A Study on Research Trend Analysis Using Social Network Analysis Methods

Yoonjin Lee

Department of Statistics

The Graduate School of Sungshin Women's University

A Study on Research Trend Analysis Using Social Network Analysis Methods

A Master's Thesis
Submitted to the
Graduate School of Sungshin Women's University

in partial fulfillment of the requirements
for the degree of
Master of Statistics

Yoonjin Lee

November, 2024

This is to certify that we have examined the
Master's Thesis of
Yoonjin Lee
Submitted to Department of Statistics

Approved as to style and content:

Thesis Advisor Hohyun Jung

Committee Chairman Seongkeon Lee

Committee Member Frederick Kin Hing Phoa

Committee Member Taehwa Choi

The Graduate School of Sungshin Women's University

Abstract

Research trend analysis identifies key topics and keywords that play a significant role in a specific field while efficiently processing information from large-scale data. Social network analysis visualizes the interactions between words in a network structure, enabling the understanding of the connectivity and structure of complex systems. Additionally, it allows the discovery of clusters in the network and identify groups that share similar characteristics. In this study, we propose a methodology for conducting research trend analysis using social network analysis methods. We construct a word co-occurrence network from large-scale text data and introduce a novel weighting method that incorporates both the frequency of word occurrences and the distances between words within a document. The proposed edge weight incorporates concordance, proximity, density, and anti-universality, complementing traditional co-occurrence-based methods. Using this network, we perform community detection to identify the themes represented by each cluster. To validate this methodology, we applied it to paper abstracts extracted from the Semantic Scholar database to demonstrate its effectiveness in identifying meaningful research topics.

Keywords : social network analysis, research trend analysis, word co-occurrence network, centrality, community detection

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Research trend analysis is a systematic method for identifying the development of research topics within a specific field or examining research trends across an entire discipline. Research trend analysis enables the observation of emerging topics, new research directions, or changes in trends by analyzing academic articles, conference papers, journals, patents, or various academic documents. This approach utilizes methods such as natural language processing and data mining for handling large-scale text data.

Research trend analysis provides insights across various fields. It is used in the academic field as it delivers knowledge about specific areas to researchers. For example, using citation network analysis, researchers can identify papers that are likely to become highly cited, allowing them to focus on cutting-edge studies in their field (Asatani et al., 2018). In the industrial field, it is applied to offer indicators of promising fields for investment. For instance, a bibliometric analysis of Industry 4.0 technologies has shown how such tools can highlight areas with significant innovation, guiding companies toward future investments (Nguyen et al., 2021). It is also utilized in policy-making, where it can help shape government policies or national research agendas. By analyzing trends in scientific research, policy-makers can adjust funding priorities and develop strategic initiatives in areas such as science and technology (Sivanandham et al., 2021; Lee and Song, 2020).

Social network analysis(SNA) is a powerful method used to study relationships and interactions among entities within a network. SNA is traditionally applied to human or organizational networks and can also be adapted to analyze relationships between words in a text to create a word co-occurrence network. In this network, words are represented as

nodes, and their co-occurrence within a given context such as a sentence, paragraph, or document forms the edges between them. Word co-occurrence networks are valuable for identifying key terms, uncovering thematic structures, and understanding how concepts are interconnected within a large amount of text. This approach falls within the domain of natural language processing (NLP), particularly useful in text mining. By utilizing NLP techniques such as tokenization, lemmatization, and part-of-speech tagging, researchers can preprocess large textual datasets and then apply network analysis techniques to study the frequency and proximity of word pairings. These NLP methods enhance the ability to extract meaningful patterns and connections from vast collections of text.

In this paper, we propose a methodology for analyzing trends in large-scale text data and apply it using paper abstracts extracted from Semantic Scholar. Semantic Scholar, a free AI-powered academic search engine, provides a means to access paper information, extract meanings, and identify connections among scientific literature. Utilizing abstracts from 2023, we constructed a word co-occurrence network where nodes represent frequently occurring words and edges represent co-occurrence relationships.

To enhance the representational power of the network, we introduced a novel edge-weighting method called Distance-IDF weight, which integrates the co-occurrence frequency of word pairs, the distance between words within a document, and the IDF (Inverse Document Frequency) value. This approach was demonstrated to be more effective in identifying key topics compared to conventional weighting methods.

The key contribution of this study is that the proposed association measure between two words, used as the edge weight, satisfies the following four desirable properties.

- **Concordance**: The more frequently two words co-occur in documents, the stronger their association.

Co-occurrence frequency is a simple yet powerful indicator of whether two words are used

in the same topic or context. If the association measure does not satisfy this property, word pairs that are contextually unrelated may be overestimated. Moreover, frequent co-occurrence provides confidence that the relationship between the two words is based on actual association rather than random chance or noise. For instance, the words “cell” and “cancer” often co-occur in biology or medical-related documents, making it reasonable to evaluate their association as strong.

- **Proximity**: The closer two words are located within a document, the stronger their association.

Words that are closer together are more likely to have been used in the same context. If this property is not considered, the association measure may fail to accurately capture contextual relationships. Additionally, proximity helps filter out word pairs that appear in the same document but are used in entirely different contexts. For example, while “artificial” and “intelligence” may appear in the same document, if they are far apart, they are more likely to belong to unrelated contexts. On the other hand, when the two words appear close together, as in “artificial intelligence”, it strongly suggests they are being used with related meanings.

- **Density**: The more frequently two words appear together within a single document, the stronger their association.

When two words are repeatedly mentioned in the same document, it indicates a strong contextual association rather than a random co-occurrence. Ignoring density could lead to missing such critical information. For instance, in the document “The doctor spoke with the patient. The patient described their symptoms to the doctor.”, the words “doctor” and “patient” are repeatedly mentioned together, reflecting a strong contextual relationship. In contrast, in the document “The doctor explained the procedure. The patient visited the clinic.”, the two words appear only once each, suggesting weaker contextual ties or mere coincidence.

- **Anti-universality**: The less universally frequent each word is, the stronger their association.

In other words, the fewer documents each word appears in, the stronger their association. If this property is not satisfied, universally frequent words that lack meaningful content may distort the association measure or lead to the underestimation of contextually significant word pairs. For example, if universally frequent words such as ”the” and ”is” are indiscriminately considered, relationships like ”artificial” and ”the” or ”intelligence” and ”is” could be overestimated. Conversely, ”artificial” and ”intelligence” are used more closely in specific contexts, so excluding the influence of universal words allows for a more accurate measurement of their true association.

Community detection was conducted using the Louvain algorithm, allowing us to cluster the network into distinct communities. Each community's content was analyzed to identify its thematic focus. By leveraging the Distance-IDF weight, we successfully highlighted more precise and representative keywords within each community, facilitating the identification of significant topics for the year. This methodology provides an efficient framework for trend analysis and can be extended to other domains requiring large-scale text analysis.

In Chapter 2, we introduce research related to word co-occurrence networks. Chapter 3 introduces the concepts applied in the methodology. In Chapter 4, we explain our proposed methodology, and Chapter 5 applies a case study to demonstrate our methodology. Finally, Chapter 6 contains a summary and conclusion of the research.

Chapter 2

Related work

A word co-occurrence network is applied in the fields of NLP and text analysis. It can be used in areas such as semantic analysis to find patterns of sentiment in text (Bermingham et al., 2009), bias detection to identify biases (Keidar et al., 2021), research field mapping to discover relationships or related fields in research (Tibaná-Herrera et al., 2018), and theme and topic identification to find clusters or key themes in a network (Radhakrishnan et al., 2017).

Word Co-occurrence Networks have been widely used to explore research topics and related fields (Radhakrishnan et al., 2017). In these networks, each word is represented as a node, and words that co-occur are connected by edges. The weight of the edges is determined by the frequency of the co-occurrences of word pairs, allowing researchers to identify patterns or the strength of connections between words.

Frequency-based word co-occurrence networks are widely used to analyze relationships between terms in textual data. When weighting the edges, the weight of an edge between two words is determined by the number of times they co-occurred in same document. This approach has been applied in various studies to uncover knowledge structures and research trends.

Lozano et al. (2019) conducted a network analysis of word co-occurrences in Data Envelopment Analysis (DEA) literature from 2008 to 2017. They standardized raw keywords to enhance consistency and constructed a weighted, undirected network where edge weights represented the frequency of keyword co-occurrences. This paper provided insights into the evolving nature of DEA research and identified emerging topics within the field.

Yuan et al. (2022) utilized a keyword co-occurrence network to review trends in intelligent manufacturing research. They analyzed keywords from 84,041 papers published between 2000 and 2020, constructing a network where nodes represented keywords and edges indicated co-occurrence frequencies. This method allowed them to systematically map the knowledge components and structures in intelligent manufacturing, revealing significant research trends and shifts over two decades.

Based on the extensive research that has been conducted on calculating network weights in keyword co-occurrence networks using frequency, we propose a novel method for calculating weights that considers both frequency-based distance and the number of occurrences of words within the document.

