Visualization and Analysis of Multidisciplinary Research Networks

A Case Study of UNIST's College of Information & Biotechnology

Sumin Park

School of Industrial Engineering
UNIST

suminpark@unist.ac.kr

Abstract

This study addresses the challenges of identifying relevant research labs amidst the rise of interdisciplinary research, which has blurred academic boundaries. Focusing on UNIST's College of Information & Biotechnology, known for fostering interdisciplinary research, we collected data from lab web pages and research databases, including URLs, professor names, research fields, interests, descriptions, and recent publications. After preprocessing for consistency, we applied text mining techniques, comparing SCI-BERT, BERT, and TF-IDF. Despite BERT achieving the highest silhouette score, SCI-BERT was selected for its optimization with scientific texts. We employed K-means clustering, favored over DBSCAN due to parameter tuning and performance issues. The analysis revealed distinct clusters between information and bio-related research, with SCI-BERT and K-means effectively visualizing these trends. Notably, Cluster 8 was identified as relevant for students interested in Human-Computer Interaction (HCI), demonstrating the method's practical utility. This approach aids students in lab selection, enhances understanding of research trends, and promotes interdisciplinary collaboration. Future work will improve data quality, explore diverse clustering algorithms, establish objective evaluation criteria, and compare global research trends.

I. Introduction

Recent interdisciplinary research has shown that the content of laboratory research is often not confined to a single department. This trend indicates that the boundaries of research are becoming increasingly blurred, leading to the emergence of new fields through the fusion of various disciplines. This phenomenon is closely related to the modern research trend of leveraging knowledge from diverse fields to solve complex problems. However, this trend also presents a challenge in comprehensively understanding the research content of different labs. For example, a student interested in Human-Computer Interaction (HCI) must explore labs across multiple departments such as design, industrial engineering, computer science, and biomedical engineering. This process can be time-consuming and confusing for students. The task of visiting each department's website or meeting with professors to understand their research topics is inefficient and can pose significant difficulties for students when choosing a lab.

UNIST's College of Information & Biotechnology is a unique school that integrates 'information' and 'biotechnology'. This college encourages interdisciplinary research and acts as a hub for such studies. In this environment, there is a high likelihood of active collaboration between labs, providing students with diverse research opportunities. However, the integrated structure can make it even more challenging to identify labs that focus on specific research topics.

II. Literature Review

Graph data visualization has been employed to effectively understand research trends across labs. This method represents the research topics and interests of each lab as nodes and edges, making the information visually clear. Graph visualization aids in intuitively understanding complex data and has the advantage of allowing users to see the relationships between different labs at a glance. Previous studies, such as the KCSS (KAIST Collaborative Study System) developed by the ALIN lab at KAIST, have already utilized graph data to visualize research trends. This system helps researchers easily find others conducting similar research and plays a significant role in promoting interdisciplinary collaboration【KAIST ALIN KCSS】. However, these studies primarily classified research topics based on the keywords of papers published under professors' names.

Fig1 : KCSS (https://alinlab.kaist.ac.kr/KCSS)

In this study, I aim to analyze the research topics and goals defined by the labs themselves with higher weighting. This

approach is taken because the research goals set by the labs directly reflect their research direction more accurately, enabling a more precise analysis of research trends. This approach will help in better understanding the unique research goals and interests of the labs and increase the potential for collaboration between them. Our study also offers an opportunity to assess whether the education, research, industry and entrepreneurship, and globalization goals of the UNIST School of Information and Biotechnology align with its strategic objectives. This assessment can serve as a crucial reference for setting future research directions by evaluating the congruence between the college's strategic goals and actual research activities.

Other research trend analyses use statistical analysis to examine the research trends of universities or specific fields. For example, the Academic Relationship Analysis Service (https://sam.riss.kr/organResearch.do) analyzes research topics at universities and visualizes collaborations in graph form. However, this service does not include UNIST. Additionally, existing studies often classify research based on the keywords of papers published under professors' names, which may not fully reflect the unique research goals and directions of the labs.

