electronics

MDPI

Article

Technology Keyword Analysis Using Graphical Causal Models

Sunghae Jun

Department of Data Science, Cheongju University, Chungbuk 28503, Republic of Korea; shjun@cju.ac.kr; Tel.: +82-10-7745-5677; Fax: +82-43-229-8432

Abstract: Technology keyword analysis (TKA) requires a different approach compared to general keyword analysis. While general keyword analysis identifies relationships between keywords, technology keyword analysis must find cause–effect relationships between technology keywords. Because the development of new technologies depends on previously researched and developed technologies, we need to build a causal inference model, in which the previously developed technology is the cause and the newly developed technology is the effect. In this paper, we propose a technology keyword analysis method using casual inference modeling. To understand the causal relationships between technology keywords, we constructed a graphical causal model combining a graph structure with causal inference. To show how the proposed model can be applied to the practical domains, we collected the patent documents related to the digital therapeutics technology from the world patent databases and analyzed them by the graphical causal model. We expect that our research contributes to various aspects of technology management, such as research and development planning.

Keywords: technology keyword; graph structure; causal inference; Poisson regression; patent document; digital therapeutics

Citation: Jun, S. Technology Keyword Analysis Using Graphical Causal Models. *Electronics* **2024**, *13*, 3670. https://doi.org/10.3390/electronics13183670

Academic Editors: Boni García, Mario Munoz-Organero, Patricia Callejo and Miguel Ángel Hombrados

Received: 13 August 2024
Revised: 11 September 2024
Accepted: 13 September 2024
Published: 15 September 2024

1. Introduction

Technology keyword analysis (TKA) involves analyzing the keywords extracted from various technological documents such as patents and papers for technology management. Up to now, TKA has been conducted across diverse fields [1–5]. The patent system grants the applicant exclusive rights to use the developed technology. Patents contain more specific and detailed descriptions of the developed technology compared to papers. Therefore, many studies have been conducted on technology analysis models using patents [6–12]. TKA is a popular method for patent technology analysis [1,2,6,9,11]. Jun (2024) used the synthpop and the generative adversarial network (GAN) to analyze the patent keywords [1]. This paper tried to combine the methods of statistics and machine learning algorithms to build an analytical model for patent keyword analysis. The synthpop and the GAN are popular generative models based on statistics and machine learning, respectively. Kim and Jun (2015) studied a patent keyword analysis model based on a graphical causal inference and copula regression model [11]. In the proposed model, they mainly focused on the copula model as a graph inference model. In addition, they used the patent documents related to the technologies of Apple. Xue and Shao (2024) conducted a patent technology analysis to understand the hydrogen energy industry and technology using text mining [6]. The authors applied vectorization to their searched patent documents. In addition, they performed topic clustering and constructed the associated structure between the patent keywords to find the technological trends in the developed technology. Park and Jun (2022) proposed a statistical method of patent keyword analysis for technology management, such as technology forecasting and new product development [10]. The authors analyzed patent keyword data using Bayesian inference-based analysis and network modeling. Park and Jun (2023) also considered the zero-inflated problem in the patent keyword data [9]. In general, the number of keywords in patent data is much larger than the number of documents, which causes the zero-inflated problem. To solve this problem, they proposed a

compound Poisson regression model. That is, the authors studied patent keyword analysis using a compound Poisson model because it was difficult to analyze zero-inflated patent data using a general Poisson model.

Most of the existing patent keyword analysis models focused on the association structure between keywords. That is, the technology structure was identified through the correlation coefficients between connected keywords. However, technology is often composed of preceding technology that has been developed and subsequent technology that needs to be developed. In other words, there is a need to model the causal structure of technology to be able to understand the technologies that cause and those that that result. In other words, we need a technology analysis model that can understand the causal structure of cause technology and effect technology. To overcome the gap between research on patent keyword analysis and its application in real-world domains, in this paper, we carry out a modeling study on the causal relationships between technologies. To perform a TKA task, we have to collect the technology documents. In our research, we search the patent documents from the world patent databases and preprocess them to build a structured data for performing TKA. Using text mining as a preprocessing tool, we construct a matrix consisting of patent documents and keywords, called the patent–keyword matrix [13,14]. Each element of the matrix represents the frequency value of a keyword occurring in a document. We analyze the data matrix using a graphical causal model (GCM). The GCM is composed of the graph structure and causal inference, so we represent the technology keywords and their relationships as vertexes and edges in the graph. The edges show the causal relationships between the technology keywords. We expect that our research results will contribute to various tasks in technology management that require the quantitative analysis of technology trends. When the government or a company tries to establish the research and development (R&D) strategy for a new technology, they can build a more efficient R&D strategy for the target technology using the result of our proposed model.

This paper consists of the following sections. In Section 2, we address our research background, including graph theory and causal inference modeling. We propose a TKA by the GCM in Section 3. In the next section, we carry out the experiments using practical patent documents related to digital therapeutics to show how this research can be applied to real-life domains. We explain the conclusions and contributions of our paper in the conclusion section.

2. Graph and Causal Inference

2.1. Graph Structure

A graph is a data representation that shows the relationships between data objects [15–17]. In general, a Graph G consists of nodes and edges, as described in (1) [15].

$$G(V, E), v \in V, e \in E \quad (1)$$

where V and E are the sets of vertices (nodes) and edges, respectively. The v and e are each vertex and edge, respectively. In a graph model, the vertices can be various objects, such as variables and keywords. In addition, edges represent the relationship between different data objects. Graphs are divided into directed and undirected graphs, depending on the direction of the edges connecting the nodes. In technology keyword data analysis using the graph structure, each node is a technology keyword and an edge represents the direction of technological relationships between keywords [11]. In order to connect the different nodes in the graph structure, we need a measure to calculate the association between the nodes. In general, we use the correlation coefficient or the conditional probability as a measure for computing associations. We can also consider causal inference to compute the causal relationships between the nodes in a graph. In the next section, we explain causal inference.