Chapter 3

Preliminaries

3.1 Concept of Network

3.1.1 Network

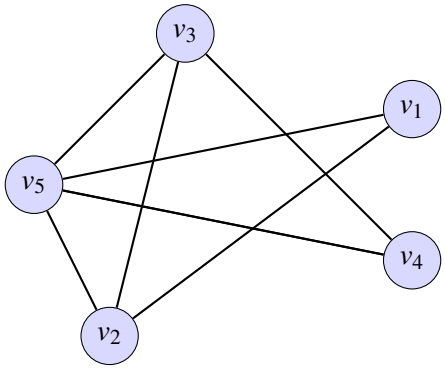

Figure 1: Network G with five nodes

Figure 1 represents a network G. In a network or graph G, the structure is represented as $G = (V, E)$, where V is a set of nodes and E is a set of edges between these nodes. The set of nodes is denoted as $V = \{v_1, v_2, v_3, v_4, v_5\}$ representing individual elements or entities that construct the network. The edges is represented as $E = \{(v_1, v_2), \cdots, (v_4, v_5)\}$. The edges are pairs of nodes that connect specific nodes to each other and shows relationships or interactions between them within the network.

3.1.2 Word Co-occurrence Network

A Word Co-occurrence Network is a graph-based representation where nodes correspond to words, and edges connect words that co-occur within the same document or sentence. Word co-occurrence potentially reflects relationships between words, aiding in visualizing and understanding the structure of topics within a document. By analyzing the associations between words in word co-occurrence network, it can uncover trends, themes, and knowledge structures in a specific field or across multiple disciplines.

The foundation of a word co-occurrence network lies in the idea that words or phrases frequently appearing together within the same context signify a semantic or topical connection. This theoretical framework is rooted in co-occurrence term, which suggests that terms that co-occur within a given context—such as a document, paragraph, or sentence—are likely related in meaning or theme. These connections are then represented as networks to illustrate these implicit relationships, where frequent co-occurrences highlight stronger or more central themes in the data.

In word co-occurrence networks, the relationships between words are non-directional, meaning they only capture whether two terms co-occur but not the order of their appearance. This framework is often employed in bibliometric studies, topic modeling, and text mining to analyze the structural relationships of terms, helping researchers identify key concepts, thematic clusters, and even emerging trends within a dataset.

3.2 Centrality measures

Centrality measures are calculated to identify important or central words within a network. Centrality measures include degree centrality, betweenness centrality, and closeness centrality.

3.2.1 Degree centrality

The degree centrality (Opsahl et al., 2010) of node i in weighted graph can be formalized as:

$$C_D^w(i) = \sum_j^N w_{ij} \tag{3.1}$$

j refers to all other nodes, N is the total number of nodes and w is the weighted adjacency matrix.

Degree centrality is a measure used to evaluate the importance of a node within a network, meaning that the more a particular node is connected to many other nodes, the higher its centrality.

3.2.2 Betweenness centrality

To calculate betweenness centrality and closeness centrality, the shortest path in a weighted graph has to be identified. The length of the shortest path between two nodes in weighted graph can be calculated by Dijkstra's algorithm (Dijkstra, 2022) as:

$$d^w(i,j) = min\left(\frac{1}{w_{ih}} + \cdots + \frac{1}{w_{hj}}\right) \tag{3.2}$$

The shortest path between two nodes i and j is the path among the various possible routes from node i to node j that has the minimum sum of the reciprocals of the weights.

The betweenness centrality of node i in weighted graph (Opsahl et al., 2010) can be formalized as:

$$C_B^w(i) = \frac{g_{jk}^w(i)}{g_{jk}^w} \tag{3.3}$$

g_{jk}^{w} is the number of shortest paths between two nodes j, k and $g_{jk}^{w}(i)$ is the number of those paths that go through node i.

Betweenness centrality identifies nodes that act as key connectors or intermediaries in the network. Nodes with high betweenness centrality lie on a large number of shortest paths between other nodes, meaning they facilitate communication and flow across different parts of the network. Such nodes are crucial for network cohesion, as they bridge otherwise disconnected regions or groups within the network.

3.2.3 Closeness centrality

The closeness centrality of node i in weighted graph (Opsahl et al., 2010) can be formalized as:

$$C_C^w(i) = \left[\sum_j^N d^w(i,j) \right]^{-1} \tag{3.4}$$

Closeness centrality reflects how quickly a node can access all other nodes in the network. Nodes with high closeness centrality have shorter average distances to other nodes, indicating that they are more central in terms of accessibility. These nodes are strategically positioned to efficiently reach others, making them effective in spreading information or influence throughout the network.

3.3 Community Detection

The Louvain algorithm (Blondel et al., 2008) is widely used due to its ability to detect meaningful community structures efficiently in large and complex networks, making it useful for applications networks. This algorithm identifies communities by maximizing modularity, a measure of the density of edges within communities compared to edges be-

tween communities.

Modularity (Newman, 2006) is a measure used in community detection to evaluate how well a network is divided into communities. Modularity is calculated as:

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j) \tag{3.5}$$

In formula 3.5, $2m$ is sum of all edge weights in the weighted network, A is the adjacency matrix of graph, k_i, k_j is the weighted degree of node i, j. $\delta(c_i, c_j)$ is 1 if node i and j are in the same community, otherwise 0.

In the Louvain algorithm, each node initially forms its own community. For each node, the algorithm evaluates the change in modularity when the node is moved to the community of each of its neighbors. The node is then assigned to the neighboring community that provides the highest modularity gain, provided the gain is positive. This step iterates through the nodes until no further modularity improvements can be achieved by moving nodes. This marks the end of the first phase of the Louvain algorithm.

After this process, each community is condensed into a single node, reducing the size of the network. Edges between nodes within the same community become self-loops, while edges between different communities become inter-community edges. This process creates a smaller network structure based on the detected communities from the previous phase. The first and second phases repeat until no further increases in modularity are achieved.

Chapter 4

Proposed method

4.1 Word Co-occurrence Network

4.1.1 Construction

Figure 2 is a flowchart that illustrates the entire process of the methodology for conducting research trend analysis.

The methodology for constructing and analyzing a word co-occurrence network involves several steps, including data collection, preprocessing, network construction, and network analysis.

Data collection involves gathering text data that will serve as the input for the word co-occurrence network. Common sources include research paper abstracts, full-text articles, social media posts, news articles, or any extensive collection of text where words are present. To extract proper words and enhance the network's associations and clarity, a preprocessing step is applied to the raw text. In each document, the words are converted to lowercase and then processed using a NLP model to obtain tokenized and POS-tagged words. Stopwords are removed. To handle two-word compounds, adjective-noun and noun-noun combinations are extracted among the POS-tagged words. To ensure meaningful compounds, only those that appear above a certain frequency threshold within the document are considered. Finally, single nouns are also extracted. Through the preprocessing step, meaningful compounds and nouns are extracted, which become the words.

After preprocessing, word pairs are generated from the preprocessed words in each document. For each word pair, the Distance-IDF weight is calculated. The method for cal-

culating the Distance-IDF weight is introduced in 4.1.2.

In the word co-occurrence network, nodes represent unique words extracted from documents, serving as concepts or terms that represent the dataset. Edges represent co-occurrence relationships between nodes. If two words appear together within a single document, they are said to co-occur, and an edge is formed between these two words.

Using the word co-occurrence network constructed through this process, community detection is performed using the Louvain algorithm. After community detection, the degree, betweenness, and closeness centrality of each cluster are calculated to identify the topics of the clusters.

4.1.2 Distance-IDF Weight

In traditional word co-occurrence networks, the edge weighting typically represents the frequency of co-occurrence between two words within the same context or document. While this approach captures the basic relational strength between terms, it does not fully account for the complex interactions across multiple documents. To address this limitation, we propose a novel method for calculating edge weights that incorporates not only the co-occurrence frequency but also the distance between words in documents. This approach aims to provide a more nuanced representation of term associations, reflecting their influence and connectivity within a broader corpus.

- Let $v_i, i = 1, 2, \ldots, N$, where N is the number of words in the vocabulary.
- Let $d, d = 1, 2, \ldots, D$, where D is the number of documents in the corpus.
- Using the document frequency $df(v_i)$, the IDF is defined as:

$$idf(v_i) = \frac{1}{df(v_i)}$$

- If word v_i appears $n_{d,i}$ times in document d, the words in document d can be denoted as $v_{i,1}, \ldots, v_{i,n_{d,i}}$.

- Distance-Idf formula is given by:

$$w_{DI}(v_i, v_j) = idf(v_i) \cdot idf(v_j) \cdot \sum_{d=1}^{D} \left(\sum_{k=1}^{n_{d,i}} \frac{1}{\min_{l=1,\ldots,n_{d,j}} dist(v_{i,k}, v_{j,l})} + \sum_{l=1}^{n_{d,j}} \frac{1}{\min_{k=1,\ldots,n_{d,i}} dist(v_{i,k}, v_{j,l})} \right) \tag{4.1}$$

In this formula, let v_i represent the words in the vocabulary, where $i = 1, 2, \ldots, N$, and N is the total number of words. Let d denote the documents in the corpus, where $d = 1, 2, \ldots, D$, and D represents the number of documents. The document frequency $df(v_i)$ refers to the number of documents in which the specific word v_i appears. The inverse document frequency (IDF) for a word v_i is defined using the document frequency, calculated as the reciprocal of its document frequency.