To overcome these limitations, our study plans to analyze the research topics and goals defined by the labs themselves with higher weighting. This approach is taken because the unique research goals of the labs more accurately reflect their research directions. As a result, we can achieve a more precise analysis of research trends and contribute to increasing the potential for collaboration between labs.

III. Method

1. Data

Data was collected from the web pages of various laboratories at UNIST and research databases. The collected items include:

Laboratory URL , Professor's name, Laboratory name, Research field, Research interests, Laboratory description, Titles and keywords of the 10 most recent papers (92 labs)

The data was cleaned for consistency by removing special characters, converting text to lowercase, and standardizing terms.

2. Method

In this study, we compared three text analysis techniques: SCI-BERT, BERT, and TF-IDF. To evaluate the characteristics and performance of each method and determine the most effective analysis approach, we followed the procedure outlined below.

The features and advantages of each candidate model are as follows:

1. SCI-BERT

- Justification: SCI-BERT is a pre-trained language model based on scientific paper data, which allows it to understand scientific terms and expressions well. This capability reflects the specificity of scientific papers, enabling accurate keyword extraction.

- Suitability: SCI-BERT can accurately understand the contextual meaning of scientific papers to extract key keywords, making it highly suitable for the purpose of this study.

2. BERT

- Justification: BERT is a general-purpose language model pre-trained on general text, capable of understanding the meaning of words by considering the context before and after the word. This provides the advantage of understanding a variety of topics and contexts.

- Suitability: BERT, being a versatile model, is useful for analyzing papers on diverse topics.

3. TF-IDF

- Justification: TF-IDF is a technique that calculates the importance of each word in a document based on the word's frequency and inverse document frequency. It has the advantage of being relatively simple and fast to compute.

- Suitability: While TF-IDF has the limitation of not understanding contextual meaning, it is useful for simple frequency-based keyword analysis.

By comparing SCI-BERT, BERT, and TF-IDF, we can identify the strengths and weaknesses of each method and select the most effective keyword analysis approach. We used TF-IDF, BERT, and SCI-BERT as text embedding methods and evaluated the clustering performance of each method using silhouette scores. The silhouette score measures the cohesion and separation of clusters, with higher scores indicating better clustering quality.