2.2. Causal Inference Models

A causal model represents the relationships between causes and their effects, finding the outcomes of interventions [18,19]. This cause–effect relationship is dependent on

causal inference. Causal inference is the process of inferring a causal relationship between one variable and another [16–20]. Correlation analysis involves finding a simple correlation structure between two variables, while causal inference models the cause-and-effect relationship between two variables. For example, a causal inference for two nodes v1 and v2 is expressed as v1→v2. We refer to v1 and v2 as the parent and the child. In addition, v1 and v2 perform the roles of cause and effect in the causal inference model.

Most previous keyword analyses focused on the correlation between keywords [2–5,9]. Therefore, most keyword analyses models tried to find the efficient correlations between keywords. However, since we cannot identify the cause–effect relationship between keywords through the association analysis results, there are limitations in the use of keyword analysis results by association. Especially in the field of technology management, where new technologies are planned and developed based on previously researched and developed technologies, the TKA we conduct must be based on causal inference. To solve this problem, we propose a TKA based on causal inference in the next section.

3. Proposed Method

In general, the correlation analysis can only confirm the correlation structure between two variables, so it cannot find the cause-and-effect relationship between variables. Therefore, we have limitations in identifying relationships between the keywords using the correlation or association analyses. In particular, in the TKA, the keywords representing preceding technology influence the keywords describing subsequent technology, so we need a model that can explain the cause and effect between keywords. To overcome the limitation of TKA, we use a causal inference model in this paper. The investigation of causal relationships between technology keywords is a very interesting field of inquiry in patent technology analysis. This is because, by understanding the cause–effect relationships between the keywords included in patent documents, we can identify the technology structure between sub-technologies for developing target technology. For this reason, we propose a TKA using a causal inference model. Most statistical and machine learning analysis techniques require structured input data for learning purposes. That is, the columns of the constructed data should be represented as variables and the rows as observations. Therefore, using the text mining techniques, we have to transform the patent documents into structured data for the keyword analysis. First, we search patent documents related to target technology from the patent databases in the world because patents contain diverse and detailed information about the developed technology. Using the following steps, we can ontain the patent–keyword matrix data for technology analysis.

(Step 1) Collecting patent document data:

- (1-1) Selecting target technology;
- (1-2) Searching patents related to target technology;
- (1-3) Determining valid patents.

(Step 2) Preprocessing patent text data:

- (2-1) Performing tokenization;
- (2-2) Performing normalization: stemming, lemmatization, lowercasing, deleting stop words, removing punctuation and meaningless characters.

(Step 3) Building a patent document–keyword matrix:

- (3-1) Finding unique vocabulary from corpus;
- (3-2) Constructing a data matrix:
  rows = patent documents, columns = keywords from the vocabulary;
  matrix values = frequency values of each keyword in a document.

Once the target technology is determined, we first search for patent documents related to this technology. Among the searched patent documents, we go through the process of selecting only the valid patents that will be used for the final analysis. Next, we preprocess the patent documents by performing tokenization and normalization. In this process, we conduct stemming, lemmatization and lowercasing. We also carry out tasks such as

deleting stop words, removing punctuation and meaningless characters. Finally, we build the patent document–keyword matrix by finding unique vocabulary from the corpus. The row and column represent patent documents and keywords, respectively. The matrix value is the frequency of each keyword in a document. Therefore, Figure 1 shows our structured data for the TKA.

Matrix	$Keyword_1, Keyword_2, \cdots, Keyword_p$
$Patent_1$	$X_{ij} = 0, 1, 2, \ldots$
$Patent_2$	$i = 1, 2, \ldots, n$
⋮	$j = 1, 2, \ldots, p$
$Patent_n$	X_{ij}: $frequency\ of\ Keyword_j\ included\ in\ Patent_i$

Figure 1. Patent–keyword matrix constructed from patent documents.

In Figure 1, we use a patent–keyword matrix for the TKA for the causal inference models. This matrix consists of patents as rows and keywords as columns, respectively. In addition, the element of the matrix, X_{ij}, is the frequency of $Keyword_j$ included in $Patent_i$. As a result, we analyze the matrix data for our model. To visually represent the structural causal relationships between the patent keywords, we use a graph causal inference model based on the graph structure. Table 1 shows the graph structure and the patent–keyword matrix.

Table 1. Comparing the graph structure and patent–keyword matrix.

Graph	Representation	Patent–Keyword Matrix
Node	Variable	Keyword
Edge	Cause	Technology relationship

A graph consists of a pair of V and G, represented by G(V, E), where V and G are a set of nodes and edges [16]. E is a set of the conditional probabilities between variables. A node in the graph is a variable representing a technology keyword in the patent–keyword matrix. In addition, an edge representing the cause refers to the relationship between the keywords that represent sub-technologies. For constructing the GCM, we consider two approaches: learning the causal structure and estimating causal effects. The proposed method is based on the PC (Peter and Clark) algorithm using a conditional independence test (CIT) [16,17]. We start from fully connected nodes in an undirected graph. In this graph, we examine the conditional independence of a pair of nodes through the CIT. We do this until finishing the CIT for all node pairs.