Formula 4.1 represents the weight between two words, v_i and v_j, defined as the 'distance-idf weight'. If a word v_i appears $n_{d,i}$ times in a document d, the occurrences of the word within the document can be expressed as $v_{i,1}, \ldots, v_{i,n_{d,i}}$. The weight is defined as the product of their inverse document frequencies, $idf(v_i)$ and $idf(v_j)$, multiplied by the sum over all documents from $d = 1$ to D. The first component in the sum is the sum over each occurrence k of v_i in document d, where each term is the reciprocal of the minimum distance to any occurrence l of v_j in the same document. The second component is the sum over each occurrence l of v_j in document d, where each term is the reciprocal of the minimum distance to any occurrence k of v_i in the document. This overall formulation quantifies the weight $w_{DI}(v_i, v_j)$ by considering how close the two words appear together across all documents, weighted by their respective IDF values. This formula reflects the IDF values of both words and aggregates the distances between their occurrences across all documents.

- Consider three documents with the following words:
  - Document 1: [ai, deep learning, parameter, model, accuracy, regression, ai]
  - Document 2: [dataset, accuracy, loss, ai, optimization]
  - Document 3: [training, ai, hyperparameter, accuracy, evaluation]
- To calculate the weight $w_{DI}(\text{ai}, \text{accuracy})$ between the words “ai” and “accuracy” across these documents, we’ll first determine the IDF values:

$$idf(\text{ai}) = \frac{1}{df(\text{ai})} = \frac{1}{3} \quad \text{and} \quad idf(\text{accuracy}) = \frac{1}{3}$$

  since both words appear in three documents.
- For each document, we find the minimum distances between occurrences of ”ai” and ”accuracy.”
- First, the distances between all occurrences of “ai” and “accuracy” are calculated with respect to the word “ai”.

  The process for calculating $dist(ai, accuracy)$ is as follows:
  - In Document 1, the first occurrence of “ai” is at position 1 and “accuracy” at position 5, so $dist(\text{ai}, \text{accuracy}) = |1-5| = 4$. The second occurrence of “ai” is at position 7 and “accuracy” at position 5, so $dist(\text{ai}, \text{accuracy}) = |7-5| = 2$.
  - In Document 2, “ai” is at position 4 and “accuracy” at position 2, so $dist(\text{ai}, \text{accuracy}) = |4-2| = 2$.
  - In Document 3, “ai” is at position 2 and “accuracy” at position 4, so $dist(\text{ai}, \text{accuracy}) = |2-4| = 2$.
- Second, the distances between all occurrences of “ai” and “accuracy” are calculated

with respect to the word “accuracy”.

The process for calculating $dist(accuracy, ai)$ is as follows:

- In Document 1, “accuracy” is at position 5 and “ai” is at position 1 and 7. Considering the minimum distance between the two words, $dist(\text{accuracy}, \text{ai}) = |5-7| = 2$
- In Document 2, “accuracy” at position 2 and “ai” is at position 4, so $dist(\text{accuracy}, \text{ai}) = |2-4| = 2$.
- In Document 3, “accuracy” at position 4 and “ai” is at position 2, so $dist(\text{accuracy}, \text{ai}) = |4-2| = 2$.

- Calculating these distances into Formula 4.1, we get:

$$\begin{aligned} w_{DI}(\text{ai}, \text{accuracy}) &= idf(\text{ai}) \cdot idf(\text{accuracy}) \cdot \\ &\sum_{d=1}^{3} \left(\frac{1}{\min \text{dist}(\text{ai}, \text{accuracy})} + \frac{1}{\min \text{dist}(\text{accuracy}, \text{ai})} \right) \\ &= \left(\frac{1}{3}\right) \cdot \left(\frac{1}{3}\right) \cdot \left(\frac{1}{4} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2}\right) \\ &= \frac{1}{9} \cdot \left(\frac{1}{4} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2}\right) \approx 0.361. \end{aligned} \tag{4.2}$$

Compared to the traditional method of calculating edge weights by simply counting the number of co-occurring documents, our proposed weight calculation method offers several advantages for better capturing the importance and relevance of edges:

- **Concordance:** The more documents in which two words co-occur, the higher the weight value.
- **Proximity:** The closer the two words are located within a document, the higher the weight becomes.

- **Density:** Reflects how closely the words are grouped together.
- **Anti-universality:** The less universally used a word is, the higher its weight. In other words, if a word appears in fewer documents, its weight increases.

Concordance ensures that the weight increases as the number of documents where two words co-occur rises. This highlights frequently co-occurring word pairs across multiple documents, as their relationship is repeatedly validated in the corpus. Proximity emphasizes the importance of words located closer together within a document, assigning higher weights to such word pairs to capture their semantic or contextual closeness effectively. Density considers how tightly grouped two words are within a given context, reflecting not only their proximity but also whether they are surrounded by similar terms, thereby highlighting clusters of related words. Finally, Anti-universality assigns higher weights to words that appear in fewer documents, emphasizing their uniqueness and specificity. This discourages overly common words from dominating the network and allows rare but significant terms to play a central role in the analysis.

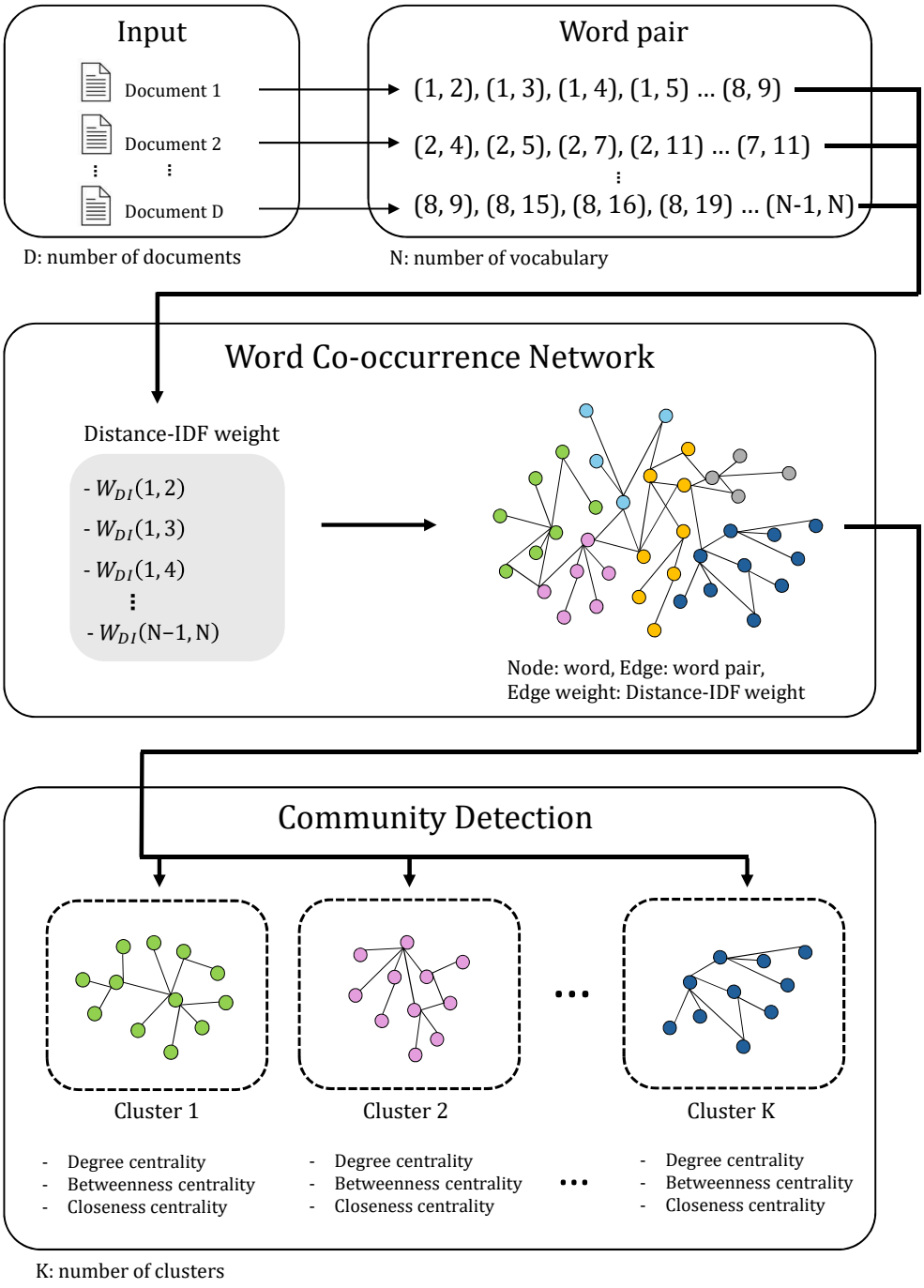

Figure 2: The flowchart of the proposed method

Chapter 5

Experiments

We applied the proposed methodology to research paper data extracted from the Semantic Scholar database.