- TF-IDF: 0.0148

- BERT: 0.1094

- SCI-BERT: 0.0597

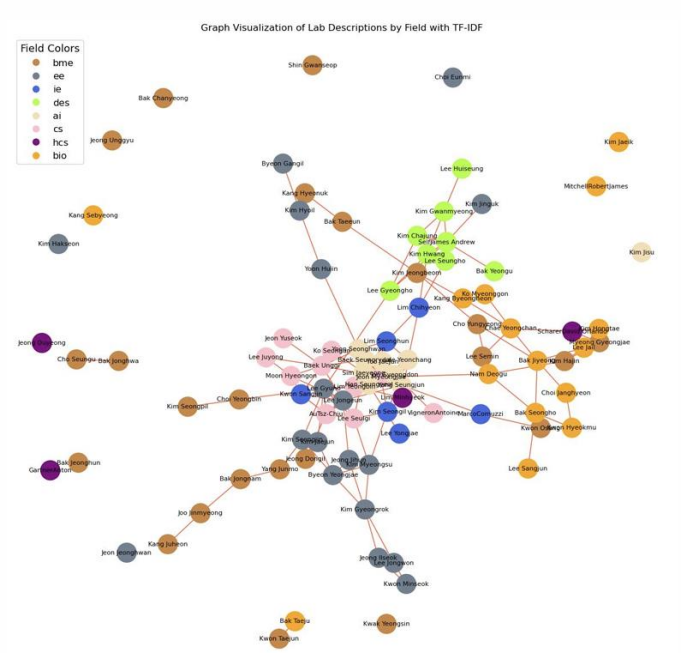

Fig2 : Graph of Lab Description by TF-IDF

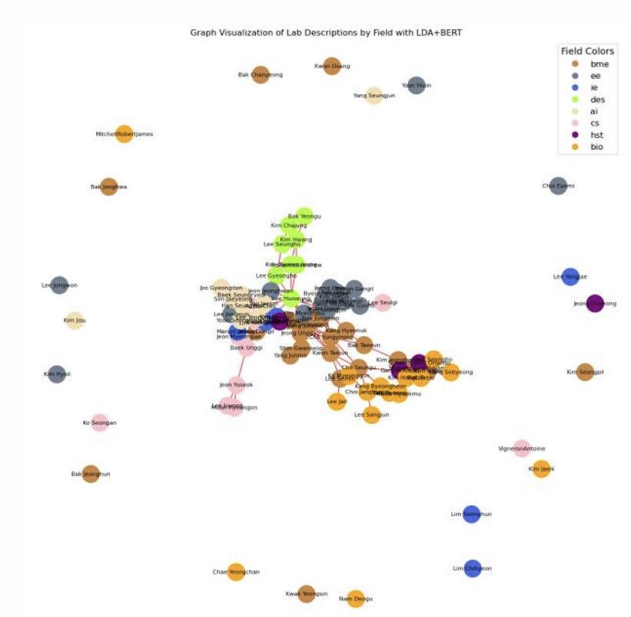

Fig3: Graph of Lab Description by BERT

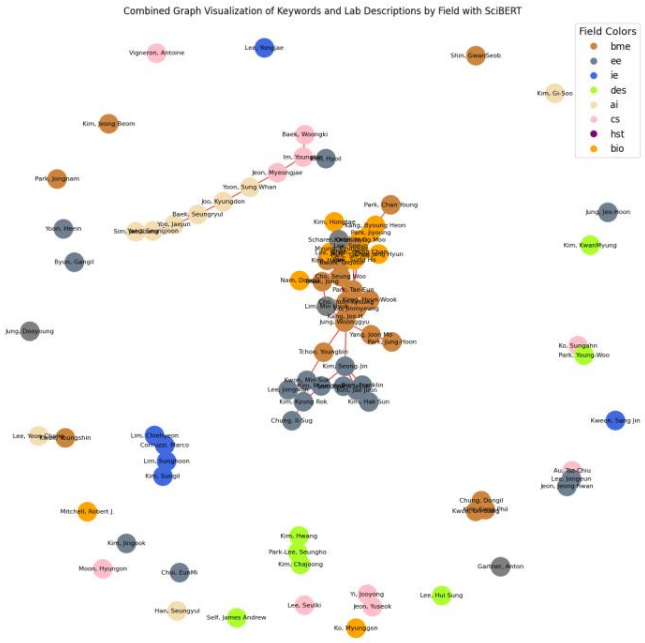

Fig4: Graph of Lab Description by SCI-BERT

Although BERT recorded the highest silhouette score, we ultimately chose SCI-BERT to reflect the specificity of scientific paper data. SCI-BERT is optimized for scientific paper texts, allowing it to more accurately reflect domain-specific keywords and research topics.

In this study, we selected two representative clustering algorithms, K-means and DBSCAN, to visualize and cluster the results of scientific paper keyword analysis. We compared these two algorithms, evaluated the pros and cons of each, and ultimately chose K-means.

1. K-means: A. K-means is an unsupervised learning algorithm that partitions data into K clusters. Each cluster has a centroid, and data points are assigned to the nearest centroid. B. It is fast to compute and efficient for large datasets. Consistent results can be obtained if the initial number of clusters is specified. C. The number of clusters must be specified in advance, and it is sensitive to outliers.

2. DBSCAN (Density-Based Spatial Clustering of Applications with Noise): A. DBSCAN is a density-based clustering algorithm that forms clusters in high-density areas and treats low-density areas as outliers. B. The number of clusters does not need to be specified in advance, and it can discover clusters of arbitrary shapes. It is robust to outliers. C. Appropriate values for the density parameter (ε) and the minimum number of points (MinPts) must be set, and performance can degrade if data density varies significantly.

We initially considered DBSCAN to be more advantageous than K-means because DBSCAN treats low-density areas as outliers, which was expected to be useful for complex datasets like scientific paper data. Additionally, DBSCAN can discover clusters of arbitrary shapes, making it effective for clustering data with various topics and forms. However, practical experimentation revealed difficulties in applying DBSCAN. It was challenging to set appropriate values for ε and MinPts, and the performance degraded due to the uneven density of the data.