A representative method that performs causal inference between the nodes included in a graph structure using the CIT as a basic component is the PC algorithm [16,17]. That is, the PC algorithm uses the CIT to infer a causal structure model from the data. The PC algorithm refines the graph structure iteratively while depending on the CIT. The algorithm starts with a complete graph and removes the meaningless edges according to the CIT results. Therefore, we use the PC algorithm in this paper. The process of the PC algorithm for the TKA is as follows:

(Step 1) Building a complete undirected graph based on technology keywords.
(Step 2) Performing a conditional independence test:
  (2-1) Determining the significance level α;
  (2-2) Using subsets of adjacency sets related to all pairs of technology keywords.
(Step 3) Deleting the edges of a pair of nodes with conditional independence.
(Step 4) Constructing a graphical causal model using the remaining edges.

A graphical model is a dependence structure based on probability distributions. The nodes in the graph are random variables, and the edges represent the dependence between the variables (keywords). In this paper, we also consider a DAG as a GCM. Because the DAG represents conditional dependencies between nodes, we can use the former node (A), representing the cause, to infer the latter node (B), representing the result A→B. That is, the relationship A→B, describing the frequency value of the keyword B, is directly influenced by the value of the keyword A. In this relationship, A is a parent of B and B is a child of A. Using the PC algorithm, we perform the graphical causal modeling, such as constructing a DAG.

A graphical model of the keywords $(W_1, W_2, \ldots, W_p)$ is represented by the conditional independence relationship, as follows [16,17]:

$$P(W_1, W_2, \ldots, W_p) = \prod_{i=1}^{p} P(W_i|PA_i) \tag{2}$$

where PA_i is a subset node of the keywords that precede the keyword W_i. We construct a DAG represented by the parents of W_i, which directly influence W_i. In addition, we apply Equation (2) to the independencies and the carry out the CIT as follows [16,17,20]:

(Step 1) Skeleton of directed acyclic graph (DAG):

(1-1) Estimating the DAG skeleton;
(1-2) Starting with a complete undirected graph.

(Step 2) Testing the constraint of each edge (between W_i and W_j):

(2-1) Defining the conditional set C;
(2-2) Deleting W_i and W_j if W_i and W_j are conditionally independent given C;
(2-3) Building the separation set (W_i,W_j).

In the graphical causal inference for the TKA, the graph model is represented as $\mathrm{M}(W, P)$. In this model, W is a keyword set, and P is the edge representing the conditional probability between the keywords. That is, P is defined by the ith keyword W_i and the parent keywords of W_i, PA_i. For example, if four keywords, W_1, W_2, W_3 and W_4 have the same graph structure as $W_1 \rightarrow W_2 \rightarrow W_3 \rightarrow W_4$, we can use the joint probability distribution to perform graph causal inference, as in Equation (3).

$$P(W_1, W_2, W_3, W_4) = P(W_1)P(W_2|W_1)P(W_3|W_2)P(W_4|W_3) \tag{3}$$

In the $\mathrm{M}(W, P)$, the keyword included in W is a random variable. We can also see how a particular keyword affects another keyword through a set of edges, P. That is, P represents the causal relationships between the technology keywords. Therefore, using the process of the graphical causal modeling, we can build the visualization of the technological keywords. If there is an edge between W_c and W_e, $W_c \rightarrow W_e$, we consider a generalized linear model to understand the causal relationship in more detail about this edge. In this relationship, W_c and W_e are cause and effect keywords, respectively. To examine the relationship between the two keywords in more detail, we build a generalized linear model (GLM) where W_e is the dependent variable and W_c is the independent variable. In this paper, we use the Poisson regression model as a GLM for the edge because the given data are count-type. In the Poisson regression model, the W_e is a random variable with the following Poisson distribution [21–24]:

$$f(W_e) = \frac{e^{-m} m^{W_e}}{W_e!}, \; W_e = 0, 1, 2, \ldots \tag{4}$$

In the Equation (4), m is the parameter of Poisson distribution and represents the mean value. Therefore, the Poisson regression model is represented as follows [25–28]:

$$log(m) = \beta_0 + \beta_1 W_{c1} + \beta_2 W_{c2} + \cdots + \beta_p W_{cp} \tag{5}$$

In the regression model of Equation (5), $W_{c1}, W_{c2}, \ldots, W_{cp}$ are the independent variables (keywords), and $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the model parameters. Using Equation (5), we can obtain the regression equation for m, as follows:

$$m = e^{\beta_0 + \beta_1 W_{c1} + \beta_2 W_{c2} + \cdots + \beta_p W_{cp}} \quad (6)$$

According to Equation (6), we estimate the model parameters from the given data and compute m to predict W_e. In addition, we carry out the hypothesis testing to find the statistical significance of each parameter, as follows [26]:

$$H_0 : \beta_i = 0 \quad \text{vs.} \quad H_1 : \beta_i \neq 0 \quad (7)$$

When we reject H_0 in Equation (7), we determine that the keyword W_{ci}, corresponding to β_i, is statistically significant. In this paper, we use the p-value to decide to reject H_0 or not. The threshold of p-value is determined according to significance level $\alpha (0 \leq \alpha \leq 1)$. For example, in the 95% confidence interval, the value of α is 0.05 and we can reject H_0 when the value of α is less than 0.05.

In this paper, our proposed method performs the TKA by four sub-methods from M1 to M4, as follows:

(M1) Building the patent–keyword matrix;
(M2) Graphical causal inference modeling;
(M3) Poisson regression modeling;
(M4) Constructing the keyword diagram of cause and effect technologies.