5.1 Dataset

Table 1 provides the number of the papers and words from Semantic Scholar Database. A total of 1,315,518 paper abstracts were extracted from the year 2023. From these papers, 38,169 words were identified by selecting only words that appeared more than 100 times across the entire dataset and were present in less than 10% of all documents. Table 2 shows some of the words selected from the papers after preprocessing.

Year	# of paper	# of words
2023	1,315,518	38,169

Table 1: Number of papers and words of year 2023

Word	**Word**	**Word**	**Word**
algorithm	assessment	back pain	body
configuration	context	estimation	deep learning
...	...	...	...
processing	covariance	quantum channel	training

Table 2: Words after preprocessing paper abstract

5.2 Experimental procedure

Based on Figure 2, paper abstracts extracted from the Semantic Scholar data were considered as individual documents and used as input values. After extracting the 2023 paper abstracts, the preprocessing steps are performed according to 4.1.1.

Word pairs are generated according to Table 3. Table 3 shows the process of tokenizing a single abstract and forming word pairs. The first abstract in Table 3 is “Mitochondria produce the energy necessary for the cell’s survival.” After preprocessing, the words mitochondria, energy, cell, and survival are tokenized. The number of word pairs that can be formed using these four words is 6, calculated as ${}_4C_2$. Therefore, the word pairs are (mitochondria, energy), (mitochondria, cell), (mitochondria, survival), (energy, cell), (energy, survival), and (cell, survival). The second abstract is ”The cell requires a constant supply of energy for its survival.” After preprocessing, the words cell, supply, energy, and survival are tokenized. The word pairs (cell, supply), (cell, energy), (cell, survival), (supply, energy), (supply, survival), and (energy, survival) are generated, making a total of 6 pairs.

Through this process, word pairs are generated from abstracts for each year. The words from the generated word pairs become the nodes in the word co-occurrence network, and the pairs themselves represent the edges connecting these nodes. The edges are weighted by a value defined as our proposed weight. The edges are assigned weights based on our proposed Distance-IDF weight.

After this process, a single word co-occurrence network is generated. The degree, betweenness, and closeness centralities of the generated word co-occurrence network are calculated. To identify the clusters within the word co-occurrence network, community detection is performed using the Louvain algorithm. The degree, betweenness, and closeness centralities of the detected clusters are calculated to identify the topics addressed within each cluster.

To verify that the word co-occurrence network generated using our proposed Distance-IDF weight allows for a better understanding of detailed topics, we compared it with a network where weights were based on document frequencies and calculated degree, betweenness, and closeness centralities. Table 4 shows the document frequencies based on table 3.

Paper	Process	Details
1	Abstract	Mitochondria produce the energy necessary for the cell's survival.
	Tokenization	mitochondria, energy, cell, survival
	Word Pair	(mitochondria, energy), (mitochondria, cell), (mitochondria, survival), (energy, cell), (energy, survival), (cell, survival)
2	Abstract	The cell requires a constant supply of energy for its survival.
	Tokenization	cell, supply, energy, survival
	Word Pair	(cell, supply), (cell, energy), (cell, survival), (supply, energy), (supply, survival), (energy, survival)

Table 3: Tokenization and Word Pairs for Paper Abstracts

Word Pair	Document Frequency
(energy, survival)	2
(cell, survival)	2
(mitochondria, energy)	1
(mitochondria, cell)	1
(mitochondria, survival)	1
(energy, cell)	1
(cell, supply)	1
(cell, energy)	1
(supply, energy)	1
(supply, survival)	1

Table 4: Document Frequency of Word Pairs

5.3 Results

5.3.1 Word Co-occurrence network

Year	# of nodes	# of edges
2023	16,568	674,108

Table 5: Word Co-occurrence Network statistics for 2023

No	word 1	word 2	compound words
1	artificial	intelligence	artificial intelligence
2	risk	management	risk management
3	linear	regression	linear regression
4	experimental	design	experimental design
5	correlation	coefficient	correlation coefficient
6	clinical	trial	clinical trial
7	deep	learning	deep learning
8	reinforcement	learning	reinforcement learning
9	hazard	ratio	hazard ratio
10	social	network	social network
11	random	sampling	random sampling
12	time	series	time series
13	decision	tree	decision tree
14	gene	expression	gene expression
15	risk	assessment	risk assessment
16	cell	line	cell line
17	support	vector	support vector
18	random	forest	random forest
19	computational	cost	computational cost
20	feature	extraction	feature extraction

Table 6: Compound Words Table

Table 5 shows the number of nodes and edges in the 2023 word co-occurrence network. The network consists of 16,568 nodes and 674,108 edges. Table 6 presents a selection of compound words extracted from the preprocessed data following the steps in 4.1.1. Examples of compound words include artificial intelligence, risk management, linear regression,

Rank	Distance-IDF Weights			Frequency-based Weights		
	Degree	**Betweenness**	**Closeness**	**Degree**	**Betweenness**	**Closeness**
1	novartis	treatment	novartis	treatment	treatment	order
2	pfizer	order	pfizer	order	disease	treatment
3	galapagos	questionnaire	galapagos	quality	network	material
4	gilead	disease	gilead	disease	cell	quality
5	universal soil	material	bms	material	material	disease
6	loss equation	parameter	celgene	participant	parameter	participant
7	systolic excursion	network	sandoz	network	participant	design
8	copyright holder	education	biontech	design	image	network
9	sanofi	annual meeting	chugai	parameter	quality	form
10	roche	temperature	abbvie	form	order	questionnaire
11	listserv	quality	battery management	activity	design	activity
12	kuhn	confidence interval	gsk	effectiveness	temperature	influence
13	type communication	neural network	roche	influence	gene	cell
14	annular plane	cell	celltrion	cell	activity	way
15	secretory phenotype	participant	sanofi	management	expression	neural network

Table 7: Degree, Betweenness, and Closeness Centrality Comparison for Word Co-occurrence Network Using Distance-IDF Weights and Frequency-based Weights

experimental design, correlation coefficient, clinical trial, deep learning, and reinforcement learning.

Table table:whole wcn shows the top 15 nodes with the highest values in each of the three centrality measures(degree, betweenness, closeness) for word co-occurrence network using Distance-IDF weight. For instance, in terms of degree centrality, the top 15 nodes are novartis, pfizer, galapagos, gilead, universal soil, loss equation, systolic excursion, copyright holder, sanofi, roche, listerv, kuhn, type communication, annualr plane, secretory phenotype. These words are commonly used by major pharmaceutical and biotech companies. Words derived using the Distance-IDF weight are found to be central to specific fields such as pharmaceuticals, life sciences, and intellectual property.

Below are the explanations for the top 15 words of degree centrality:

1. **novartis**: A multinational pharmaceutical company headquartered in Switzerland, known for developing innovative medicines.
2. **pfizer**: An American pharmaceutical giant famous for its vaccines (e.g., COVID-19 vaccine) and other healthcare products.
3. **galapagos**: A European biotechnology company focused on the discovery and devel-

opment of medicines.

4. **gilead**: A biopharmaceutical company known for its research on antiviral drugs, including treatments for HIV and hepatitis.
5. **universal soil**: Likely refers to a concept or tool used in environmental science or agriculture to study or standardize soil characteristics.
6. **loss equation**: A mathematical or scientific model used to calculate losses in various contexts, such as energy, information, or resource efficiency.
7. **systolic excursion**: A medical term describing the movement of the heart's structures during the systolic phase of the cardiac cycle, often used in echocardiography.
8. **copyright holder**: Refers to the individual or entity that holds the legal rights to a copyrighted work.
9. **sanofi**: A French multinational pharmaceutical company, well-known for its vaccines and treatments for rare diseases.
10. **roche**: A Swiss healthcare company focused on pharmaceuticals and diagnostics, renowned for cancer treatments and diagnostic tools.
11. **listserv**: A software application used to manage electronic mailing lists, commonly used in professional communication.
12. **kuhn**: Likely refers to the Kuhn-Tucker conditions, a fundamental concept in mathematical optimization and statistics, which generalize the method of Lagrange multipliers for constrained optimization problems.
13. **type communication**: Refers to the methods or modes of communication, such as verbal, written, digital, etc.
14. **annular plane**: A medical term related to the anatomy of the heart, specifically referring to the annulus (ring-like structure) seen in imaging.
15. **secretory phenotype**: A biological term describing cells' behavior when they produce and secrete substances, often studied in immunology and cancer research.

