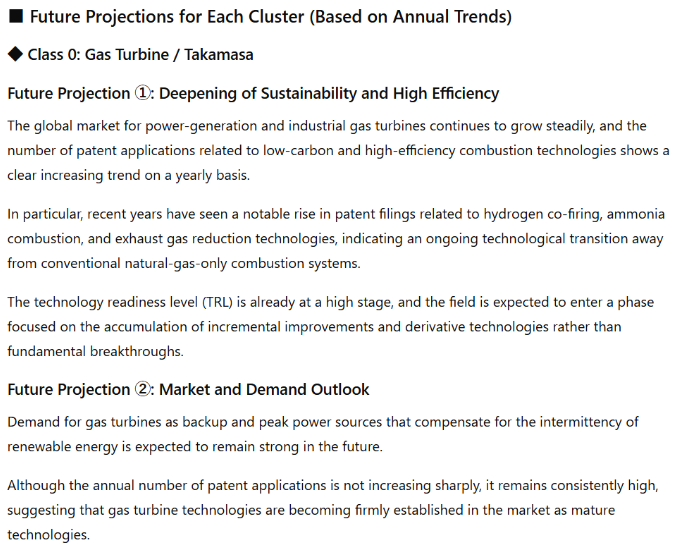
IV. Results and Discussion

A. Results of the End-to-End Processing Pipeline from Preprocessing to Clustering

In this study, we developed an integrated processing pipeline that collects patent information from Google Patents using multiple user-provided keywords, performs document vectorization with Sentence-BERT, applies dimensionality reduction with UMAP, and conducts clustering with the K-medoids method. This section describes the entire preprocessing procedure, including the web scraping stage and the resulting clustering outcomes.

First, an OR-based keyword search is executed on Google Patents to retrieve relevant patent identifiers. Because Google Patents imposes a limit of 1,000 records per request, the present study divides the target period into one-year intervals and assigns each year to an individual thread. This parallel scraping strategy enables efficient collection of patent texts across multiple years, and the aggregated results are combined to construct the final patent dataset.

Next, the collected text data are encoded into document vectors using Sentence-BERT. The resulting high-dimensional vectors are then reduced to 15 and 2 dimensions using UMAP. The 15-dimensional vectors are used for clustering and interpreting cluster contents, as this representation retains sufficient semantic information while suppressing noise, resulting in stable clustering. In contrast, 2-dimensional vectors are used to visualize the clustering results. The reason for adopting 2-dimensional vectors for visualization is that, compared with three-dimensional or higher-dimensional vector spaces, the distances and density relationships between data points plotted in a 2D space can be more intuitively perceived by humans. Furthermore, in visualizing similarities among data points, a 2D space makes it easier to recognize how data points are grouped within each cluster. It provides a more straightforward overview of the dataset's overall structure. The UMAP



■ Future Projections for Each Cluster (Based on Annual Trends)

◆ Class 0: Gas Turbine / Takamasa

Future Projection ①: Deepening of Sustainability and High Efficiency

The global market for power-generation and industrial gas turbines continues to grow steadily, and the number of patent applications related to low-carbon and high-efficiency combustion technologies shows a clear increasing trend on a yearly basis.

In particular, recent years have seen a notable rise in patent filings related to hydrogen co-firing, ammonia combustion, and exhaust gas reduction technologies, indicating an ongoing technological transition away from conventional natural-gas-only combustion systems.

The technology readiness level (TRL) is already at a high stage, and the field is expected to enter a phase focused on the accumulation of incremental improvements and derivative technologies rather than fundamental breakthroughs.

Future Projection ②: Market and Demand Outlook

Demand for gas turbines as backup and peak power sources that compensate for the intermittency of renewable energy is expected to remain strong in the future.

Although the annual number of patent applications is not increasing sharply, it remains consistently high, suggesting that gas turbine technologies are becoming firmly established in the market as mature technologies.

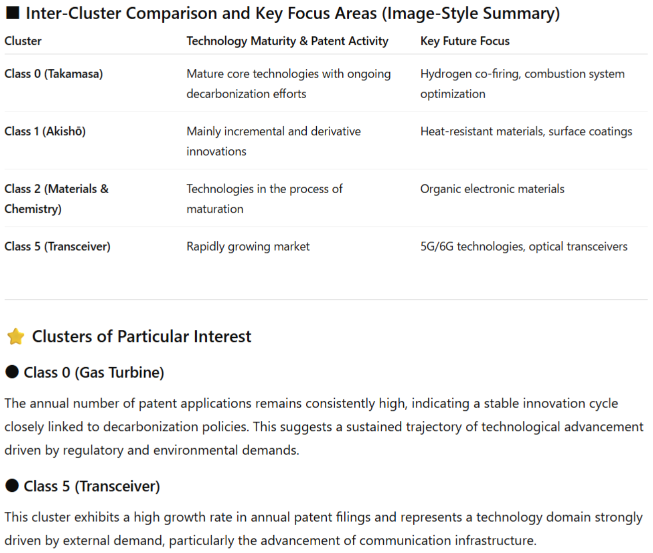
Fig. 3. Future Projections by Cluster (Cluster 0)

parameters were tuned to preserve both the local and global structures of the patent corpus.