In the first sub-method (M1), we build the patent–keyword matrix from the searched patent documents using text mining techniques. Next, we apply the GCM to the patent–keyword matrix in the second sub-method (M2), and we find the causal structure for the target technology. In the third sub-method (M3), we perform Poisson regression modeling using the cause and effect keywords from the GCM results. Finally, we construct the technology keyword diagram of causes and effects in the fourth sub-method (M4). We can use these results in various fields of technology management. In various technology domains, once the target technology is decided on, we collect relevant patent documents from patent databases around the world and perform text preprocessing by selecting only valid patents. The patent–keyword matrix data constructed through this process is used in graph causal inference modeling. Using the results of graph causal inference modeling, we perform Poisson regression analysis and ultimately construct a technology diagram for the target technology. We expect that the technology keyword analysis method proposed in this paper can be utilized in various fields of technology management.

4. Experimental Results

4.1. Experimental Data

We used the patent documents in our experiments for the technology keyword analysis. We searched the patents related to digital therapeutics technology from the world patent databases [29,30]. Digital therapeutics refer to software provided to patients to prevent, manage, and treat diseases and disorders [31–33]. The technology of digital therapeutics is still in the early stages of development worldwide, and much research on this technology is underway in various fields. In this paper, we collected data from patent documents related to digital therapeutics technology and analyzed them using the graph causal inference model. First, we extracted technology keywords from the patent documents and constructed a patent–keyword matrix, shown in Figure 2.

As shown in Figure 2, we finally selected 12 technology keywords from 2685 valid patents. The elements of this matrix represent the frequency of specific keywords included in each patent document. $Fq_{i,j}$ represents the frequency of $Keyword_i$ included in $Patent_j$. In addition, the selected keywords are as follows: analysis, compute, digit, generate, intelligent, learn, machine, network, sensor, signal, smart, therapeutics. In the next section,

we built a graph causal inference model using the patent–keyword matrix data in Figure 2. We used the data language R and its provided packages for our experiments [14,34,35].

Matrix	$Keyword_1$	$Keyword_2$	⋯	$Keyword_{12}$
$Patent_1$	$Fq_{1,1}$	$Fq_{1,2}$	⋯	$Fq_{1,12}$
$Patent_2$	$Fq_{2,1}$	$Fq_{2,2}$	⋯	$Fq_{2,12}$
⋮	⋮	⋮	$Fq_{i,j}$	⋮
$Patent_{2685}$	$Fq_{2685,1}$	$Fq_{2685,2}$	⋯	$Fq_{2685,12}$

Figure 2. Patent–keyword matrix for graphical causal modeling.

4.2. Analyzing Technology Keyword Data Using Graphical Causal Modeling

A single technology field is made up of a causal structure in which many different sub-technologies influence each other. Likewise, the digital technology field also exhibits a causal structure, with causal technologies influencing other technologies and the resulting technologies being influenced in turn. Therefore, we tried to construct the causal structure of digital therapeutics technology using our proposed method in our experiments. To understand the technology of digital therapeutics, we have three questions, as follows:

(Q1) What are the causes within the field of digital therapeutics technology?
(Q2) What are the effects within the field of digital therapeutics technology?
(Q3) What is the mediator that connects cause and effect in the field of digital therapeutics?

In questions 1 and 2 (Q1 and Q2), we look for patent technology keywords that correspond to causes and effects within the field of digital therapeutics. Additionally, in question 3 (Q3), find the patent keywords corresponding to mediating technologies that connect the causes and effects in the field of digital therapeutics. To address the three questions, we analyzed patent–keyword matrix data using the graph causal inference model.

First of all, to analyze the technology keyword data, we applied the GCM to analyze the patent–keyword matrix. We carried out the GCM according to the significance level of the individual conditional independence tests. Figure 3 shows the GCM result with $\alpha = 0.05$.

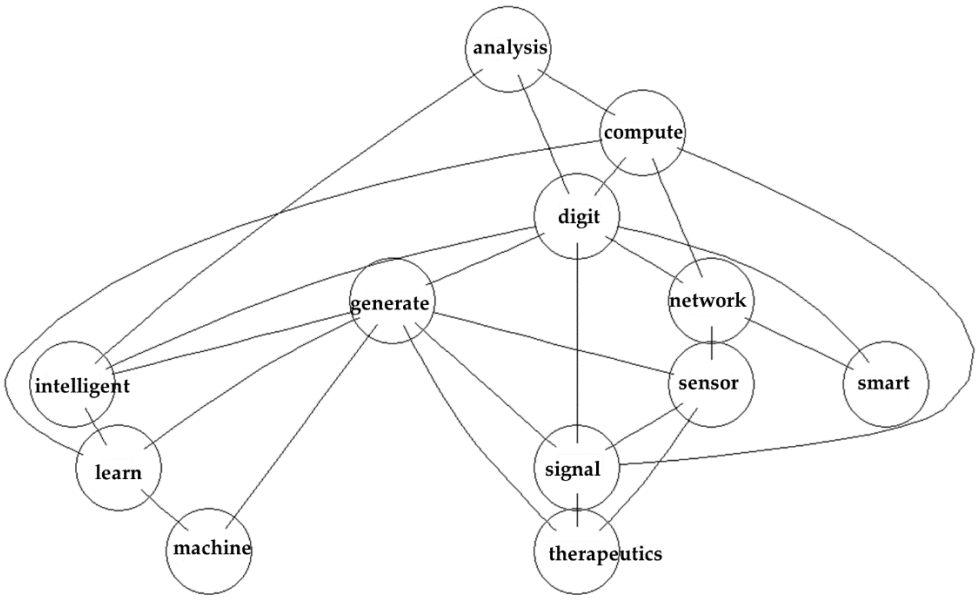

Figure 3. GCM result of the initial skeleton with $\alpha = 0.05$.