Degree Centrality measures how well-connected a node is to other nodes in the network. In table 7, nodes such as treatment, order, quality is high ranked based on degree centrality. These terms are frequent and widely used but they do not provide detailed insights into the domain of study. Degree Centrality based on Distance-IDF weight focuses on specific entities like pharmaceutical companies that are highly relevant to the research domain, whereas frequency weight highlight overly broad terms.

Betweenness Centrality identifies nodes that act as connection between different clusters in the network. Terms such as 'treatment', 'disease', and 'network' in 7 rank high, indicating their role as connectors across various research subfields. Additionally, words like 'parameter' and 'education' show that the method identifies both technical and interdisciplinary bridging topics. In table 7 'treatment' and 'disease' remain significant, other terms like 'cell' and 'activity' dominate. These terms, while important, lack the same level of specificity and contextual relevance.

Closeness Centrality measures how quickly a node can reach all other nodes in the network. In table 7 nodes like 'novartis', 'pfizer', 'galapagos', and 'gilead' are central, highlighting the focus on pharmaceutical fields. In table 7 nodes such as 'order', 'treatment', and 'material' rank highly. While relevant, these terms are too general to provide meaningful insights into the specific focus of the research.

The comparison across all three centrality measures demonstrates that using our proposed weight to interpret network is more effective in identifying domain-specific and meaningful terms. It prioritizes entities and concepts relevant to the research context, such as pharmaceutical companies ('novartis', 'pfizer') and interdisciplinary terms ('network', 'education'). In contrast, using frequency based weight tend to highlight general-purpose terms ('treatment', 'order') that, while frequent, offer limited insights into the specificity of the domain. This highlights the advantage of using our proposed 'distance-IDF' weight for more precise topic identification and analysis in the network.

5.3.2 Community detection

A total of 37 clusters with more than 100 nodes were identified and we visualized 4 clusters.

5.3.2.1 The First Cluster

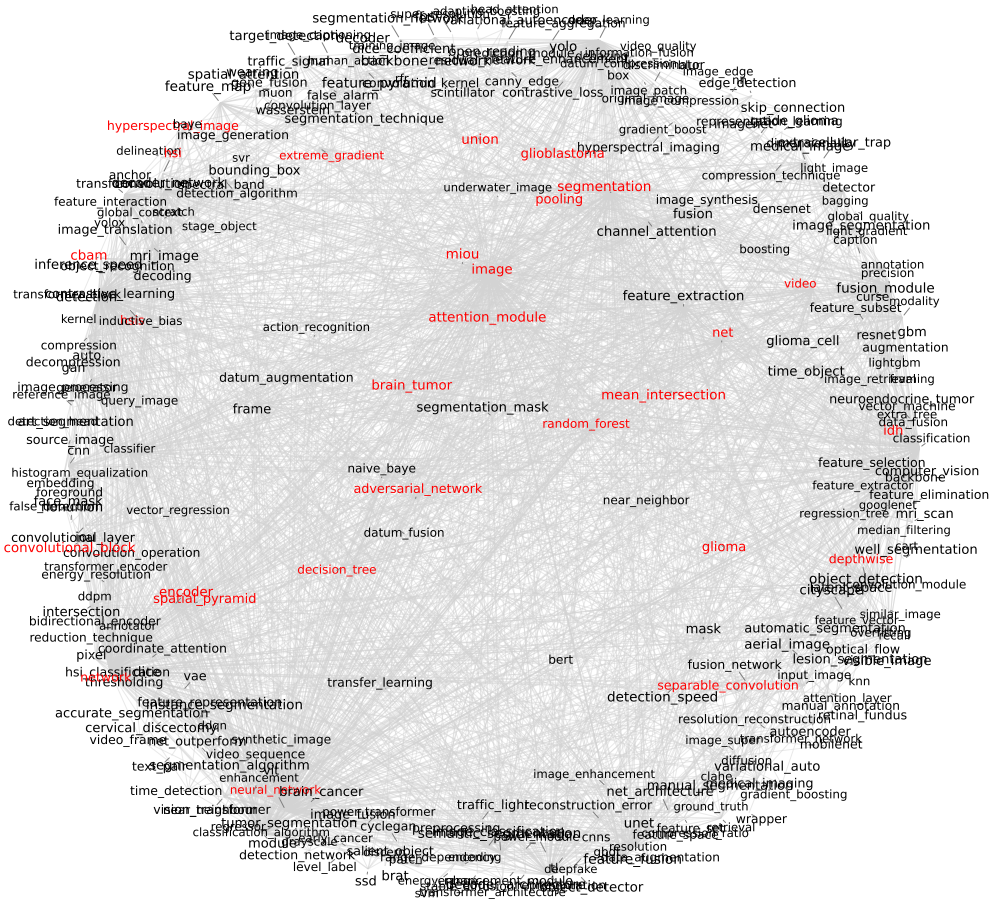

Figure 3: The first cluster by the community detection

Rank	Distance-IDF Weights			Frequency-based Weights		
	Degree	**Betweenness**	**Closeness**	**Degree**	**Betweenness**	**Closeness**
1	synthetic minority	logistic regression	feature pyramid	image	network	image
2	oversampling technique	network	feature fusion	network	image	network
3	separable convolution	image	backbone network	neural network	neural network	neural network
4	depthwise	neural network	object detection	detection	detection	detection
5	hsi	detection	feature map	deep learning	classification	deep learning
6	hyperspectral image	resolution	semantic segmentation	classification	video sequence	classification
7	image	loss	segmentation	algorithm	algorithm	algorithm
8	glioma	algorithm	miou	machine learning	deep learning	machine learning
9	convolutional block	machine learning	fusion module	module	video	module
10	segmentation	diffusion	instance segmentation	machine	segmentation	machine
11	cbam	machine	detection speed	loss	loss	loss
12	attention module	deep learning	mean intersection	precision	machine	segmentation
13	encoder	classification	union	feature extraction	machine learning	precision
14	mean intersection	modality	unet	segmentation	module	feature extraction
15	idh	representation	attention module	representation	original image	representation

Table 8: Degree, Betweenness, and Closeness Centrality for Cluster 1: Comparison Between Distance-IDF Weights and Frequency-based Weights

Below are the explanations for the top 15 words of degree centrality for Distance-IDF Weights:

1. **synthetic minority**: Refers to methods for balancing imbalanced datasets, such as the Synthetic Minority Oversampling Technique (SMOTE), often used in machine learning to improve model performance.
2. **oversampling technique**: A data augmentation strategy to increase the representation of minority classes in imbalanced datasets, commonly applied in classification problems.
3. **separable convolution**: A type of convolution operation in deep learning that reduces computational cost by separating spatial and depth-wise operations, often used in lightweight neural networks.
4. **depthwise**: Refers to depthwise separable convolutions, an efficient form of convolution that applies a single filter per input channel, frequently utilized in mobile and efficient neural networks.
5. **hsi**: Stands for Hyperspectral Imaging, a technique that captures spectral information across a wide range of wavelengths, commonly used in remote sensing and medical

imaging.

6. **hyperspectral image**: An image containing spectral data for each pixel across multiple wavelengths, widely applied in fields such as agriculture, environmental monitoring, and medical diagnostics.
7. **image**: Refers to visual data or digital representations used in computer vision and imaging systems, foundational in tasks like object detection, segmentation, and classification.
8. **glioma**: A type of tumor occurring in the brain or spinal cord, commonly studied in medical imaging and diagnostic applications.
9. **convolutional block**: A key component of convolutional neural networks (CNNs), typically including convolutional layers, activation functions, and normalization layers, designed to extract features from images.
10. **segmentation**: Refers to the process of partitioning an image into meaningful regions or segments, commonly used in medical imaging, computer vision, and remote sensing.
11. **cbam**: Convolutional Block Attention Module, a mechanism that enhances feature learning by applying spatial and channel-wise attention, often integrated into CNN architectures.
12. **attention module**: A deep learning mechanism that dynamically adjusts the weight of features or data, often improving the performance of tasks like image recognition, translation, and segmentation.
13. **encoder**: A component of deep learning models that maps input data into a feature representation, commonly used in architectures like autoencoders and transformers.
14. **mean intersection**: Likely refers to the Mean Intersection Over Union (mIoU), a metric used to evaluate segmentation accuracy in computer vision tasks.
15. **idh**: Stands for Isocitrate Dehydrogenase, a metabolic enzyme whose mutations are studied in cancer biology, particularly gliomas.

Figure 3 primarily focuses on advanced techniques in computer vision and deep learning models, as well as specific application areas such as medical imaging and diagnostic technologies.

In table 8, words like 'synthetic minority', 'oversampling technique', and 'feature pyramid' show advanced techniques specific to machine learning and image analysis. Word 'glioma' and 'hyperspectral image' suggest a deeper focus on real-world applications, such as medical imaging and remote sensing. In contrast, in table 8, words like 'image', 'network', and 'neural network' frequently appear, reflecting their importance but lacking contextual specificity.