Subsequently, silhouette analysis was conducted on the 15-dimensional vectors to estimate the optimal number of clusters, followed by K-medoids clustering. K-medoids was chosen over K-means because it is more robust to outliers and defines cluster centers as actual data samples, enabling more interpretable, stable clustering results in non-Euclidean vector spaces such as those produced by Sentence-BERT. To facilitate interpretation, the ten documents closest to each cluster medoid were extracted, and essential terms were identified through the termextract library. Based on this, three representative terms were selected for each cluster, enabling explicit characterization of the underlying themes and technological domains.

The clustering output was then visualized as a 2D scatter plot using pyvis, with each cluster assigned a distinct color and marker shape to improve interpretability. The visualization also displays representative terms for each cluster, enabling users to intuitively select areas of interest (refer to Fig. 1). After a cluster is selected, a co-occurrence network is constructed using the Simpson coefficient, and both 2D and 3D graphs of the term network are generated.

Overall, the proposed system provides a fully automated end-to-end pipeline that spans keyword input, patent retrieval, text preprocessing, dimensionality reduction, clustering, and visualization. This enables users to intuitively grasp cluster structures and their content, while seamlessly progressing to more detailed analyses of specific technical fields. The integrated results presented in this section also serve as essential inputs for subsequent stages, such as future-trend prediction and idea-generation support, contributing significantly to exploratory analysis in targeted technological domains.



■ Inter-Cluster Comparison and Key Focus Areas (Image-Style Summary)

Cluster	Technology Maturity & Patent Activity	Key Future Focus
Class 0 (Takamasa)	Mature core technologies with ongoing decarbonization efforts	Hydrogen co-firing, combustion system optimization
Class 1 (Akishō)	Mainly incremental and derivative innovations	Heat-resistant materials, surface coatings
Class 2 (Materials & Chemistry)	Technologies in the process of maturation	Organic electronic materials
Class 5 (Transceiver)	Rapidly growing market	5G/6G technologies, optical transceivers

⭐ Clusters of Particular Interest

● Class 0 (Gas Turbine)

The annual number of patent applications remains consistently high, indicating a stable innovation cycle closely linked to decarbonization policies. This suggests a sustained trajectory of technological advancement driven by regulatory and environmental demands.

● Class 5 (Transceiver)

This cluster exhibits a high growth rate in annual patent filings and represents a technology domain strongly driven by external demand, particularly the advancement of communication infrastructure.

Fig. 4. Comparison results by cluster

B. Future Prediction of each Cluster Using Generative AI

In this study, we conducted clustering on a set of patents containing the terms “wind” and “turbine” to identify the technological characteristics of each cluster. Subsequently, we used generative AI to predict future trends. The purpose of the future prediction was to estimate the developmental directions of technological domains for each cluster, providing insights valuable for research and development strategies and investment decision-making.

Inputs to the generative AI were formatted according to the template shown in Figure 2, with three representative terms describing each cluster. The template instructs the AI to define the future technological trends of each cluster, compare clusters when necessary, and explain the rationale for its predictions based on historical trends such as changes in patent filings or the maturity of related technologies. This approach succinctly conveys the key points of the cluster while facilitating the AI’s inference of the underlying technological context. The generative AI, based on statistical knowledge derived from past technical literature, news, and patent trends, generated textual outputs predicting future trends for each cluster (refer to Fig. 3). Figure 3 shows an example output for Cluster 0, which summarizes AI-generated predictions in Japanese regarding short- to mid-term (3–5 years) technological developments, supporting indicators such as patent filing trends and citation counts, and suggested strategic actions for researchers and companies. The outputs were organized according to the following aspects:

- Direction of annual trends
- Emerging research topics
- Expected application areas
- Societal and industrial impact
- Technical challenges and potential risks

Through this process, the generative AI provided concrete descriptions of the anticipated 5-10 years development potential and possible barriers to adoption for each cluster. The outputs also supported inter-cluster comparisons, highlighting clusters with high development potential as recommended areas (refer to Fig. 4). In this way, the future prediction results function not merely as supplementary information but as a guide for identifying technological domains likely to become strategically important.

By integrating these predictions, we organized the technological positioning of each cluster, providing a framework for constructing technology scenarios in the wind power domain. Notably, some clusters showed a clear upward trend in patent filings, indicating potential market expansion in the coming years. In contrast, mature domains faced challenges in technological differentiation, suggesting a need to shift focus to peripheral technologies such as material performance and environmental resilience.