This is an undirected and fully connected graph of the initial skeleton. Overall, we were able to find that the three keywords, digit, generate and signal, were highly connected

to other keywords. The keyword digit is connected to the five keywords, analysis, compute, generate, network and signal. That is, the technology of digit is related to the technologies based on the five keywords. Next, we show the GCM result by PC algorithm with $\alpha = 0.05$.

In Figure 4, we represent the direction of each keyword with in- and out-directions. The keywords with the out-direction perform the role of cause node (vertex) in the graph. In addition, the keywords with the in-direction are influenced by the keywords with the out-direction. Therefore, the keyword signal is affected by the keywords compute, generate, sensor and therapeutics. The keyword digit is influenced by the keywords analysis, compute, signal and smart. In addition, the keyword generate receives influence from the keywords digit, intelligent, learn, machine, sensor and therapeutics. Next, Figure 5 represents the initial skeleton result of GCM with $\alpha = 0.01$.

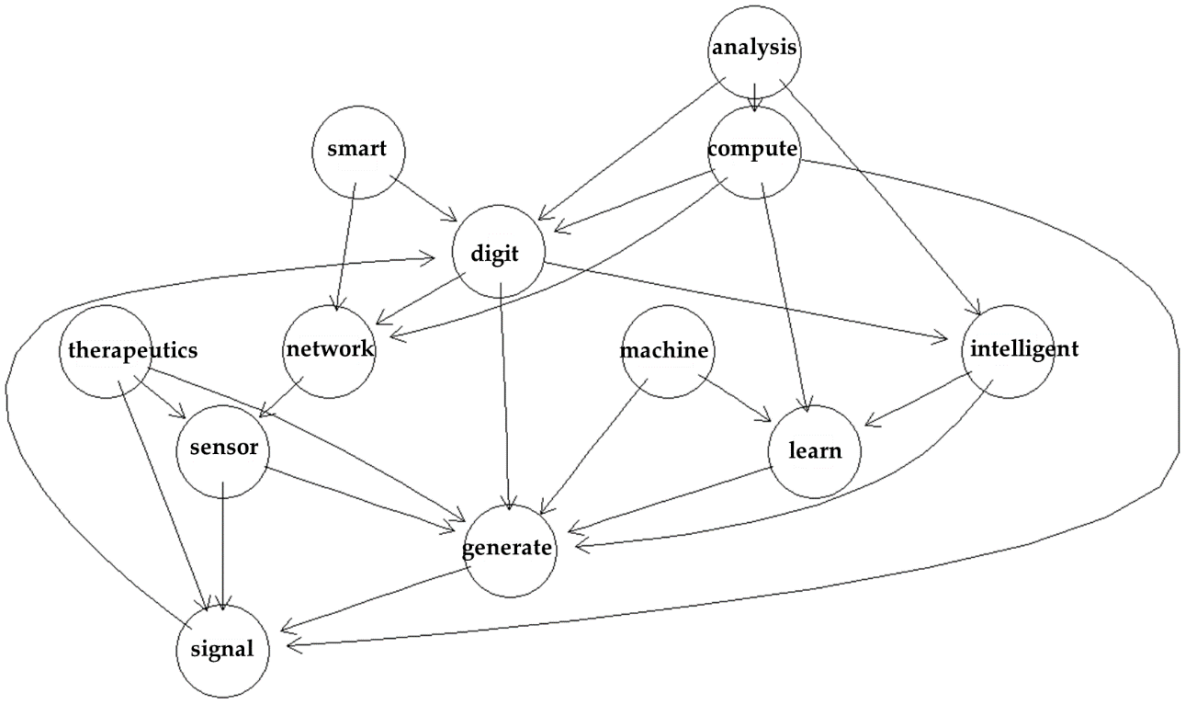

Figure 4. GCM result based on the PC algorithm with $\alpha = 0.05$.

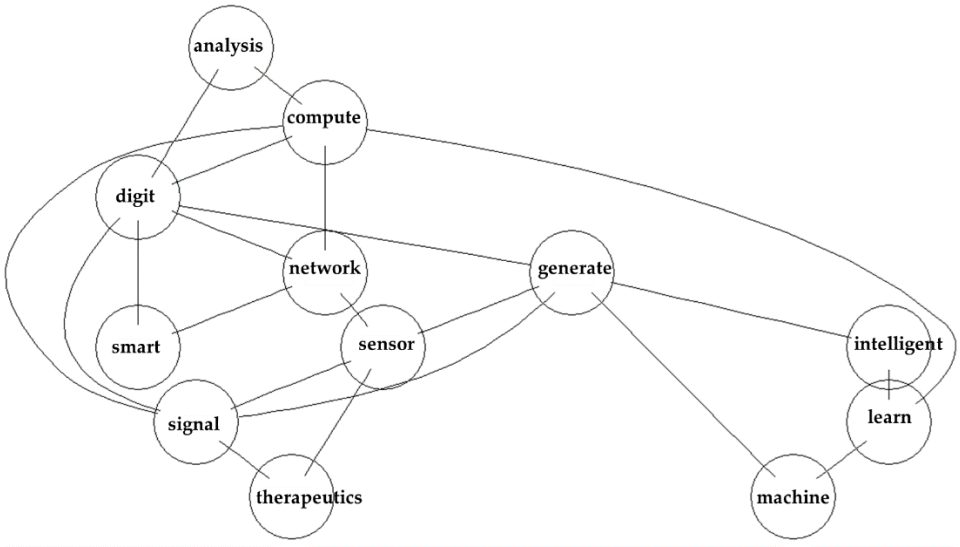

Figure 5. GCM result of initial skeleton with $\alpha = 0.01$.