5.3.2.2 The Ninth Cluster

Rank	Distance-IDF Weights			Frequency-based Weights		
	Degree	**Betweenness**	**Closeness**	**Degree**	**Betweenness**	**Closeness**
1	cluster head	sensor	cluster head	node	sensor	node
2	graph	graph	chs	sensor	graph	cluster
3	packet delivery	cluster	ch	graph	node	sensor
4	wsn	node	network lifetime	cluster	cluster	clustering
5	knowledge graph	delay	wsn	protocol	data transmission	graph
6	wsns	protocol	residual energy	wireless sensor	delay	protocol
7	wireless sensor	spatio	node	clustering	protocol	wireless sensor
8	node	knowledge graph	wsns	delay	clustering	gnn
9	network lifetime	wireless sensor	sensor node	throughput	network lifetime	delay
10	graph attention	arrival	wireless sensor	routing	spatio	throughput
11	sensor node	gaussian mixture	metastatic lymph	wsn	ch	neighbor
12	gat	cluster head	routing	wsns	routing	wsn
13	kgs	gold nanoparticle	cervical lymph	cluster head	node embedding	vertex
14	chs	chart	node representation	sensor node	knowledge graph	routing
15	end delay	packet delivery	sensor network	datum transmission	partition	wsns

Table 9: Degree, Betweenness, and Closeness Centrality for Cluster 9: Comparison Between Distance-IDF Weights and Frequency-based Weights

Below are the explanations for the top 15 words in degree centrality for Cluster 9 using Distance-IDF weights:

1. **cluster head**: Refers to a key node in a wireless sensor network (WSN) responsible for managing and coordinating communication within its cluster to optimize energy

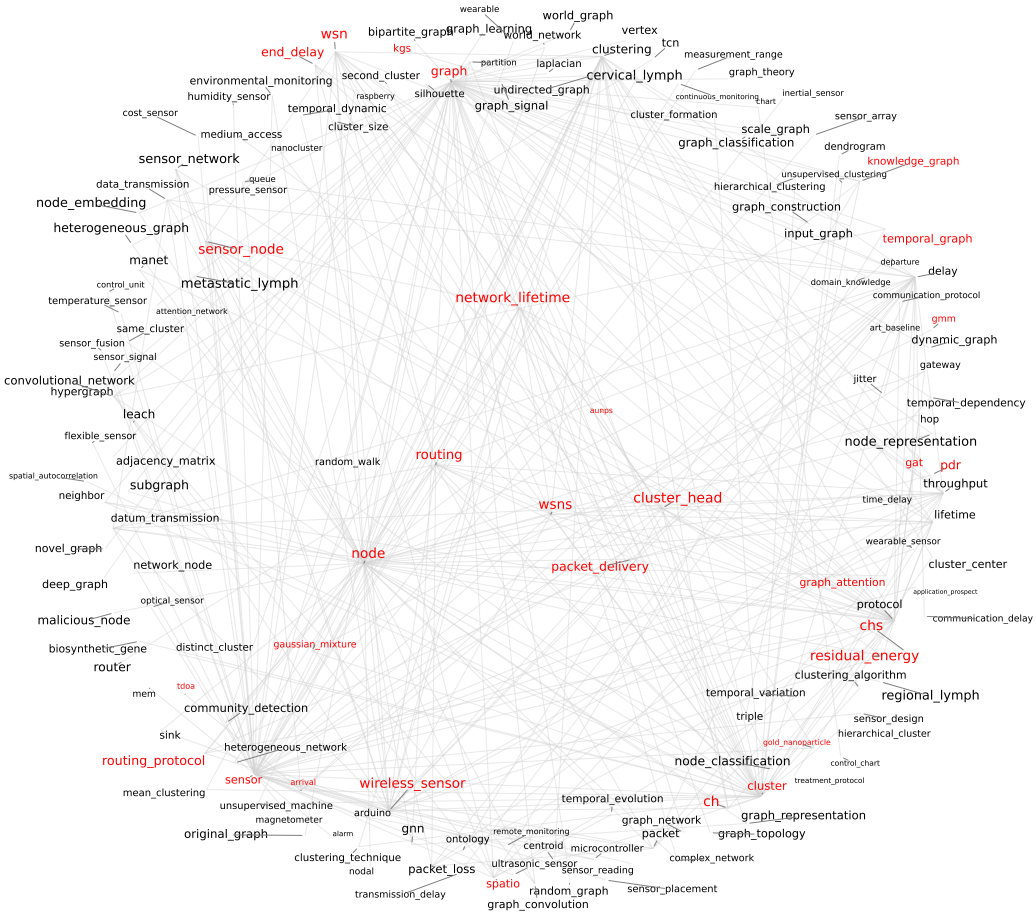

Figure 4: The ninth cluster by the community detection

consumption and data transmission.

2. **graph**: Represents graph-based models used for network topology representation and analysis in wireless sensor networks or similar systems.
3. **packet delivery**: Indicates the successful delivery of data packets within a network, a key metric in assessing network performance.
4. **wsn**: Stands for Wireless Sensor Network, a distributed network of sensors used for

monitoring and collecting data in various applications.

5. **knowledge graph**: A structured representation of information in graph form, often used to model relationships between data points for intelligent decision-making.
6. **wsns**: Plural of WSN, emphasizing the collective nature of multiple wireless sensor networks in a study.
7. **wireless sensor**: Refers to individual sensors within a WSN, typically used for collecting and transmitting data.
8. **node**: A fundamental component of a network, representing devices or points that connect to form the WSN.
9. **network lifetime**: The duration for which a WSN operates efficiently before the energy of its nodes is depleted.
10. **graph attention**: A mechanism used in graph neural networks (GNNs) to assign different importance to nodes or edges for more effective data processing.
11. **sensor node**: A node in a WSN equipped with sensors for monitoring specific parameters and transmitting data.
12. **gat**: Stands for Graph Attention Network, a type of neural network designed to operate on graph-structured data.
13. **kgs**: Likely refers to Knowledge Graph Systems, which integrate data from multiple sources in a structured form.
14. **chs**: Short for Cluster Heads, critical elements in WSNs that manage communication within a cluster.
15. **end delay**: The time taken for data to travel from the source node to the destination node in a network.

Figure 4 primarily focuses on energy-efficient communication, network optimization, and advanced computational techniques in wireless sensor networks (WSNs).

In table 9, words like ‘cluster head’, ‘graph attention’, and ‘network lifetime’ represent advanced methodologies specific to managing energy consumption and optimizing WSN performance. Terms such as ‘knowledge graph’ and ‘gat’ (Graph Attention Network) suggest the use of machine learning and graph-based approaches for enhancing network functionality.

In contrast, in table 9, words like ‘node’, ‘sensor’, and ‘graph’ frequently appear, emphasizing fundamental components of WSNs but lacking the depth and specificity to represent cutting-edge techniques.

5.3.2.3 The Tenth Cluster

Rank	Distance-IDF Weights			Frequency-based Weights		
	Degree	**Betweenness**	**Closeness**	**Degree**	**Betweenness**	**Closeness**
1	backdoor attack	security	attack	security	security	security
2	attack	aggregation	authentication	attack	attack	scheme
3	backdoor	robustness	security	internet	internet	thing
4	intrusion detection	thing	password	scheme	scheme	privacy
5	denial	internet	authentication scheme	thing	vulnerability	attack
6	medical thing	scheme	encryption	privacy	privacy	internet
7	encryption	attack	intrusion detection	edge	scheduling	iot
8	iomt	perturbation	decryption	cloud	client	attacker
9	watermark	edge	user authentication	computing	robustness	cloud
10	watermarking	privacy	membership inference	computation	defense	edge
11	network intrusion	scheduling	idss	server	intrusion detection	cyber
12	security	vulnerability	image encryption	latency	encryption	robustness
13	ddos	client	iot network	iot	server	computation
14	nid	computing	authorization	robustness	cloud	authentication
15	service attack	computation	ddos	vulnerability	computing	vulnerability

Table 10: Degree, Betweenness, and Closeness Centrality for Cluster 10: Comparison Between Distance-IDF Weights and Frequency-based Weights

Below are the explanations for the top-ranking words in degree centrality for Cluster 10 using Distance-IDF weights:

1. **backdoor attack**: Refers to a type of cyberattack where attackers bypass normal authentication to gain unauthorized access to systems, often exploiting vulnerabilities to insert malicious code.