C. Visualization of Selected Clusters and Idea Generation with Generative AI

For user-selected clusters, we perform 2D and 3D visualizations using UMAP and conduct a detailed analysis of the structural characteristics of term groups. These visualizations highlight key terms defining the cluster, peripheral concepts, relative distances to other clusters, and regions with high semantic density within the cluster, thereby enabling an intuitive understanding of the internal structure of the technical domain and the relationships among terms.

In particular, 3D visualizations can represent complex relationships and multi-layered structures in space, helping capture potential interactions between clusters and understand the developmental directions of technologies. In contrast, 2D visualizations have lower information density, making them suitable for grasping the overall structure and for rapid interpretation. By combining both 2D and 3D visualizations, the precision and efficiency of visual understanding can be enhanced.

Furthermore, in this study, generative AI is leveraged to support idea generation and technological interpretation of clusters. Specifically, representative and frequently occurring terms from each cluster are fed into a language model, which automatically generates descriptive summaries that capture latent relationships among terms, related technologies, potential applications, and future technological trends.

This approach enables the extraction of subtle inter-cluster relationships and novel technological directions that conventional co-occurrence analysis or statistical methods may fail to reveal. Moreover, the outputs of generative AI can be evaluated and refined by human experts, improving both the reliability and interpretive accuracy of the analysis.

V. CONCLUSION

In this study, we developed an end-to-end analytical pipeline that integrates patent retrieval, text vectorization, dimensionality reduction, clustering, visualization, and generative AI–based interpretation. By combining Sentence-BERT embeddings, UMAP, and K-medoids clustering, the proposed system successfully extracted meaningful cluster structures from patents related to wind-turbine technologies. Representative terms were identified for each cluster, enabling explicit characterization of technological domains and facilitating the interpretability of the clustering results.

Furthermore, by employing generative AI to predict future developments for each cluster, we demonstrated the applicability of large language models for estimating emerging technological directions, potential application areas, and expected challenges. The results provided valuable insights for technology forecasting and strategic decision-making. Visualization with both 2D and 3D UMAP further enhanced interpretability by enabling intuitive exploration of cluster structures and inter-cluster relationships.

Overall, the proposed pipeline significantly streamlines the process of exploratory patent analysis and supports early-stage R&D planning by integrating automated data processing with AI-driven interpretation. Future work includes expanding the data sources to incorporate scientific publications and real-world news, integrating temporal trend analysis, and developing an interactive decision-support system for industrial and academic users.

REFERENCES

[1] B. Taskan, A. Junça-Silva and A. Caetano, “Clarifying the conceptual map of VUCA: a systematic review”, *International Journal of Organizational Analysis*, pp. 196-217, 2022

[2] K. Han, A. Xiao and E. Wu, et al. “Transformer in Transformer”, pp. 1-12, 2023,

[3] Y. Cao, S. Li and Y. Liu, et al., “A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT”, arXiv:2303.04226 [cs.AI], 2023

[4] G. Perrone, J. Unpingco and H. Lu, “Network visualizations with Pyvis and VisJS”, arXiv:2006.04951 [cs.SI], 2020

[5] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embedding using Siamese BERT-Networks”, *ArXiv e-prints*, 1908. 10084, 2019

[6] J. Devlin, M. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, vol. 1, pp. 4171–4186, 2019

[7] L. Mclnnes, J. Healy, J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”, 1802. 03426, 2018

[8] S. M. R. Zadegan, M. Mirzaie and F. Sadoughi, “Ranked k-medoids: A fast and accurate rank-based partitioning algorithm for clustering large datasets”, *Knowledge-Based Systems*, vol. 39, pp. 133-143, 2013.

[9] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”, *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53-65, 1986

[10] Y. MAatsuo and M. Ishizuka, “Keyword Extraction from a Single Document Using Word Co-Occurence Statistical Information”, *International Journal on Artificial Intelligence Tools*, vol. 13, No. 1, pp. 157-169, 2004

[11] L. R. Dice, “Measures of the Amount of Ecologic Association Between Species”, *Ecology*, vol. 26, No. 03, 1945, pp. 297-302

[12] Robert K. Peet, “The Measurement of Species Diversity”, *Annual Review of Ecology and Systematics*, vol. 5, pp. 285-307, 1974

[13] T. Brown, B. Mann, N. Ryder, et al., “Language Models are Few-Shot Learners”, *Advances in Neural Information Processing Systems*, 33, pp. 1-25, 2020

[14] Z. Wang, J. Shang and R. Zhong, “Goal-Driven Explainable Clustering via Language Descriptions”, arXiv preprint arXiv:2305. 13749, 2023