In Figure 5, the GCM result of the initial skeleton with $\alpha = 0.01$ is similar to the GCM result with $\alpha = 0.05$. However, there is a difference in the size of in- and out-directions. In this graph, the edge size of the keyword signal is the largest. On the other hand, in the GCM graph result with $\alpha = 0.05$, the keyword generate had the largest number of edges, while the keyword signal had the second largest number of edges. Next, we show the GCM result based on the PC algorithm with $\alpha = 0.01$ in Figure 6.

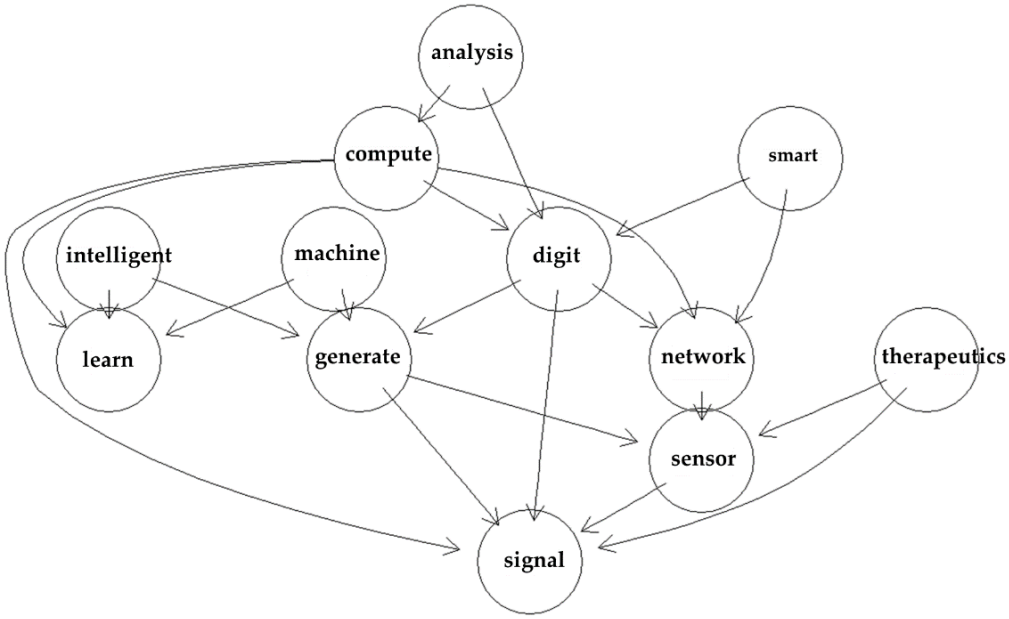

Figure 6. GCM result based on PC algorithm with $\alpha = 0.01$.

This adds the in- and out-directions to the graph of Figure 5. We found that the number of in-directions for the keyword signal is five, so this keyword is influenced by the five keywords compute, digit, generate, sensor and therapeutics. The four keywords digit, generate, network and sensor all have three in-directions each. That is, these keywords are influenced the most by the graph structure of Figure 6, following the keyword signal. Therefore, we represent the in- and out-directions of all keywords in Table 2.

Table 2. Causes and effects of technology keywords.

Keyword	$\alpha = 0.05$		$\alpha = 0.01$	
	In	Out	In	Out
analysis	0	3	0	2
compute	1	4	1	4
digit	4	3	3	3
generate	6	1	3	2
intelligent	2	2	0	2
learn	3	1	3	0
machine	0	2	0	2
network	3	1	3	1
sensor	2	2	3	1
signal	4	1	5	0
smart	0	2	0	2
therapeutics	0	3	0	2

This shows the in- and out-edges according to the keywords at significance levels of $\alpha = 0.05$ and 0.01. First, at the significance level of $\alpha = 0.05$, the in-direction of the keyword generate is the largest. The keywords of digit and signal are second largest. Therefore, we found that the keywords digit, generate and signal are mainly affected by other keywords.

Next, in the out-direction, the numbers associated with the keywords compute, digit and therapeutics are larger than those of the other keywords. Therefore, these three keywords perform the role of cause in the GCM. Second, at the significance level of $\alpha = 0.01$, we found that the keywords digit, generate, network, sensor and signal have relatively larger values compared to the others. Therefore, these keywords are influenced by other keywords in the graph model. The keywords compute and digit have relatively larger numbers of the out-direction. We also see that the keyword digit acts as both a cause and effect because it has large values in both in- and out-directions. We identified which keywords were acting as cause and effect. Using this, we performed a Poisson regression analysis with keywords corresponding to cause and effect as the independent and dependent variables, respectively. Table 3 shows the result of the Poisson regression analysis, using the GCM results with a significance level of $\alpha = 0.05$.

Table 3. Poisson regression model using the GCM results with $\alpha = 0.05$.

Dependent	Independent	$\lvert \beta \rvert$	*p*-Value
digit	analysis	0.7398	<0.0001
	compute	0.0395	0.0503
	signal	0.1952	<0.0001
	smart	0.6709	<0.0001
generate	digit	0.3594	<0.0001
	intelligent	0.5361	<0.0001
	learn	0.3201	<0.0001
	machine	0.2507	<0.0001
	sensor	0.1344	<0.0001
	therapeutics	0.0110	0.6130
signal	compute	0.1045	<0.0001
	generate	0.3181	<0.0001
	sensor	0.1756	<0.0001
	therapeutics	0.1421	<0.0001

First, we found that in the model with the keyword digit as the dependent variable, the remaining three keywords, analysis, signal and smart, excluding compute, statistically significantly explained the keyword digit. Next, in the case of the regression model where the keyword generate was the dependent variable, the remaining five keywords, digit, intelligent, learn, machine and sensor, excluding the keyword therapeutics, statistically significantly explained the keyword generate. Finally, in the model where the dependent variable was the keyword signal, all four keywords, compute, generate, sensor and therapeutics, used as independent variables, had a statistically significant effect on the keyword signal. We also represent the Poisson regression result using the GCM results, with $\alpha = 0.01$ in Table 4.