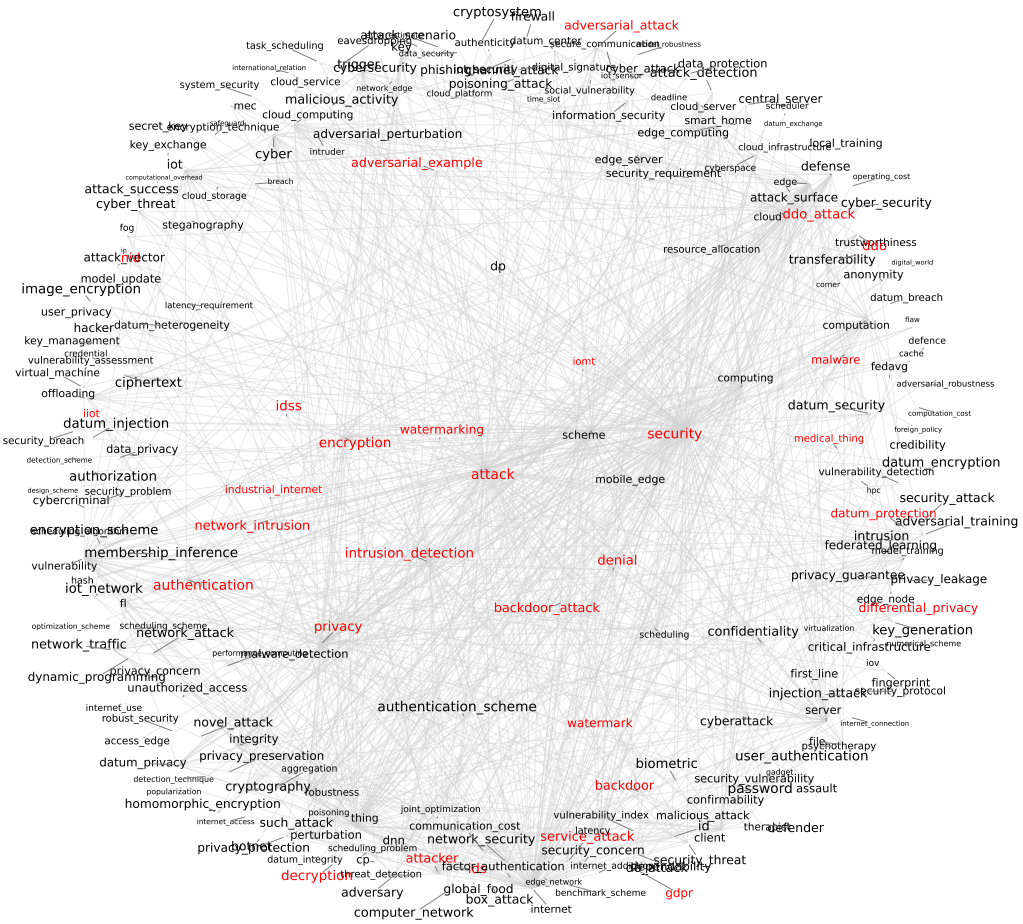

Figure 5: The tenth cluster by the community detection

2. **attack**: A general term for actions aimed at compromising the security of systems, including denial-of-service (DoS), malware, and phishing attacks.
3. **backdoor**: Denotes a method of bypassing security controls to access a system or network, often installed maliciously by attackers to maintain persistent access.
4. **intrusion detection**: Refers to techniques or systems designed to monitor network traffic for suspicious activity and detect unauthorized access or anomalies.

5. **denial**: Likely refers to Denial-of-Service (DoS) attacks, where attackers overwhelm a system with traffic to disrupt its functionality.
6. **medical thing**: Related to Internet of Medical Things (IoMT), which involves secure data communication in medical devices connected to the internet.
7. **encryption**: The process of converting data into a coded format to prevent unauthorized access, commonly used to secure communication in IoT networks.
8. **iomt**: Stands for Internet of Medical Things, a network of connected medical devices that require robust security to protect sensitive health data.
9. **watermark**: Refers to digital watermarking techniques used to embed information into data (e.g., images, videos) for authentication and copyright protection.
10. **watermarking**: The broader concept of embedding information into digital content to ensure data integrity, authentication, and traceability.
11. **network intrusion**: Refers to unauthorized access or activities within a network, often a precursor to cyberattacks like data breaches or system compromise.
12. **security**: A broad term encompassing measures and techniques to protect systems, data, and networks from unauthorized access and cyber threats.
13. **ddo**: Likely a shorthand for Distributed Denial-of-Service (DDoS) attacks, where multiple compromised systems are used to flood a target system or network.
14. **nid**: Stands for Network Intrusion Detection, a type of system that identifies malicious activity within a network to prevent or mitigate cyber threats.
15. **service attack**: Refers to attacks targeting services, such as denial-of-service (DoS) or Distributed Denial-of-Service (DDoS) attacks, aimed at disrupting the availability of services.

Figure 5 primarily focuses on cybersecurity topics, including attacks, defenses, and privacy measures in Internet of Things (IoT) networks.

In table 10, words like ‘backdoor attack’, ‘intrusion detection’, and ‘encryption’ highlight advanced methodologies for identifying and mitigating security breaches in IoT systems. Terms such as ‘watermark’ and ‘authentication scheme’ suggest specific techniques for ensuring data integrity and secure access in IoT environments.

In contrast, in table 10, words like ‘security’, ‘attack’, and ‘scheme’ frequently appear, reflecting their foundational importance but lacking the specificity to detail advanced methods. Additionally, terms like ‘privacy’ and ‘internet’ show broad concepts without delving into the technical nuances present in Distance-IDF results.

5.3.2.4 The Thrity-Third Cluster

Rank	Distance-IDF Weights			Frequency-based Weights		
	Degree	**Betweenness**	**Closeness**	**Degree**	**Betweenness**	**Closeness**
1	cell	cell	immune cell	cell	cell	cell
2	gene	gene	hub gene	gene	gene	gene
3	tissue expression	target	immune infiltration	expression	protein	protein
4	gtex	etiology	ssgsea	protein	expression	expression
5	expression	protein	sample gene	pathway	tumor	tumor
6	methylation	expression	luad	tumor	pathway	pathway
7	tumor	relaxation	cytoscape	tissue	gene expression	gene expression
8	mirna	exclusion	deg	gene expression	damage repair	mouse
9	protein	clinical presentation	expression network	mouse	cell line	activation
10	lncrna	tumor	lung adenocarcinoma	activation	tissue	tissue
11	paraffin	package	enrichment	target	assay	proliferation
12	mouse	pathway	infiltration	prognosis	target	genome
13	relu	tissue	cancer genome	enrichment	human protein	transcription
14	linear unit	chip	wgcna	transcription	strand	enrichment
15	circrna	activation	protein interaction	proliferation	mouse	immune cell

Table 11: Degree, Betweenness, and Closeness Centrality for Cluster 33: Comparison Between Distance-IDF Weights and Frequency-based Weights

Below are the explanations for the top-ranking words in degree centrality for Cluster 33 using Distance-IDF weights:

1. **cell**: Refers to the basic structural and functional unit of life, often studied in the context of cellular mechanisms and molecular interactions.
2. **gene**: Represents a unit of heredity, commonly analyzed in molecular biology to understand genetic expression and regulation.

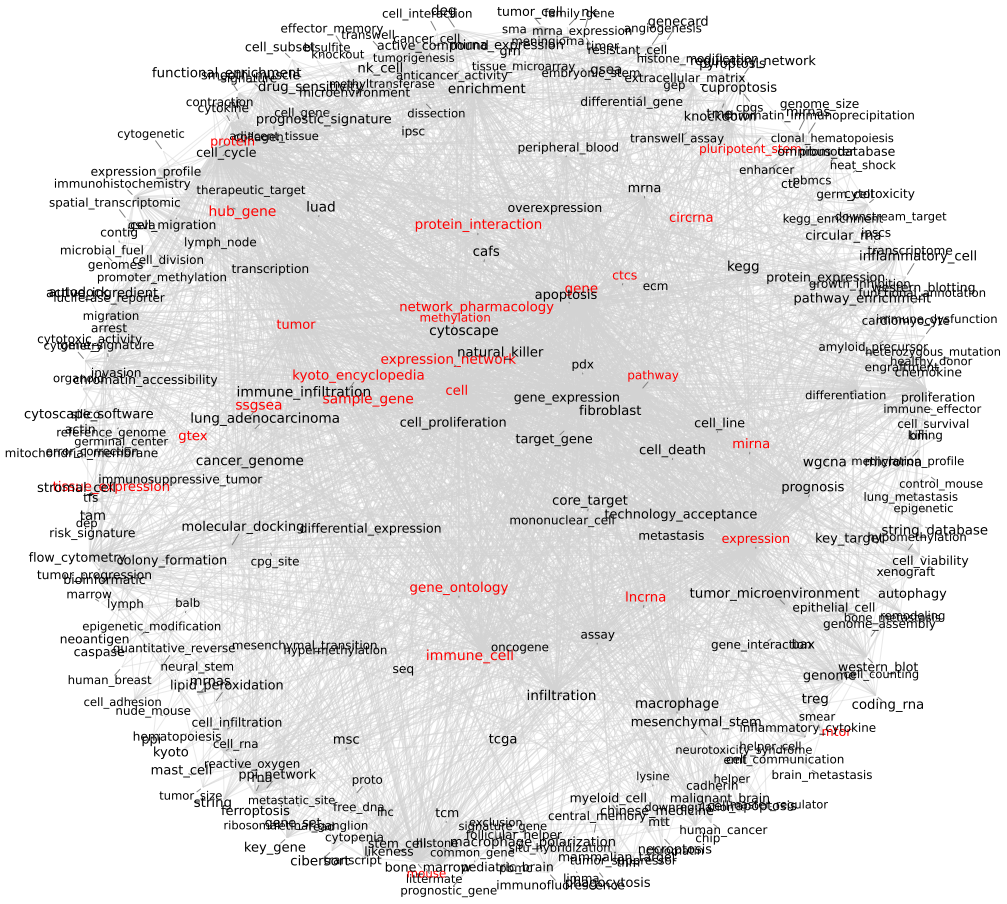

Figure 6: The thirty-third cluster by the community detection

3. **tissue expression**: Refers to the study of gene expression patterns specific to various tissue types, critical in understanding tissue-specific functions and diseases.
4. **gtex**: Stands for the Genotype-Tissue Expression project, a dataset that provides insights into tissue-specific gene expression and regulatory mechanisms.
5. **expression**: Indicates the process by which genes produce functional products (e.g., proteins), commonly studied in transcriptomics.