In contrast, the K-means algorithm provided appropriate results for the following reasons:

1. K-means produced meaningful clusters, unlike DBSCAN, which means that all data points could be included in clusters.

2. It allows for the selection of specific clusters: specific clusters generated by the K-means algorithm can be selected to introduce relevant research labs to users. This is very useful for understanding research trends and finding related labs.

Therefore, we decided to use the K-means algorithm to cluster the paper keyword data. K-means is suitable for generating consistent clusters and selecting specific clusters, making it effective for introducing research trends and related labs to users.

IV. Experiment

The experimental process of this study was conducted in the following sequence: data preprocessing, text mining, visualization analysis, and clustering. During the visualization analysis phase, we employed dimensionality reduction techniques to visualize high-dimensional vector data. Specifically, we used T-SNE (t-Distributed Stochastic Neighbor Embedding) and PCA (Principal Component Analysis) to reduce the data to two or three dimensions. This reduced data was then visualized in 2D or 3D space to visually confirm the clustering structure of the data.

V. Results and Discussion

We visualized and analyzed the results of scientific paper keywords and lab descriptions separately, as well as the combined data. During the visualization process, edges were created when the similarity score between nodes exceeded 0.815.

The visualization of paper keywords showed a more cohesive structure overall, with particularly high similarity in the bio-related fields. This indicates that bio-related papers are clustered around specific keywords. In contrast, the lab descriptions visualization showed that the descriptions of labs in CS, AI, and EE fields were clustered together, indicating shared research interests in these fields. The combined visualization of paper keywords and lab descriptions revealed well-separated clusters for information and bio-related fields, demonstrating distinct research topics and keywords in each area.