Table 4. Poisson regression model using the GCM results with $\alpha = 0.01$.

Dependent	Independent	$\lvert \beta \rvert$	*p*-Value
signal	compute	0.0975	<0.0001
	digit	0.2816	0.0006
	generate	0.3122	<0.0001
	sensor	0.1682	<0.0001
	therapeutics	0.1533	<0.0001

Since the number of the in-direction decreases as the α decreases in the GCM, only one regression model with the keyword signal as the dependent variable was performed in Table 4. In this regression model, we confirmed that all five keywords, compute, digit, generate, sensor and therapeutics, used as independent variables, statistically significantly explained the keyword signal. Using all the experimental results conducted so far in this

paper, we created the technological relationship structure related to digital therapeutics depicted below.

In Figure 7, the technology of digital therapeutics consists of two technological groups, according to effect and cause. The effect group is represented by technologies based on the sub-technologies generate and signal. In addition, the technology corresponding to the cause group is based on the sub-technologies analysis, compute, intelligent, machine, smart and therapeutics. We also found that the sub-technology related to digit corresponds to both the effect and cause groups. In the following sections, we present our conclusions of this paper and suggest directions for future research related to our research.

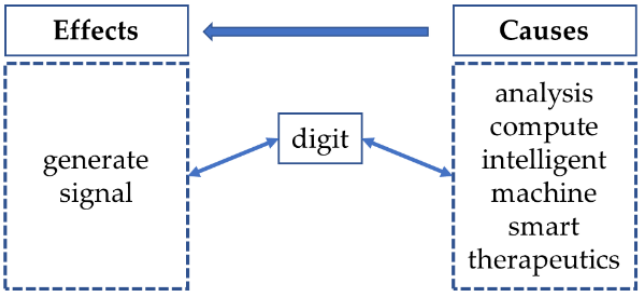

Figure 7. Technological relationship structure in digital therapeutics using GCM.

5. Conclusions

The goal of this paper was to create the causal structure for target technology using TKA. In this paper, we proposed a TKA using graph modeling and causal inference, called the GCM. We searched the patent documents from the world patent databases to collect technology keyword data. Using text mining techniques, we preprocessed the patent documents to construct the patent–keyword matrix. We used the patent document data for our research because a patent contains various information related to the developed technology, such as title, abstract, claims, citations, etc. Therefore, we applied the GCM to the patent–keyword matrix to build the causal structure of technologies. We chose digital therapeutics as our target technology for the TKA. Using the patent–keyword matrix related to digital therapeutics technology, we developed the causal structure of the technology based on the results from the PC algorithm, in- and out-directions and Poisson regression modeling. In our experimental results, we found that the keywords analysis, compute, intelligent, machine, smart and therapeutics correspond the cause technologies in the field of digital therapeutics. In addition, we confirmed that the keywords that play an effect role are generate and signal. Finally, we could see that the keyword digit serves the role of mediator between cause and effect.

The significance of our study is that it models the causal structure that explains cause and effect between technology keywords. Most existing technology keyword analyses have focused on identifying correlation structures between keywords. However, since most technologies are developed based on previously developed preceding technologies, the causal analysis of technology is important. Therefore, our research can contribute to the analysis and management of technology across various fields. A company can establish an efficient strategy for new R&D related to target technology using our proposed method.