6. **methylation**: Refers to a chemical modification of DNA that regulates gene expression without altering the DNA sequence, often studied in epigenetics.
7. **tumor**: Denotes an abnormal growth of cells, a central focus in cancer research to understand mechanisms of oncogenesis.
8. **mirna**: Refers to microRNAs, small non-coding RNA molecules that regulate gene expression post-transcriptionally, playing critical roles in cancer and other diseases.
9. **protein**: Represents functional biomolecules encoded by genes, often studied for their role in biological processes and diseases.
10. **lncrna**: Stands for long non-coding RNA, a class of RNA molecules involved in gene regulation, with implications in cancer and other diseases.
11. **paraffin**: Refers to formalin-fixed paraffin-embedded (FFPE) tissue samples, commonly used in histology and molecular studies.
12. **mouse**: Represents laboratory mice used as model organisms for studying genetic, molecular, and disease mechanisms.
13. **relu**: Refers to the Rectified Linear Unit, a common activation function in neural networks, possibly indicating computational analysis in bioinformatics.
14. **linear unit**: Likely refers to linear transformations in computational or statistical models used for analyzing biological data.
15. **circrna**: Stands for circular RNA, a type of non-coding RNA involved in gene regulation and associated with various diseases, including cancer.

Figure 6 primarily focuses on molecular biology topics, including gene expression analysis, tumor biology, and protein interaction networks.

In table 11, words like ‘tissue expression’, ‘methylation’, and ‘gtex’ highlight specific molecular biology methods and datasets, emphasizing advanced techniques for analyzing gene expression and tissue-specific data. Terms like ‘immune infiltration’ and ‘cytoscape’ suggest the cluster’s focus on cancer research and bioinformatics tools for network visualization and analysis.

In contrast, in table 11, words like ‘cell’, ‘gene’, and ‘protein’ frequently appear, reflecting their foundational importance but lacking the specificity to detail advanced molecular biology techniques. Additionally, terms like ‘pathway’ and ‘tumor’ highlight broad concepts without delving into the specific datasets or tools present in the Distance-IDF results.

Chapter 6

Conclusion

This study presents two main contributions. First, we propose a systematic method for research trend analysis using social network methods. Starting with documents as input, word pairs are generated, and the proposed Distance-IDF weights are calculated. Subsequently, a word co-occurrence network is constructed. Words are represented as nodes, edges denote co-occurrence relationships, and edge weights are assigned using the Distance-IDF values. Based on this network, community detection is performed to identify multiple clusters. Finally, degree, betweenness, and closeness centrality are calculated to identify the topics or key terms discussed within each cluster.

Second, we demonstrate a data mining approach by applying our methodology to analyze statistical papers from 2023 using the Semantic Scholar database. By employing word co-occurrence networks, we identify word pairs and compile a list of significant terms in the field of statistics. Additionally, we applied community detection techniques to identify clusters within the network and analyzed the topics covered by these clusters.

In this word co-occurrence network, we identified several topics based on different clusters. Cluster 1 focuses on advanced techniques in computer vision, including image segmentation, deep learning architectures, and applications in medical imaging. Cluster 9 shows wireless sensor networks (WSNs), emphasizing energy-efficient communication, graph-based methods, and routing strategies. Cluster 10 centers on cybersecurity in IoT networks, addressing cyberattacks, encryption techniques, and privacy measures. Finally, Cluster 33 delves into molecular biology, covering gene expression, cancer research, and bioinformatics tools.

Through our results, we confirmed the effectiveness of our proposed weight calculation method in capturing four key aspects of word relationships in the co-occurrence network. First, **concordance** was evident as words that co-occurred in more documents were assigned higher weights. Second, **proximity** was reflected in the higher weights for words located closer together within individual documents. Third, **density** was observed in the way closely grouped words formed stronger connections in the network. Finally, **anti-universality** was demonstrated by the higher weights assigned to less universally used words, highlighting their significance within specific contexts.

References

Asatani, K., Mori, J., Ochi, M., and Sakata, I. (2018). Detecting trends in academic research from a citation network using network representation learning. *PloS one*, 13(5):e0197260.

Bermingham, A., Conway, M., McInerney, L., O'Hare, N., and Smeaton, A. F. (2009). Combining social network analysis and sentiment analysis to explore the potential for online radicalisation. In *2009 International Conference on Advances in Social Network Analysis and Mining*, pages 231–236. IEEE.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Dijkstra, E. W. (2022). A note on two problems in connexion with graphs. In *Edsger Wybe Dijkstra: his life, work, and legacy*, pages 287–290.

Keidar, D., Zhong, M., Zhang, C., Shrestha, Y. R., and Paudel, B. (2021). Towards automatic bias detection in knowledge graphs. *arXiv preprint arXiv:2109.10697*.

Lee, M. and Song, M. (2020). Incorporating citation impact into analysis of research trends. *Scientometrics*, 124(2):1191–1224.

Lozano, S., Calzada-Infante, L., Adenso-Díaz, B., and García, S. (2019). Complex network analysis of keywords co-occurrence in the recent efficiency analysis literature. *Scientometrics*, 120:609–629.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582.

Nguyen, V. T., Kravets, A. G., and Duong, T. Q. (2021). Predicting research trend based on bibliometric analysis and paper ranking algorithm. *Cyber-physical systems: digital technologies and applications*, pages 109–123.

Opsahl, T., Agneessens, F., and Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social networks*, 32(3):245–251.

Radhakrishnan, S., Erbis, S., Isaacs, J. A., and Kamarthi, S. (2017). Novel keyword co-occurrence network-based methods to foster systematic reviews of scientific literature. *PloS one*, 12(3):e0172778.

Sivanandham, S., Sathish Kumar, A., Pradeep, R., and Sridhar, R. (2021). Analysing research trends using topic modelling and trend prediction. In *Soft Computing and Signal Processing: Proceedings of 3rd ICSCSP 2020, Volume 1*, pages 157–166. Springer.

Tibaná-Herrera, G., Fernández-Bajón, M. T., and de Moya-Anegón, F. (2018). Mapping a research field: Analyzing the research fronts in an emerging discipline. In *Scientometrics*. IntechOpen.

Yuan, C., Li, G., Kamarthi, S., Jin, X., and Moghaddam, M. (2022). Trends in intelligent manufacturing research: a keyword co-occurrence network based review. *Journal of Intelligent Manufacturing*, 33(2):425–439.

국문 초록

소셜네트워크 분석법을 활용한 연구동향 분석에 관한 연구

성신여자대학교
대학원
통계학과
이윤진

연구 동향 분석은 특정 분야에서 중요한 역할을 하는 핵심 주제와 키워드를 파악하고, 대규모 데이터에서 효율적으로 정보를 처리할 수 있도록 합니다. 사회 네트워크 분석은 단어 간의 상호작용을 네트워크 구조로 시각화하여 복잡한 시스템의 연결성과 구조를 이해할 수 있게 합니다. 또한, 네트워크에서 클러스터를 발견하여 유사한 특성을 공유하는 그룹을 식별할 수 있습니다. 본 연구에서는 사회 네트워크 분석 기법을 활용하여 연구 동향 분석을 수행하는 방법론을 제안합니다. 구체적으로, 대규모 텍스트 데이터에서 동시출현 단어 네트워크를 생성하고, 문서 내 단어 출현 빈도와 단어 간 거리를 반영한 새로운 가중치 계산 방식을 제안하였습니다. 제안한 에지 가중치는 동시출현성, 근접성, 밀집성, 반보편성을 모두 반영하여, 기존의 동시출현성에 기반한 방법을 보완하였습니다. 이를 바탕으로 커뮤니티 탐지를 수행하여 각 클러스터별로 나타나는 주제를 식별하는 과정을 제시하였습니다. 이 방법론의 유효성을 검증하기 위해 Semantic Scholar 데이터베이스에서 추출한 논문 초록에 적용하였으며, 의미 있는 연구 주제를 효과적으로 식별할 수 있음을 입증하였습니다.

핵심용어 : 소셜네트워크 분석, 연구 동향 분석, 단어 동시출현 네트워크, 중심성, 커뮤니티 탐지