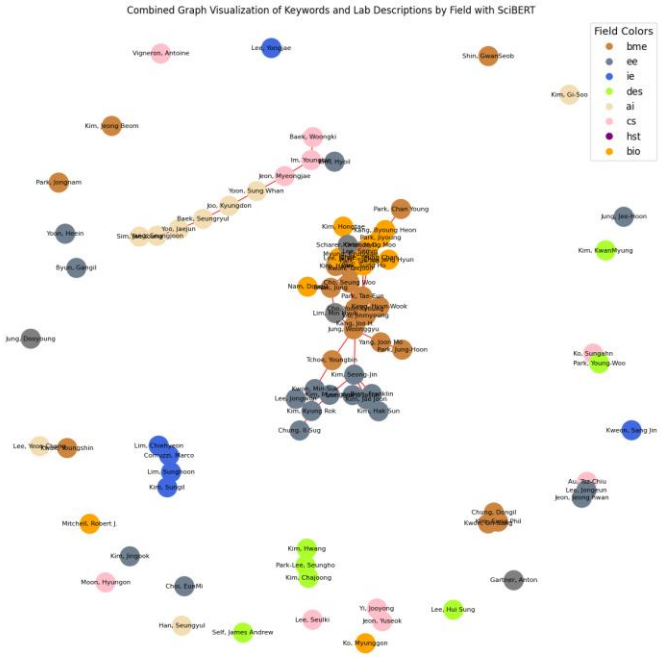

Fig5: Graph of Keywords and Lab Descriptions

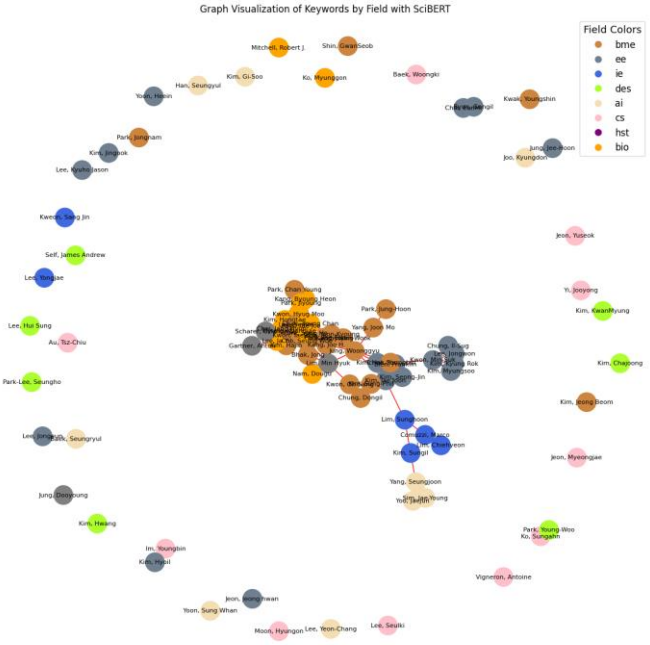

Fig6: Graph of Keywords

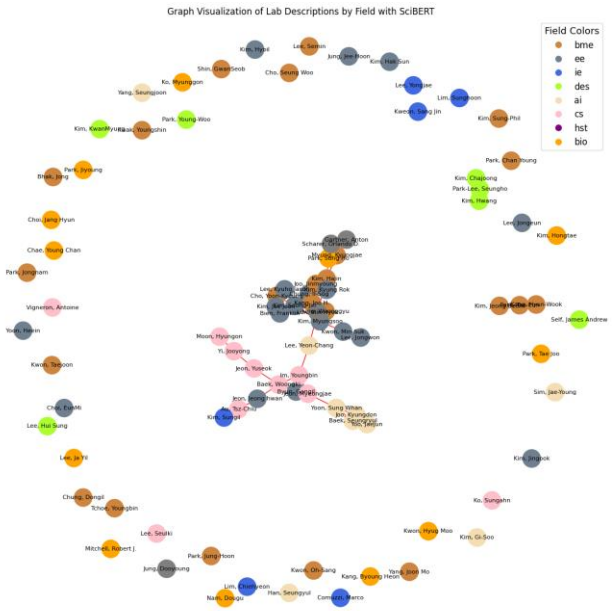

Fig7: Graph of Lab Descriptions

The 3D dimensionality reduction results further clarified these distributions. Labs within the same department were generally located close to each other and connected, while labs from different departments but with similar research topics were also connected. This illustrates both the similarity of research topics within departments and the connections between similar research topics across departments.

Notably, clustering based on paper keywords and lab descriptions clearly distinguished between the research in information and bio fields. This suggests that the self-defined research goals of the labs yield different results from the keyword-based trends observed in existing research, highlighting the unique goals and directions of the labs.

The visualization results confirmed that using SCI-BERT for K-means clustering is effective in understanding UNIST's research trends. The clustering based on paper keywords and lab descriptions clearly showed the distinction between research in the information and bio fields, making it easier for researchers to understand research trends and find related labs.

This study also includes search results for students interested in Human-Computer Interaction (HCI). Analysis indicated that labs in Cluster 8 are the most suitable, with keywords such as 'interaction, mobility, future, electronic, human-robot, driving, theory, robots, robotics, self'. The following labs belong to Cluster 8. These results suggest that labs share similar research topics and keywords beyond departmental boundaries, demonstrating the importance and activation of interdisciplinary research. Particularly for students seeking HCI-related research, labs in Cluster 8 may be suitable.

Comparing clusters with existing departments showed that the Bio department has the most concentrated research areas, while the Computer Science (CSE), Electrical Engineering (EE), and Artificial Intelligence (AI) departments are more dispersed. The departments in the same cluster could be divided into two groups: the first group consists of Bio, Biomedical Engineering (BME), and Human Cognitive Science (HCS) departments, while the second group includes AI, Computer Science (CS), and EE departments, with some Design (DES) and Industrial Engineering (IE) labs also included in the second group.

Fig8: Relationship between Clusters and Fields

VI. Conclusion

This study analyzed and visualized research trends at UNIST by using SciBERT embeddings and K-means clustering to analyze the keywords and descriptions of research labs. This approach reflects the characteristics of the labs and presents a method to form interpretable clusters.

6.1 Justification of Learning and Analysis Methods:

In this study, we used SciBERT embeddings and K-means clustering. SciBERT, a pre-trained model specialized in scientific texts, more accurately reflects the characteristics of research labs compared to simpler methods like TF-IDF or LDA. K-means clustering is a straightforward and effective algorithm for grouping data, making it suitable for showing the similarities between research labs. We chose these methods because UNIST's research lab data is relatively structured, necessitating accurate clustering based on this structure.