In our future work, we will apply topological data analysis (TDA) to the GCM for a more advanced understanding of the causal structure between technology keywords. Topology applies low-dimensional geometric structure to high-dimensional complex data to show how local parts are connected globally [36,37]. Therefore, we use TDA to identify the topological structure between technology keywords. This allows us to explain not only the causal relationships between individual keywords, but also the causal structure between all keywords. We call this new method TDA-GCM. By using the TDA-GCM method, we can understand the overall causal structure of technology and, based on this, conduct various technology management activities for target technologies.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Jun, S. Keyword Data Analysis Using Generative Models Based on Statistics and Machine Learning Algorithms. *Electronics* **2024**, *13*, 798. [CrossRef]
2. Jun, S. Patent Keyword Analysis Using Bayesian Zero-Inflated Model and Text Mining. *Stats* **2024**, *7*, 827–841. [CrossRef]
3. Jun, S. Zero-Inflated Text Data Analysis using Generative Adversarial Networks and Statistical Modeling. *Computers* **2023**, *12*, 258. [CrossRef]
4. Shin, H.; Lee, H.J.; Cho, S. General-use unsupervised keyword extraction model for keyword analysis. *Expert Syst. Appl.* **2023**, *233*, 120889. [CrossRef]
5. Bzhalava, L.; Kaivo-oja, J.; Hassan, S.S. Digital business foresight: Keyword-based analysis and CorEx topic modeling. *Futures* **2024**, *155*, 103303. [CrossRef]
6. Xue, D.; Shao, Z. Patent text mining based hydrogen energy technology evolution path identification. *Int. J. Hydrog. Energy* **2024**, *49*, 699–710. [CrossRef]
7. Reher, L.; Runst, P.; Thomä, J. Personality and regional innovativeness: An empirical analysis of German patent data. *Res. Policy* **2024**, *53*, 105006. [CrossRef]
8. Coccia, M.; Roshani, S. Path-Breaking Directions in Quantum Computing Technology: A Patent Analysis with Multiple Techniques. *J. Knowl. Econ.* **2024**, 1–34. Available online: https://link.springer.com/article/10.1007/s13132-024-01977-y (accessed on 1 September 2024).
9. Park, S.; Jun, S. Zero-Inflated Patent Data Analysis Using Compound Poisson Models. *Appl. Sci.* **2023**, *13*, 4505. [CrossRef]
10. Park, S.; Jun, S. Patent Analysis Using Bayesian Data Analysis and Network Modeling. *Appl. Sci.* **2022**, *12*, 1423. [CrossRef]
11. Kim, J.-M.; Jun, S. Graphical causal inference and copula regression model for apple keywords by text mining. *Adv. Eng. Inform.* **2015**, *29*, 918–929. [CrossRef]
12. Uhm, D.; Jun, S. Zero-Inflated Patent Data Analysis Using Generating Synthetic Samples. *Future Internet* **2022**, *14*, 211. [CrossRef]
13. Julia, S.; Robinson, D. *Text Mining with R.*; O'Reilly: Sebastopol, CA, USA, 2017.
14. Feinerer, I.; Hornik, K. *Package 'tm' Version 0.7-13, Text Mining Package*; CRAN of R Project; R Foundation for Statistical Com-puting: Vienna, Austria, 2024.
15. Goodrich, M.T.; Tamassia, R.; Goldwasser, M.H. *Data Structures and Algorithms in Python*, 1st ed.; Wiley: Hoboken, NJ, USA, 2013.
16. Sucar, L.E. *Probabilistic Graphical Models Principles and Applications*; Springer: New York, NY, USA, 2015.
17. Kalisch, M.; Mächler, M.; Colombo, D.; Maathuis, M.H.; Bühlmann, P. Causal Inference Using Graphical Models with the R Package pcalg. *J. Stat. Softw.* **2012**, *47*, 1–26. [CrossRef]
18. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; MIT Press: Cambridge MA, USA, 2012.
19. Theodoridis, S. *Machine Learning A Bayesian and Optimization Perspective*; Elsevier: London, UK, 2015.
20. Pearl, J. Causal diagrams for empirical research. *Biometrika* **1995**, *82*, 669–710. [CrossRef]
21. Hogg, R.V.; McKean, J.M.; Craig, A.T. *Introduction to Mathematical Statistics*, 8th ed.; Pearson: Upper Saddle River, NJ, USA, 2018.
22. Bruce, P.; Bruce, A.; Gedeck, P. *Practical Statistics for Data Scientists*; O'Reilly Media: Sebastopol, CA, USA, 2020.
23. Hilbe, J.M. *Modeling Count Data*; Cambridge University Press: New York, NY, USA, 2014.
24. Cameron, A.C.; Trivedi, P.K. *Regression Analysis of Count Data*, 2nd ed.; Cambridge University Press: New York, NY, USA, 2013.
25. Li, X.-J.; Tian, G.-L.; Zhang, M.; Ho, G.T.S.; Li, S. Modeling Under-Dispersed Count Data by the Generalized Poisson Distribution via Two New MM Algorithms. *Mathematics* **2023**, *11*, 1478. [CrossRef]
26. Montgomery, D.C.; Peck, E.A.; Vining, G.G. *Introduction to Linear Regression Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
27. Sidumo, B.; Sonono, E.; Takaidza, I. Count Regression and Machine Learning Techniques for Zero-Inflated Overdispersed Count Data: Application to Ecological Data. *Ann. Data Sci.* **2023**, *11*, 803–817. [CrossRef]
28. Roback, P.; Legler, J. *Beyond Multiple Linear Regression: Applied Generalized Linear Models And Multilevel Models in R.*; CRC Press: Boca Raton, FL, USA, 2021.
29. USPTO, The United States Patent and Trademark Office. Available online: http://www.uspto.gov (accessed on 1 April 2024).
30. KIPRIS, Korea Intellectual Property Rights Information Service. Available online: www.kipris.or.kr (accessed on 1 April 2024).
31. Sepah, S.C.; Jiang, L.; Peters, A.L. Long-Term Outcomes of a Web-Based Diabetes Prevention Program: 2-Year Results of a Single-Arm Longitudinal Study. *J. Med. Internet Res.* **2015**, *17*, e92. [CrossRef] [PubMed]
32. Nakamura, K.A.; Kim, N. Digital Therapeutics in Hearing Healthcare: Evidence-Based Review. *J. Audiol. Otol.* **2024**, *28*, 159–166.
33. Liu, M.; Schueller, S.M. Integrating Digital Therapeutics With Mental Healthcare Delivery. *J. Health Serv. Psychol. Off. J. Natl. Regist. Health Serv. Psychol.* **2024**, *50*, 77–85. [CrossRef]
34. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020. Available online: http://www.R-project.org (accessed on 1 April 2024).
35. Kalisch, M. *Package 'pcalg' Ver. 2.7–11, Methods for Graphical Models and Causal Inference*; CRAN of R Project; R Foundation for Statistical Computing: Vienna, Austria, 2024.

36. Wasserman, L. Topological Data Analysis. *Annu. Rev. Stat. Its Appl.* **2018**, *5*, 501–532. [CrossRef]
37. Chazal, F.; Michel, B. An introduction to Topological Data Analysis: Fundamental and practical aspects for data scientists. *Front. Artif. Intell.* **2021**, *4*, 667963. [CrossRef] [PubMed]