6.2 Methodology and Model Description:

SciBERT is a pre-trained model based on BERT and further trained on scientific paper data, enabling it to better understand and reflect the context of scientific texts. K-means clustering maps each data point to a vector space and groups similar data points into clusters. In this study, we vectorized the keywords and descriptions of each research lab using SciBERT and then performed clustering with the K-means algorithm.

6.3 Value of the Research:

The significance of this study is presented from two perspectives:

Value to UNIST Students: This study can help students choose the appropriate department and research lab. For example, students interested in HCI can refer to relevant clusters to find labs that match their research interests. Analyzing lab keywords and descriptions can help students clarify their research interests and create more specific research plans. Visualizing the active research landscape enables students to understand the connections between various research fields and recognize the importance of collaborative research.

Value to Academic Stakeholders: This study provides a method to clearly understand research trends within UNIST. Academic stakeholders can use this information to comprehend current research directions and inform future research planning. Additionally, visualizing research interests and characteristics across departments offers motifs that can aid in department classification and research evaluation. This can be useful for assessing research capabilities and identifying areas for improvement. The results emphasize the importance of interdisciplinary research and promote collaborative research opportunities.

6.4. Limitations:

Data Collection Issues:There may be missing or incomplete information in the data collection process. Data collected through web scraping may not be perfect and might not reflect the latest information from some labs.

Clustering Algorithm Limitations: The results of the K-means clustering algorithm can vary based on the initial cluster centroids. This can lead to different clustering outcomes for the same dataset, affecting the reproducibility of the results.

Subjective Evaluation Involvement: The model selection process and the selection of various parameters involve subjective evaluation. The researcher's subjectivity may influence the evaluation of the visualization results. To overcome these subjective limitations, it is necessary to introduce objective evaluation criteria for visualization research.

Lack of Generalizability of the Model: One reason for using SciBERT and K-means clustering is the ease of visualization. However, this is based on subjective evaluation, and using other embedding models or clustering algorithms may yield different results.

6.5. Future Plans:

Improve Data Quality: We will enhance the data collection process to gather more reliable data. Besides web scraping, we will incorporate data directly provided by research labs and the latest paper data to improve accuracy.

Compare Various Clustering Techniques: We will apply and compare different clustering algorithms besides K-means to identify the most suitable algorithm. For example, we plan to experiment with DBSCAN, hierarchical clustering, and other methods.

Introduce Evaluation Criteria: To overcome the subjective evaluation of visualization research, we will introduce objective evaluation criteria. For instance, we will establish and apply evaluation standards based on the accuracy, interpretability, and user feedback of the visualization.

Compare Global Research Trends: We will verify the research results by comparing them with trends at other universities or global research trends. This will help us understand UNIST's research trends from a broader perspective and analyze differences from global trends.

In conclusion, using SciBERT embeddings and K-means clustering, we were able to understand research trends based on the keywords and descriptions of research labs. SciBERT embeddings, specialized for scientific texts, effectively reflected the characteristics of the labs, and K-means clustering was effective in forming interpretable clusters. This analysis method can help UNIST students choose departments and research labs, propose methods for analyzing interdisciplinary research, and provide motifs for department classification. Moving forward, we will continue to analyze research trends across various academic fields to provide a better research environment.

REFERENCES

[1] ALIN Lab. (n.d.). KAIST Collaborative Study System (KCSS). Retrieved from https://alinlab.kaist.ac.kr/KCSS

[2] RISS. (n.d.). Academic Relationship Analysis Service. Retrieved from https://sam.riss.kr/organResearch.do

[3] Borgatti, S. P., & Halgin, D. S. (2011). On network theory. Organization Science, 22(5), 1168-1181. doi:10.1287/orsc.1100.0641

[4] van Eck, N. J., & Waltman, L. (2014). Visualizing bibliometric networks. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), Measuring scholarly impact: Methods and practice (pp. 285-320). Springer. doi:10.1007/978-3-319-10377-8_13

[5] BERT: Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language

understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) (pp. 4171-4186). doi:10.18653/v1/N19-1423